中文图书分类号：TP391
UDC：39
学 校 代 码：10005

硕士专业学位论文

PROFESSIONAL MASTER DISSERTATION

论 文 题 目 ：基于人脸表情识别的学习情绪分析与应用研究

论 文 作 者 ：吴亚娜

专业类别/领域 ：电子与通信工程

指 导 教 师 ：贾克斌教授

论文提交日期 ：2022 年 5 月

UDC：39　　学校代码：　10005
中文图书分类号：TP391　　学　号：S201961241

北京工业大学硕士专业学位论文

（全日制）

题　　目：基于人脸表情识别的学习情绪分析与应用研究

英文题目：RESEARCH ON LEARNING EMOTION ANALYSIS AND APPLICATION BASED ON FACIAL EXPRESSION RECOGNITION

论　文　作　者：吴亚娜
学　　科：电子与通信工程
研　究　方　向：图像处理与模式识别技术
申　请　学　位：工程硕士专业学位
指　导　教　师：贾克斌教授
所　在　单　位：信息学部
答　辩　日　期：2022 年 5 月
授予学位单位：北京工业大学

摘 要

互联网的飞速发展为在线教育的普及提供了良好的平台，当前新冠疫情的严峻形势也更加凸显了远程在线教育的重要性。与传统线下课堂相比，在线教育虽然具有师资充足，课程选择丰富，不受时间地点约束等优点，但却不具备教师学生面对面进行情感交流的优势，教师难以观察到学生的学习情绪，然而学生的学习情绪对于教学效果有着不容忽视的影响。因此，目前对于在线教育而言，师生交互困难是急需解决的问题。学生的面部表情传达了他们对教学的直接感受和理解程度，是学生学习情绪的直接反映。同时，表情识别作为一项非入侵式技术，对于学习者而言接受度高，对仪器设备要求低。利用表情识别技术分析学生学习情绪，能帮助教师及时掌握学生学习状态，对不良情绪作出及时回应，从而设计更加合理化的教学内容，加强师生互动，提高教学质量。

目前，虽然面向在线教育应用的表情识别技术取得了相当大的进展，但还存在一些急待解决的问题。关于数据集方面，学习情绪具有其特殊性，对应的表情类别不能按照表情识别领域认可的基本表情类别一概而论，需要针对学习情绪设计相应的表情类别。并且人脸数据涉及隐私，开源的大规模数据集较少。由此导致国内目前缺乏用于学习情绪研究的数据集。关于算法设计方面，目前已有的表情识别模型为了提高识别精度，通常将模型设计得复杂庞大，导致实际应用部署时对硬件成本要求较高，需要对模型进行轻量化的研究。同时，在线教育存在大规模学生同时上课的情况，现有的表情识别模型在保证识别精度的前提下，推理速度无法满足同时识别大量学生的学习情绪的需求。这些问题为基于人脸表情识别的学习情绪研究带来了重大的挑战。

针对上述问题，本文重点研究了学习背景下的人脸表情识别算法，并开发了学习情绪分析系统。本文主要工作内容和取得的成果如下:

（1）针对目前学习背景下人脸表情数据集空缺的问题，构建了学习情境下的学生自发学习情绪人脸表情数据集。首先，通过调研教育领域对于学习情绪的认定，结合对在线课堂上学生的访谈调查，确定了愉悦、疲劳、悲伤、惊讶、厌倦、中性六种可以反映常见学习情绪的表情类别。然后，收集了真实课堂环境下的学生自发表情数据，通过模型标注和人工审核相结合的方法进行数据标注。最后，通过网络爬虫和数据增强等方式扩充数据集。最终得到了适用于学习情景下表情识别研究的人脸数据集 LE-FER，包含 10000 张人脸图像和 6 类学习表情。

（2）针对实际应用部署时对模型轻量化和泛化性的要求，设计了一种基于

多尺度特征结合注意力机制的轻量级人脸表情识别模型 Multi-scale Feature Net（MSFNet）。首先，该模型借鉴了密集连接的思想实现了特征图的重用，并利用不同尺度的卷积核使模型获取到多尺度的特征，提高了识别精度。其次，提出了一种“渐进式”轻量级结构，实现通道间信息交互的逐渐递减，在尽可能保证精度的前提下对模型的大小进行优化。最后，在模型中引入注意力机制，有助于特定通道信息的高效传播。该模型的参数量仅 0.33M，在 LE-FER 数据集上实现了 89.12%的准确率，在各开源数据集上均表现了较好的性能，实现轻量化设计的同时保证了较高的识别精度和良好的泛化性。

（3）针对在线教育系统应用时对于模型识别实时性的要求，设计了一种基于多层级特征结合动态判别机制的快速轻量级人脸表情识别模型 Multi-ghostnet。首先，该模型利用相似特征图之间存在信息共享的特性，提出用多层级的简单线性变换取代普通卷积来获得相似特征图，以此实现模型的压缩。多层级变换使得特征图之间的线性变换依次叠加，显著加深了网络，并有效减少网络中由于直接进行粗糙的单层变换而造成的性能下降。另外，提出自适应动态判别机制，对各层级特征图重要性进行权重分配，最大程度地利用有效信息。最后，引入残差连接，从而保留较多的原始人脸信息，提高模型精度。该模型在兼顾识别精度，参数量和计算量的同时，实时识别速度达到25ms/帧，综合性能较优，适用于在线教育环境下大规模学生上课时进行实时学习情绪监测。

（4）设计并实现了一种针对在线教育应用的学习情绪分析系统。基于已经训练好的表情识别模型，采用 PyQt5 框架搭建了学习情绪分析系统，将表情识别模型部署至平台上。该系统主要包括用户登录，学生注册，学生签到，单人学习情绪分析，群体分析等几个功能模块。现场试运行表明，系统每 10 帧抽取一帧画面进行识别，连续测试了数百帧，能顺利实现实时高效识别。同时，经过 PyInstaller 打包成.exe 格式后，该系统可以灵活应用于各主流操作系统，实用性和扩展性较高。教师端通过访问该平台，可以直观地获取学生的表情和学习状态，将面向在线教育的人脸表情识别技术推向了实际应用。

关键词：学习情绪；在线教育；表情识别；轻量级；卷积神经网络

Abstract

The rapid development of the Internet has provided a good platform for the popularity of online education, and the current severe situation of the new crown epidemic has further highlighted the importance of distance online education. Compared with traditional offline classes, online education has the advantages of sufficient teachers, rich course selection and not being bound by time and place, but it does not have the advantage of face-to-face emotional communication between teachers and students, and it is difficult for teachers to observe students' learning emotions, yet students' learning emotions have a significant impact on teaching effectiveness that cannot be ignored. Therefore, the difficulty of teacher-student interaction is a pressing issue for online education. Students' facial expressions convey their direct feelings and understanding of the teaching and learning process, and are a direct reflection of their emotions. At the same time, expression recognition, as a non-invasive technology, is highly acceptable to learners and requires low instrumentation. The use of expression recognition technology to analyze students' learning emotions can help teachers keep track of students' learning status and respond to undesirable emotions in a timely manner, so that they can design more rationalized teaching content, enhance teacher-student interaction and improve teaching quality.

At present, although considerable progress has been made in expression recognition technology for online education applications, there are still some pressing issues to be resolved. Regarding the dataset, learning emotions have their own specificity, and the corresponding expression categories cannot be generalized according to the basic expression categories recognized in the field of expression recognition, and corresponding expression categories need to be designed for learning emotions. Moreover, face data involves privacy, and there are few open source large-scale datasets. This has led to a lack of datasets for learning emotion research in China. Regarding algorithm design, the currently available expression recognition models are usually designed to be complex and large in order to improve recognition accuracy, resulting in high hardware cost requirements for practical application deployment and the need for lightweight research on the models. At the same time, online education has a large number of students attending classes at the same time, and the existing expression recognition models are unable to meet the demand of recognizing the

learning emotions of a large number of students at the same time while ensuring the recognition accuracy and inference speed. These problems pose significant challenges for the study of learning emotions based on face expression recognition.

To address these problems, this thesis focuses on face expression recognition algorithms in a learning context and develops a learning emotion analysis system. The main work and results achieved in this thesis are as follows:

(1) To address the current problem of the vacancy of face expression datasets in learning contexts, a face expression dataset of students' spontaneous learning emotions in learning contexts is constructed. Firstly, through research on the identification of learning emotions in education, combined with interviews and surveys of students in online classrooms, six categories of expressions reflecting common learning emotions were identified: happy, tired, sad, surprised, bored and neutral. Then, data on students' spontaneous expressions in real classroom environments were collected and the data were annotated through a combination of model annotation and manual review. Finally, the dataset was expanded by means of web crawlers and data augmentation. The final face dataset LE-FER, which is applicable to the study of expression recognition in learning situations, was obtained, containing 10,000 face images and 6 types of learning expressions.

(2) A lightweight face expression recognition model Multi-scale Feature Net (MSFNet) based on multi-scale features combined with attention mechanism was designed to address the requirements of model lightweight and generalization when deployed in practical applications. Firstly, the model is based on the idea of dense connectivity to reuse feature maps, and uses convolution kernels of different scales to obtain multi-scale features and improve recognition accuracy. Secondly, a "progressive" lightweight structure is proposed to achieve decreasing information interaction between channels and to optimize the size of the model while ensuring accuracy as much as possible. Finally, an attention mechanism is introduced into the model to facilitate the efficient propagation of channel-specific information. The model, with only 0.33M number of parameters, achieves 89.12% accuracy on the LE-FER dataset and performs well on all open source datasets, achieving a lightweight design while ensuring high recognition accuracy and good generalization.

(3) A fast and lightweight face expression recognition model, Multi-ghostnet, based on multi-level features combined with dynamic discriminative mechanism is designed to meet the requirement of real-time model recognition in online education

system applications. The model is then compressed. The multi-level transformation allows the linear transformations between feature maps to be superimposed sequentially, significantly deepening the network and effectively reducing the performance degradation caused by direct coarse single-level transformations in the network. In addition, an adaptive dynamic discrimination mechanism is proposed to assign weights to the importance of the feature maps at each level to maximize the use of valid information. Finally, a residual join is introduced so as to retain more of the original face information and improve the model accuracy. The model achieves a real-time recognition speed of 25ms/frame while taking into account the recognition accuracy, number of parameters and computational effort, and is suitable for real-time monitoring of learning emotions during large-scale student classes in an online education environment.

(4) A learning emotion analysis system is designed and implemented for online education applications. Based on the trained expression recognition model, the PyQt5 framework was used to build the learning emotion analysis system, and the expression recognition model was deployed to the platform. The system mainly includes several functional modules such as user login, student registration, student sign-in, single person learning sentiment analysis and group analysis. The trial run on site showed that the system was able to recognize one frame every 10 frames, and hundreds of frames were tested continuously, and it was able to achieve real-time and efficient recognition smoothly. At the same time, after PyInstaller packaged into .exe format, the system can be flexibly applied to all mainstream operating systems, practicality and scalability is high. By accessing the platform, the teacher side can intuitively access the students' expressions and learning status, pushing the face expression recognition technology for online education into practical application.

Keywords: Learning emotion, Online education, Facial expression recognition, Lightweight, Convolutional neural network

目　录

第 1 章 绪论

1.1 课题背景及研究意义

随着互联网技术的蓬勃发展，以人工智能，大数据等现代化技术为依托展开的教育体系改革在教育领域掀起了热潮。近几年，中国教育支出稳步增长，2020 年我国财政教育支出达 36337.18 亿元，同比增长 4.43%[1]，国家对教育事业的重视保证了在线教育行业的不断发展。自 2020 年以来，受疫情影响，教育部联合包含 27 个省市在内的区域陆续开放了国家级、省市级在线教学平台，在线教育取得了更进一步的发展。同年 7 月，国家发展改革委联合 13 个部门印发意见书，明确提出要大力发展融合性在线教育，未来将逐渐形成线上教学与线下教学相辅相成，良性互动的主流模式。2022 年 3 月 28 日，国家智慧教育平台正式上线，进一步推进了在线教育的服务智能化，数据精准化和管理全量化。

在线教育虽然具有可以突破时间地点限制，学习资源丰富多元等优点，但是与传统线下课堂相比，也有其不容忽视的局限性。首先，在线教育环境下，老师与学生没有面对面接触，缺乏情感交流，沟通效率降低。另外，在线教育由于不受上课地点限制，为节省教育资源，上课人数往往较多。对教师而言，既要完成课堂教学内容安排，又要观察把握学生学习状态，个人精力有限往往难以兼顾。这些局限性会引发一些不良情况，最终导致在线课堂教学质量达不到预期。比如由于在线教育环境更为宽松自由，学生缺乏监督，容易注意力不集中，导致学习效率低下；当学生对课堂内容产生疑惑时，难以及时向教师反映，容易滋生厌学情绪；线上教学课堂互动率低，学习环境不固定，学生极易在较为放松的环境下出现疲惫困倦状态，无法达到教学要求。这些消极情绪如果无法被及时排解，将对在线教学的质量和效率造成严重影响。鉴于以上问题，针对学生学习情绪的研究对于在线教育尤为重要。

目前的学习情绪研究方法，从采集信号类型的角度，大致可以分为基于学习者生理信号和基于学习者行为两大类。基于学习者生理信号的方法需要专业的仪器设备，并且作为入侵式的检测方法，本身会给学习者带来一定的压力。基于学习者行为的研究方法，通常需要收集学生的头部位置姿势改变和眼睛眨眼频率等信息。若学生姿势未改变，也未出现因为困倦而导致的频繁眨眼或闭眼情况，但实质上在发呆走神时，系统将无法检测到。心理学领域 [2]研究表明，人脸面部表情传达的信息量在人类情感表达中占比达 55%。利用人脸表情识别对学生学习情绪进行研究，对于学习者而言为非入侵式监测，接受度更高，同

时对于仪器设备要求而言，只需要学习者参与在线教学时配备摄像头即可，成本较低。

因此，面向在线教育进行人脸表情识别的研究，有利于弥补线上教学师生交互困难的缺陷，能辅助教师实时掌握学生的学习状态，对教学进度作出及时调整；对于推进教育信息化，普及智慧教育具有十分重要的价值。

1.2 国内外研究现状

1.2.1 学习情绪识别研究现状

在教育领域，学习者在学习过程中的情绪一直是研究人员的关注重点。积极的学习情绪会提高学生对教学内容的感兴趣程度，触发主动学习机制；而消极懈怠的学习情绪则会使学生逃避教学，给教学质量带来严重的不良影响。对于在线教育而言，由于师生无法直接接触，情绪的沟通会受到一定限制，因此在线学习系统有必要具备识别学习者的学习情绪并反馈给教师，从而引导学生进入合适的学习状态的功能。目前，以信号采集方式不同进行归类划分，情绪识别技术可以分为基于学习者生理信号进行识别和基于学习者行为进行识别两个主要类别。

基于生理信号的情绪识别需要专业仪器，通过测量学习者的心电图、脑电图、肌电图和瞳孔直径等方式实现情绪识别。AlZoubi O[3]使用EEG(Electroencephalogram)信号识别参与者的 10 种情绪。Healey J[4]等研究者将心电图,皮肤电和肌电图信号相结合，实现对 8 种情绪的识别。Belle A [5]通过采集心电图（ECG）信号，观察学习者注意力与心率之间的关系。Lee H[6]使用电极采集受试者的脑电图信息，从而监测其注意力的变化。

基于学习者行为则是通过监测学习者的表情，语音和姿态等间接的方式实现情绪的识别。D'Mello S[7]等人用来分析学习者情绪的指标有学习者座位和靠背的压力变化、语音监测信息以及面部特征等。Stanley D[8]使用 3D 体感摄影机 Kinect 采集学习者的头部位置移动和身体姿态变化等信息，评估其注意力。吴沧海[9]等人提出的情感计算方式，涉及学习者的表情、身体姿势位置、眼睛闭合状态等多种信息的采集。卢希[10]通过传感器采集学习者打哈欠频率、眨眼频率等信息，分析其学习状态。易佳玥[11]通过跟踪学习者眼角和虹膜的位置变化，评估其学习状态。熊碧辉[12]等人提出将视线变化与头部位置变化等信息进行融合来判断学习者的专注程度，目的是为了检测学习者表面上头部位置姿势正常，实则视线落在屏幕外的情况。

这些方法使学习情绪的识别具备可行性，但都或多或少存在一些局限性。

例如，采集学习者的生理信号需要专业的设备仪器，成本较高；同时对于学习者而言，是一种入侵式检测，本身会带来一定的压力，进而会影响其学习情绪。基于学习者行为的研究方法，往往通过监测学习者身体头部位置姿势的变化以及眨眼频率等，判断其学习状态。虽然此类方法适用于大多数场景，但是当学生表面上身体头部姿态正常，视线也落在屏幕上，实则分心失神时，无法检测到。

1.2.2 基于传统方法的人脸表情识别研究现状

利用人脸表情识别对学生学习情绪进行研究，对于学习者而言接受度更高，同时不需要其他专业设备，实现更简单。

表情识别算法流程主要分为三步：预处理、特征提取和表情分类。对于传统方法而言，特征提取主要通过手工特征选择完成，其结果直接影响模型最终性能。运动特征、频率特征和灰度特征是目前常用来学习的三类特征。运动特征指的是人脸处于不同表情状态时面部相应特征点的运动信息；频率特征指的是表情类别不同时，在不同的频率下的分解结果也不同，这种方法识别速度较快；灰度特征则是先将人脸图像进行灰度归一化，然后利用不同表情具有不同灰度图像的特性来识别表情。

传统表情识别技术中，常用的特征提取方法有以下几种：

（1）整体学习法，识别不同表情下的图像整体差别，如主成分分析法。

（2）局部学习法，对人脸不同部位重要性赋予不同权重。典型方法有面部动作编码法（Facial Actions Code System，FACS）[13-14]，Gabor 小波法等。

（3）形变提取法，通过提取人脸表情变化时脸部各个部位的形变情况来进行识别。典型的方法有：点分布模型法（Point Distribution Model，PDM）、主成分分析法（PCA）、运动模板法（Active Shape Model，ASM）和 Gabor 小波法等。

1.2.3 基于深度学习的人脸表情识别研究现状

复杂环境下的表情识别往往面临着各种噪声的干扰，此时传统方法识别速度慢，准确率低，难以应对。随着人工智能，深度学习的蓬勃发展，表情识别领域也迎来了技术改革，基于深度学习的方法处理大规模数据的效率远超过传统方法，同时大幅度减少了对图像预处理和特征提取的依赖，在不同环境下鲁棒性表现更佳。

近几年，表情识别领域提出了很多性能优秀的网络框架，不同网络根据数据集和应用背景的不同各有侧重。

Zhao 等人提出一种自动特征选择网络(FSN) [15] , 有效过滤掉与表情识别任

务不相关的特征。Zeng 等人[16]提出 IPA2L 的框架来获取多数据集中表情数据的真实潜在标签。Yang 等人[17]提出一种“表情残差学习”的方法滤除掉与表情无关的面部信息。Kuo 等人[18]设计了一种利用循环时间单元信息的表情识别模型，并提出了一种光照增强方案，用来缓解过拟合现象。Wang[19]等人提出了一种自愈网络(Self-Cure Network)来解决大规模表情识别不确定性问题，该网络包括自注意力重要性加权、排序正则化和重标签三个模块。Adrian[20]等人提出一种优化卷积神经网络超参数的方法，以提高面部情绪识别上下文的准确性，在超参数离散值定义的搜索空间中，利用随机搜索算法生成和训练模型，确定网络的最优超参数。Andrey[21]等人提出轻量级卷积神经网络的多任务训练，用于表情识别和面部属性(年龄，性别，种族)的分类。Amir[22]等人提出一种深度注意中心损失(DACL)方法，自适应地选择显著特征元素的子集进行增强表情识别，利用 CNN 中的中间空间特征映射作为上下文，评估各特征的重要性权重。

上述深度学习方法面向不同数据集针对性各有不同，尽管上述模型使表情识别任务取得了较大进展，但这些模型往往为了提高识别精度，不断加深加宽网络结构，在实际应用部署时，对硬件成本要求太高。另外，在线教育由于不受地点限制，可能出现大规模学生同时上课的情况，教师要及时监控到每个学生的学习状态，则需要模型推理速度尽可能快。因此，面向在线教育应用时，目前的表情识别模型分类与训练针对性不强，同时模型轻量级和推理速度也有待改进。

1.3 研究目标与主要内容

1.3.1 存在问题与研究难点

深度学习的不断崛起极大地促进了计算机视觉各领域的飞速发展，针对表情识别的研究已有大量性能优秀的模型框架被提出，但针对在线教育应用的人脸表情识别算法，目前还面临以下问题有待解决，这也是本项研究的挑战和机遇：

（1）迄今为止，虽然国内外研究人员构建了许多与学习情绪研究相关的人脸表情数据集，但存在一些不足。首先，由于学习情绪具有特殊性，表情识别领域公认的基本表情类别并不适用，对学习情绪类别认定的差异，导致数据集建立标准不统一。其次，人脸表情涉及学生隐私，学生自发表情采集困难，数据处理过程复杂，导致开源大规模数据集较少。另外，由于人种差异，不同国家人脸面部单元信息并不具备普适性，导致已有数据集适用性不佳。

（2）目前表情识别模型往往为了提高识别精度，不断加深加宽网络，但与

此同时会带来网络参数量和计算量的大幅度增加。实际应用部署时，要同时处理大量学生的视频监控数据，庞大复杂的模型对硬件成本要求非常高，这极大地限制了表情识别技术在教育领域的应用。因此，在不断提高模型识别精度的同时，对其进行轻量化的改进也是目前研究的重点和难点。

（3）由于不受地点限制，在线教育授课人数不唯一，可能出现大规模学生同时上课的情况。此时，系统若想实现实时监测每位学生的表情变化，对算法的识别速度要求较高，因此在压缩模型体积的同时，还要考虑进一步提高模型的推理速度。

综上所述，面向在线教育应用，对表情识别技术进行进一步研究尤为重要。

1.3.2 研究目标与主要内容

本文的研究目标一是建立适用于学习情绪研究的课堂背景下的人脸表情数据集；二是搭建高精度，低参数量，低运算量的表情识别模型，同时加快模型的推理速度；三是搭建学习情绪分析系统，辅助教师实时掌握学生学习状态，提高教学质量。

基于对人脸表情识别技术的研究，以及面向在线教育实际应用的需求，本文的主要工作内容包括：

（1）针对目前学习背景下人脸表情数据集空缺的问题，构建了学习情境下的学生自发学习情绪人脸表情数据集。首先，通过调研教育领域对于学习情绪的认定结合对在线课堂上学生的访谈调查，确定了愉悦、疲劳、悲伤、惊讶、厌倦、中性六种可以反映常见学习情绪的表情类别。然后，收集了真实课堂环境下学生自发表情数据，通过模型标注和人工审核相结合的方法进行数据标注。最后，通过网络爬虫和数据增强等方式扩充数据集。最终得到了适用于学习情景下表情识别研究的人脸数据集 LE-FER，包含 10000 张人脸图像和 6 类学习表情。

（2）针对实际应用部署时对模型轻量化和泛化性的要求，设计了一种基于多尺度特征结合注意力机制的轻量级人脸表情识别模型 Multi-scale Feature Net（MSFNet）。首先，该模型借鉴了密集连接的思想实现了特征图的重用，并利用不同尺度的卷积核使模型获取到多尺度的特征，提高了识别精度。其次，提出了一种“渐进式”轻量级结构，实现通道间信息交互的逐渐递减，在尽可能保证精度的前提下对模型的大小进行优化。最后，在模型中引入注意力机制，有助于特定通道信息的高效传播。该模型的参数量仅 0.33M，在 LE-FER 数据集上实现了 89.12%的准确率，在各开源数据集上均表现了较好的性能，实现轻量化设计的同时保证了较高的识别精度和良好的泛化性。

（3）针对在线教育系统应用时对于模型识别实时性的要求，设计了一种基于多层级特征结合动态判别机制的快速轻量级人脸表情识别模型 Multi-ghostnet。首先，该模型利用相似特征图之间存在信息共享的特性，提出用多层级的简单线性变换取代普通卷积来获得相似特征图，以此实现模型的压缩。多层级变换使得特征图之间的线性变换依次叠加，显著加深了网络，并有效减少网络中由于直接进行粗糙的单层变换而造成的性能下降。另外，提出自适应动态判别机制，对各层级特征图重要性进行权重分配，最大程度地利用有效信息。最后，引入残差连接，从而保留较多的原始人脸信息，提高模型精度。该模型在兼顾识别精度，参数量和计算量的同时，实时识别速度达到25ms/帧，综合性能较优，适用于在线教育环境下大规模学生上课时进行实时学习情绪监测。

（4）设计并实现了一种针对在线教育应用的学习情绪分析系统。基于已经训练好的表情识别模型，采用 PyQt5 框架搭建了学习情绪分析系统，将表情识别模型部署至平台上。该系统主要包括用户登录，学生注册，学生签到，单人学习情绪分析，群体分析等几个功能模块。现场试运行表明，系统每 10 帧抽取一帧画面进行识别，连续测试了数百帧，能顺利实现实时高效识别。同时，经过 PyInstaller 打包成.exe 格式后，该系统可以灵活应用于各主流操作系统，实用性和扩展性较高。教师端通过访问该平台，可以直观地获取学生的表情和学习状态，将面向在线教育的人脸表情识别技术推向实际应用。

论文的工作得到了以下项目的资助：

- 北京工业大学教育教学研究课题：疫情防控背景下在线教学“慕课化”的工程化方法研究（ER020B015）
- 北京工业大学教育教学研究课题：交互式教学模式在提高课堂效率与培养创新型人才中的研究与实践（002000514111027）

1.4 论文研究内容与结构安排

本文一共分为六章，研究内容和结构安排如图1-1所示，各章节内容如下：

第一章为绪论，首先论述了研究面向在线教育应用的人脸表情识别技术的背景及意义；然后详细研究了目前常用的学习情绪监测方法，并分别调研了基于传统方法和深度学习方法的表情识别技术研究现状；最后阐述了论文的重点研究内容与结构安排。

第二章为相关技术介绍，首先介绍了人脸检测的相关技术；然后介绍了人脸表情识别的完整流程及涉及到的关键技术；最后介绍了卷积神经网络的相关概念。

第三章构建了学习情绪人脸表情数据集，首先通过调研教育领域对于学习情绪的认定，结合对在线课堂上学生的访谈调查，确定了愉悦、疲劳、悲伤、惊讶、厌倦、中性六种可以反映常见学习情绪的表情类别；然后介绍了数据的采集处理和标注过程；最后列出了最终数据集的分布。

第四章搭建了基于多尺度特征结合注意力机制的轻量级人脸表情识别模型，首先介绍了建模分析过程和模型设计细节；然后对模型中涉及的相关数学运算以及模型训练过程中用到的数据集给出了详细介绍；最后对实验结果作出了详细说明。

第五章搭建了基于多层级特征结合动态判别机制的快速轻量级人脸表情识别模型，首先详细介绍了建模思想和模型各模块具体网络结构；然后介绍了模型中涉及到的相关数学运算；最后通过实验，将模型的参数量，计算量，识别精度，推理速度与目前较优模型进行了详细的对比。

第六章搭建了学习情绪分析系统，首先介绍了系统开发平台和功能模块设计；然后对用户登录、学生注册、学生签到、单人信息、群体分析等功能和界面进行了具体的介绍和展示；最后对平台性能进行了分析。

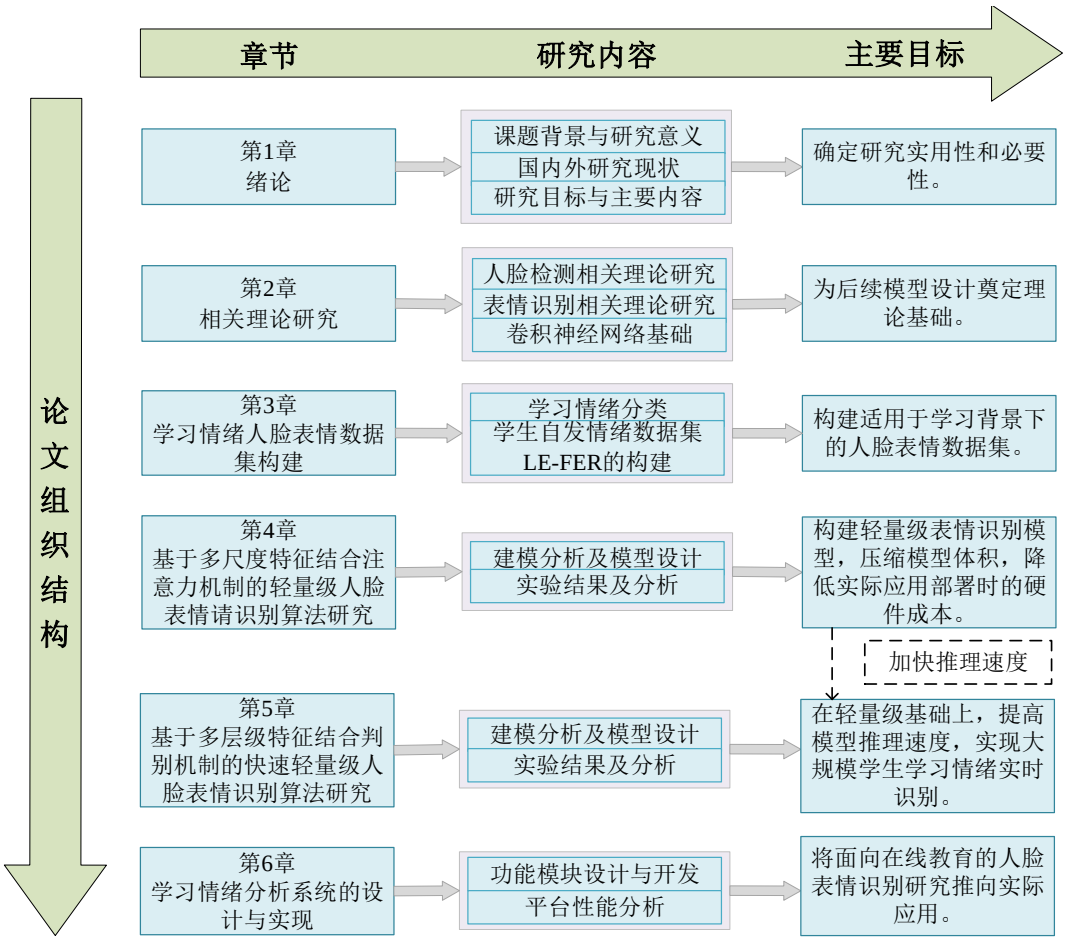


图 1-1 论文研究内容与结构安排

Fig. 1-1 The research content and structure arrangement of the thesis

第 2 章 相关理论研究

2.1 引言

在利用表情识别技术分析学生学习情绪前，需要先进行人脸检测，检测的准确性以及速度都会对最终系统的整体性能产生较大影响。检测得到候选人脸框后，进入表情识别算法流程，该算法首先对人脸进行归一化等预处理操作，然后进入特征学习阶段，模型学习与表征人脸表情相关的特征并滤除掉无关特征，最后通过表情分类得到最终的类别标签。为了为后续的算法设计建立理论基础，本章详细研究了人脸检测和表情识别流程中各关键步骤涉及到的相关技术，并研究学习了卷积神经网络的相关知识。

2.2 人脸检测相关理论研究

2.2.1 人脸检测算法流程

人脸检测的核心目的是获取人脸位置信息。传统人脸检测算法流程主要分为三步：第一步，在输入人脸图像中裁剪出一系列候选框，检测判断是否包含人脸信息；第二步，特征提取，对包含人脸信息的候选框进行人脸特征的学习；第三步，特征分类，通过分类器获得满足条件的人脸候选框的坐标位置以及分类置信度等信息。

深度学习方法由于其高效性也逐渐被广泛运用到智能人脸检测过程中。运用深度学习技术实现人脸智能检测有两个主要研究方向,一是把目标检测技术领域中用于多任务检测的算法应用集成到人脸智能检测系统中,如 Faster-RCNN 等;另有一类则是专门设计针对应用于人脸检测系统的卷积神经网络,如 Cascade CNN、MTCNN 等。

2.2.2 多任务卷积神经网络人脸检测算法

多任务卷积神经网络[23]（Multi-task Convolutional Neural Network, MTCNN）是一种可以同时完成人脸检测和人脸对齐的人脸检测模型，检测精度高，能在部分遮挡的情况下检测到人脸；检测速度快，可以实现实时处理。考虑到在线教育表情识别对于实时性和遮挡鲁棒性有较高要求，本研究选用该模型进行表情识别前的人脸检测。

MTCNN 模型的核心结构包括一个图像金字塔和三个级联的 CNN 模块。其中，图像金字塔由原始图像通过双线性插值操作进行尺度缩放获得。

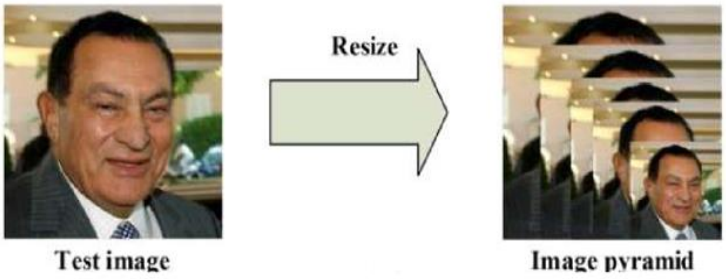


图 2-1 图像金字塔

Fig. 2-1 Image pyramid

然后依次通过三个级联的多任务 CNN 网络，如图 2-2 所示，此网络结构用来处理人脸判定，人脸候选框回归和人脸特征点定位三个任务。其中，判定时选择交叉熵损失函数，如公式（2-1）所示，候选框回归和特征点定位时选择欧氏距离损失函数，分别如公式（2-2），（2-3）所示。该网络结构同时还会实现对人脸特征由粗到细的提取过程。

$$L_i^{\det} = -(y_i^{\det}\log(p_i) + (1-y_i^{\det})(1-\log(p_i))) \tag{2-1}$$

$$L_i^{box} = \| \hat{y}_i^{box} - y_i^{box} \|_2^2 \tag{2-2}$$

$$L_i^{landmark} = \| \hat{y}_i^{landmark} - y_i^{landmark} \|_2^2 \tag{2-3}$$

首先将“图像金字塔”送入第一个浅层的 CNN 结构 P-Net，快速获得人脸候选框，并通过非极大值抑制法（NMS）合并高度重合的候选框。然后通过中间层的 CNN 结构 R-Net，对候选框进行筛选和矫正，滤除掉非人脸窗口。第三步通过更复杂的深层 CNN 结构 O-Net，对结果做进一步精化，并输出五个人脸特征点坐标信息。

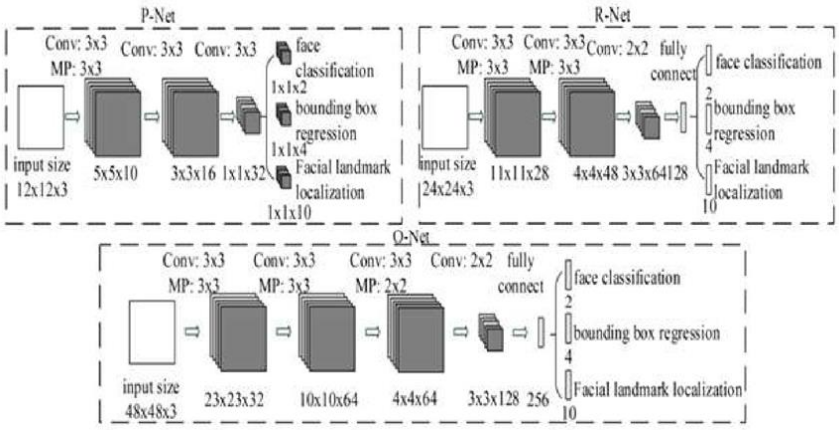


图 2-2 P-Net, R-Net 和 O-Net 网络结构图

Fig. 2-2 Network structure of P-NET, R-NET, and O-NET

2.3 表情识别相关理论研究

如图 2-3 所示，表情识别算法流程主要分为三步：预处理，特征提取，表情分类。表情识别任务应用背景复杂，在进行数据采集时，会受到光照、角度、距离等各个因素的影响，导致采集到的人脸图像出现角度偏移或包含过多噪声信息，需要进行一系列预处理，使图像满足表情识别的要求。然后是特征提取过程，模型需要学习人脸各特征点的信息，并提取出与表情相关的特征。最后是分类过程，对学习到的特征进行表情分类，得到对应的类别标签。

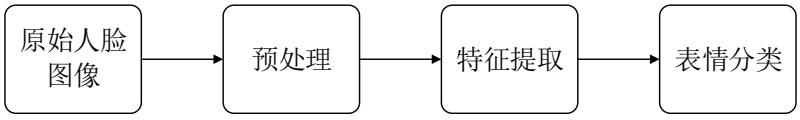


图 2-3 表情识别流程

Fig. 2-3 Facial expression recognition process

2.3.1 图像预处理

在采集学生处于学习状态下的表情时，为了得到学生自发的表情，研究各种不同的学习状态，采集时长一般长达十几到几十分钟，同时为了不给被采集者压力，摄像头与被采集者会有一定距离。此时，由于学生头部转动和光线改变等原因，采集到的图像往往包含过多噪声信息，不利于后续对其进行表情识别。通过预处理操作可以滤除噪声，保留更多有用信息，一般需要进行以下操作：

（1）人脸归一化处理

在构建自发表情数据集时，学生并非总是正视摄像头，环境光线也并不总是保持不变。因此，收集到的人脸表情很可能由于角度的偏移，光线的明暗而缺乏一些表征表情的关键区域信息，如嘴唇，眼睛等。此时，需要对人脸进行光照归一化和姿态归一化处理，获得对比度合适的正脸图像。其中，光照归一化早期通过高斯函数差分[24]和离散余弦变换[25]等方法实现，近几年的研究[26-28]结合直方图均衡化取得了不错的效果。三维重建[29]和 GAN 网络则是近几年姿态归一化任务中常用到的方法，常用的 GAN 网络有 DR-GAN[30]、FF-GAN[31]、TP-GAN[32]等。

（2）人脸图像灰度化处理

原始人脸图像因为光照差异存在明暗变化不均的问题，导致图像信号发生改变。灰度化处理能对此进行补偿，突出图像细节，有助于提取梯度信息；同时对人脸图像进行降维处理，降低模型训练的参数量。灰度化处理的主要思想

是对原始 RGB 图像的三种颜色分量进行线性组合，比如选择最大分量或者平均分量等。

（3）人脸图像直方图均衡化处理

当原始人脸图像由于过度曝光或者曝光不足，图像像素集中分布在过高或过低亮度时，图像细节模糊，整体对比度低。此时通过直方图均衡化，能改变原始图像各像素的灰度值，增强人脸图像对比度，使图像细节纹理更清晰。

2.3.2 特征提取

表情识别算法流程中，模型在特征提取阶段对图像进行特征学习，即在检测到的人脸图像中提取出表征表情的有效特征，滤除与表情无关的噪声信息。

传统特征提取算法[33]主要有基于 Harr-like 特征法、基于 Gabor 特征法、特征点跟踪法等。Harr-like[34]特征包括一系列遵循 Alfred Harr 提出的阶跃函数的特征，例如边缘特征、中心特征等；基于傅里叶变换的 Gabor 函数依据小波理论与 Gabor 特征，Lyons 等[35]在频域中分析面部图像，通过 Gabor 小波滤波器计算面部表情信息。特征点跟踪法是使用特征点的位移变化表征输入图像的表情特征。Cai[36]提出将改进的加速鲁棒特征(SURF)算法与金字塔 kade -lucas-tomasi(P-KLT)匹配算法相结合，作为一种新的特征点跟踪法。

深度学习方法中，CNN 由于其处理大批量样本的潜力，也被用在表情识别任务中。输入人脸图像矩阵经过卷积层操作后，得到特征映射矩阵，主要包括通道内特征，通道间特征和通道信息融合。然后通过池化层对特征进行降维处理。同时，CNN 可以灵活应用批量标准化[37]和 Dropout[38]等机制防止训练过程中过拟合现象的发生。

2.3.3 表情分类

表情分类，即对提取到的表情特征进行分类，并输出对应的的表情类别标签。传统的分类算法有 K 最近邻法（K-Nearest Neighbors, KNN），支持向量机法（Support Vector Machine, SVM），贝叶斯网络法等。KNN 法[39-41]的核心思想是在训练数据集中寻找与待分类图像最相似的 K 个数据，这 K 个数据有各自对应的类别，其中占比最多的类别被认定为待分类图像的类别。SVM 作为一种典型的二分类器，通常用于线性可分离数据，核心思想是寻找一个最优决策面将所有样本分成不同类别。贝叶斯网络法[42-44]是一种概率图形模型，用条件概率表征不同结果出现的可能性，具备处理不确定性问题的能力，需要学习多个变量的线性参数。 深度学习方法在进行分类时，将特征提取模块得到的特征送入池化层进行降维处理，然后通过 Softmax 分类器确定最终的样本类别。

2.4 卷积神经网络基础

2.4.1 卷积神经网络的结构

卷积神经网络（Convolutional Neural Networks，CNN）[45] 起源于人工神经网络，最早由 Fukushima[46]提出，是模仿人类神经系统结构搭建的一系列模型，包括感知机，单层神经网络，多层神经网络等。多层神经网络经过进一步优化改造得到卷积神经网络，传统的多层网络结构输入为向量信息，不包含图像相邻像素的结构信息，这导致特征的提取和融合及其有限。而 CNN 包含卷积层、池化层、全连接层等多种不同功能的隐藏层，如图 2-4 所示，这种结构能有效利用图像的空间信息，高效提取图像特征。

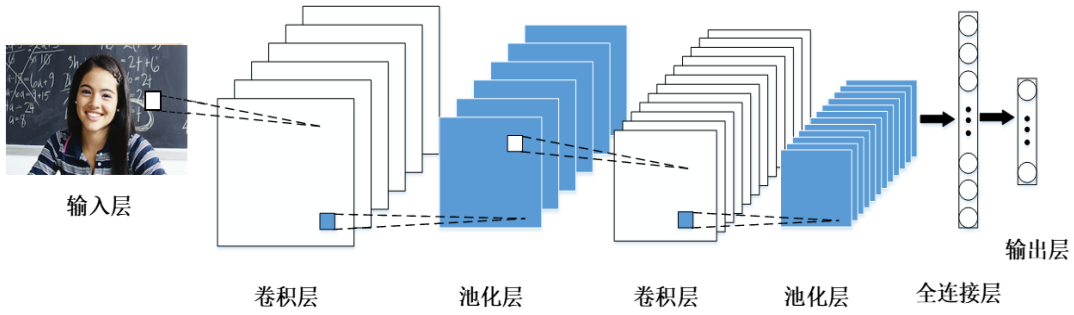


图 2-4 卷积神经网络结构图

Fig. 2-4 Convolutional neural network structure diagram

（1）输入层

CNN 中输入层一般具备两个主要功能[47]，一是读取原始数据，数据类型具有多样性，可以是数字图像或者音频信息等。二是对获取到的样本进行预处理。需要进行预处理的主要原因一是因为原始样本单位类型不统一，影响模型收敛速度、加长了训练时间；二是因为训练过程中激活函数受到值域限制，未经过预处理的原始数据得到的输出有可能超出规定的值域范围。

数据标准化是较为常见的预处理方式，常用方法有以下几种：

（a）Z-Score normalization

Z-Score 标准化是根据样本的标准差和均值进行标准化处理，如公式（2-4）所示，将输入样本转换为符合标准正态分布的数据。

$$x^{*}=\frac{x-\overline{x}}{\sigma},\sigma=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2} \tag{2-4}$$

其中 x 为原始样本，$\overline{x}$ 表示原始均值，σ 表示原始标准差，N 表示样本总数，i 表示第 i 个样本。

（b）Min-max normalization

Min-max 标准化通过线性变换将原始样本映射到指定范围内，若指定范围为[0,1]，则变换公式如（2-5）所示：

$$x^* = \frac{x - min}{\max - \min} \tag{2-5}$$

其中 max，min 分别表示样本数据的最大值和最小值，x 表示原始数据，x^* 为标准化后的数据。

（c）log 函数转换

log 函数转换公式如（2-6）所示：

$$x^* = \frac{\log(x)}{\log_{10}(\max)} \tag{2-6}$$

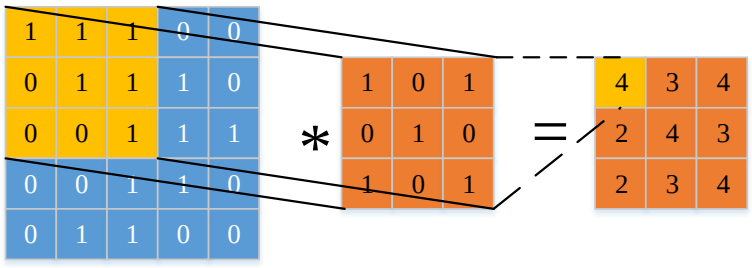
其中，x 为原始输入数据，x^* 为归一化后的数据。

（2）卷积层

CNN 的特征提取功能主要通过卷积层进行卷积操作实现，公式表示如（2-7）所示，其中，$f(x,y)$ 表示输入，$g(x,y)$ 为卷积核，m 、n 表示卷积核大小，$output(x,y)$ 表示输出。

$$output(x,y) = f(x,y) * g(x,y) = m\sum_{n} f(x-m, y-n)g(m,n) \tag{2-7}$$

图 2-5 展示了对5×5矩阵和3×3矩阵做卷积操作的过程。



1×1+1×0+1×1+0×0+1×1+1×0+0×1+0×0+1×1=4

图 2-5 卷积运算示例

Fig. 2-5 Example of convolution process

卷积核选定起始位置开始运算，将卷积核各位置的值和图像对应位置的像素值进行相乘再求和，得到输出目标位置的像素值，然后卷积核根据步长移动，在下一个位置重复上述计算过程，直至覆盖整个目标区域。

（3）池化层

随着网络的不断加深，高维数据会带来参数量的大幅度增长，此时需要池化层来对高维特征进行降维处理。常见池化方式有以下两种：

（a）最大池化（Max Pooling），实现过程如图 2-6 所示：

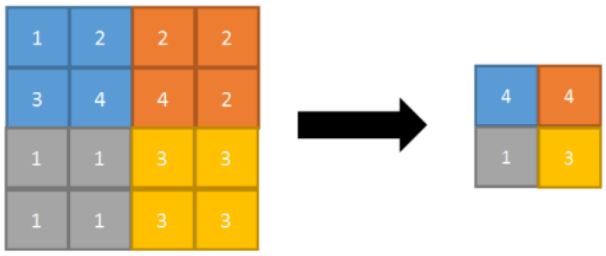


图 2-6 最大池化

Fig. 2-6 Max Pooling

输入是一个 4×4 矩阵，过滤器参数为 2×2，在原输入矩阵中，从左上角开始，过滤器取大小范围为 2×2 的邻域内的最大元素，然后以 2 为步长，依次向右向下移动，每移动一次得到一个最大元素，最终输出为 2×2 矩阵。最大池化能很好地保留纹理特征，适用于对细粒度要求高的模型。

（b）平均池化（Mean Pooling），实现过程如图 2-7 所示：

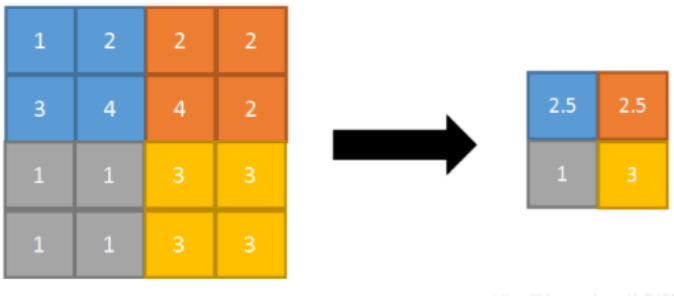


图 2-7 最大池化

Fig. 2-7 Max Pooling

与最大池化不同之处在于，平均池化取每个邻域内元素的平均值作为该邻域对应输出。平均池化能更好地保留样本的背景信息，但会降低图像质量，使图像变得模糊。

（4）全连接层

全连接层用来对前面卷积层获得的特征进行加权求和，结构如图 2-8 所示，其每个神经元结点都与上层的所有结点相连。

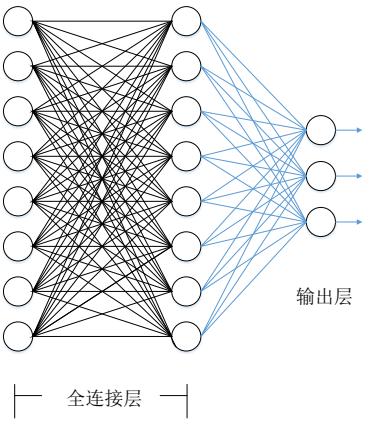


图 2-8 全连接层

Fig. 2-8 Fully connected layer

2.4.2 卷积神经网络的训练

CNN 的训练过程如图 2-9 所示，主要流程如下：

（1）模型对输入向量的权重进行初始化分配；

（2）计算各隐藏层的输出；

（3）根据输出结果计算其与目标值之间的偏量；

（4）若偏量超过允许范围，则计算各隐藏层的偏量，否则结束训练；

（5）根据上一步求得的偏量对权重进行更新。

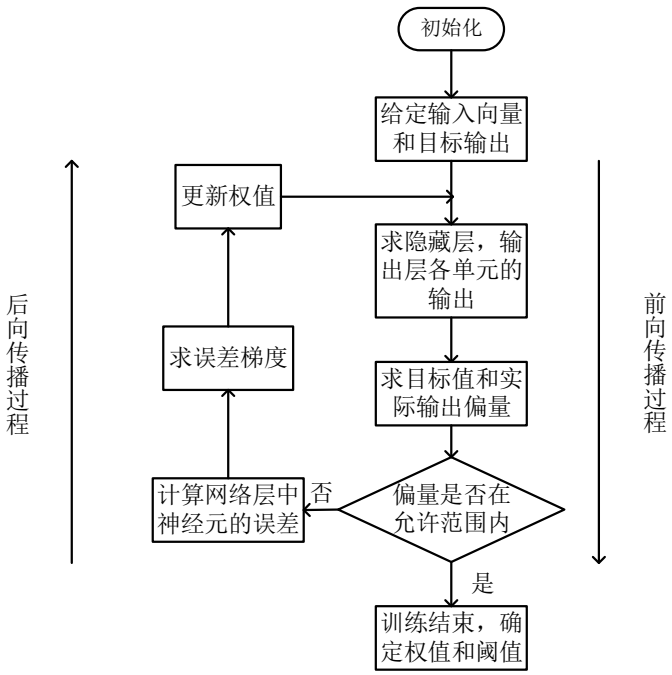


图 2-9 卷积神经网络训练过程

Fig. 2-9 Training process of convolutional neural network

2.5 小结

本章首先研究了人脸检测任务的基本流程，并重点阐述了多任务卷积神经网络人脸检测模型。接着重点研究了表情识别的算法流程，详细介绍了人脸图像预处理过程中涉及到的归一化，灰度化和直方图均衡化技术，以及特征提取和表情分类过程中常用到的传统方法和深度学习方法。最后对卷积神经网络的有关定义和运算做了详尽说明，为后续算法设计和模型优化奠定了理论基础。

第 3 章 学习情绪人脸表情数据集构建

3.1 引言

学生学习情绪是教学效果评价的重要指标。在教育领域，学习者在学习过程中的情绪变化一直是研究人员的关注重点。积极的学习情绪会提高学生对教学内容的感兴趣程度，触发主动学习机制；而消极懈怠的学习情绪则会使学生逃避教学，给教学质量带来严重的不良影响。学生的面部表情传达了他们对教学的直接感受和理解程度，是学生学习情绪的直接反映。

利用表情识别技术分析学生的学习情绪，首要需要解决的就是数据集的问题。迄今为止，虽然国内外研究人员构建了许多与学习情绪研究相关的人脸表情数据集，但还存在一些不足。首先，由于学习情绪具有特殊性，表情识别领域公认的基本表情类别并不适用，对学习情绪类别认定的差异，导致数据集建立标准不统一。其次，人脸表情涉及学生隐私，学生自发表情采集困难，数据处理过程复杂，导致开源大规模数据集较少。另外，由于人种差异，不同国家人脸面部单元信息并不具备普适性，导致已有数据集适用性不佳。

鉴于以上问题，本章构建了基于学习情境的学生自发表情数据集 LE-FER，首先，通过调研教育领域对于学习情绪的认定，结合对在线课堂上学生的访谈调查，确定了愉悦、疲劳、悲伤、惊讶、厌倦、中性六种可以反映常见学习情绪的表情类别。然后，收集了真实课堂环境下学生自发表情数据，通过模型标注和人工审核相结合的方法对数据进行标注。最后，通过网络爬虫和数据增强等方式扩充数据集。最终得到了包含 10000 张人脸图像和 6 类学习表情的适用于学习情景下表情识别研究的人脸数据集。

3.2 学习情绪分类

在线学习学生情绪分析在教学效果评价中起着重要作用。学习者的面部表情传达了他们对教学过程的直接感受和对教学内容的理解程度，通过了解学生的面部表情，教师可以根据学生的学习情绪对课程进度进行调整，最大程度的避免消极情绪对教学效果产生的不良影响。

随着教育领域对学生心理的关注度不断提高，表情识别技术由于其非侵入性和低成本设备受到了越来越多研究人员的关注。然而，经过研究发现，表情识别领域里的基本表情分类[48]并不能与学习情绪一一对应，类似愤怒、轻视等

基本表情类型在学习过程中极少出现。相反由于缺乏监管，不少学生会在学习过程中出现疲倦甚至睡眠的情况，此时基本表情类型已不再适用。因此对于学习情绪类别的确定，应该着重关注学生在真实学习环境下自发产生的出现频率较高的表情。

通过对目前国内外较为认可的学习情绪相关文献进行研究，对研究者们划分的学习情绪[49-55]类型归纳总结为表 3-1。

表 3-1 学习情绪分类表

Tab. 3-1 Classification of learning emotions

研究人员	表情类别	表情类型
Mampusti[49]	4	无聊、疑虑、专注、失落
何秀玲[50]	5	快乐、困惑、疲劳、惊讶、中性
唐康[51]	5	疑惑、倾听、理解、不屑、抗拒
Whitehill 等[52]	6	悲伤、愉悦、惊讶、厌烦、恐惧、中性
Sharma 等[53]	7	愉悦、中性、厌烦、惊讶、恐惧、愤怒、悲伤
Tonguc 等[54]	7	悲伤、厌烦、愉悦、轻视、惊讶、恐惧、愤怒
Lehman 等[55]	8	疑虑、欣喜、好奇、沮丧、专注、骄傲、享受、希望

通过分析表中信息可知，对于学习情绪的分类，虽然不同研究人员的分类准则各有不同，但基本可以分为积极，中性，消极三大类。同时部分相似情绪可以归为一类，比如愉悦、欣喜、享受等表情的类别边界比较模糊，对于模型而言很难学习到细微差异，且其划分对于教师评判学生学习状态而言差异不大。因此，综合考虑在划分学习情绪时，可以根据积极，中性，消极三大类各自延伸，保留出现频率较高的几类表情，并将相似表情合为一类。

为了进一步了解学生的学习情绪与教学内容的具体关联，本研究对疫情期间内，通过在线教育平台进行课程学习的大学生进行了访谈调查。根据学生反馈，当课程内容较为简单轻松时，学生往往会觉得心情愉悦，并能顺利跟上课程节奏；当课程内容新颖有趣时，学生的注意力会被最大程度吸引，出现惊讶等表情；当课程节奏过快，学生跟不上进度时，则会产生悲伤等情绪；而当课程内容重复多次，节奏过慢时，学生会出现厌烦心理；另外，在线教育环境下，一些自觉性较低的学生由于缺乏监管，容易感觉疲惫困倦。

因此，基于既有研究结合访谈考察，本文最终确定了包含愉悦（happy）、悲伤（sad）、惊讶（surprised）、厌烦（disgusted）、疲惫（tired），中性(neutral)在内的 6 类表情来对学习情绪进行划分。

3.3 学生自发情绪数据集 LE-FER 的构建

3.3.1 数据采集

本章学习情绪人脸表情数据集（Learning Emotion Facial Expression

Recognition, LE-FER）的构建流程如图 3-1 所示，数据来源包括两个渠道，一是通过教室摄像头采集到的课堂视频，对视频进行抽帧处理，得到包含学生人脸信息的图像；二是通过网络爬虫爬取课堂环境下特定表情类别的图像。针对上述图像，通过人脸检测获得人脸候选框后，经过归一化，灰度化和直方图均衡化等操作对人脸图像进行矫正处理，得到符合后续识别条件的标准人脸图像。再对其进行标注和扩充，最终得到课堂环境下学生自发人脸表情数据集 LE-FER。

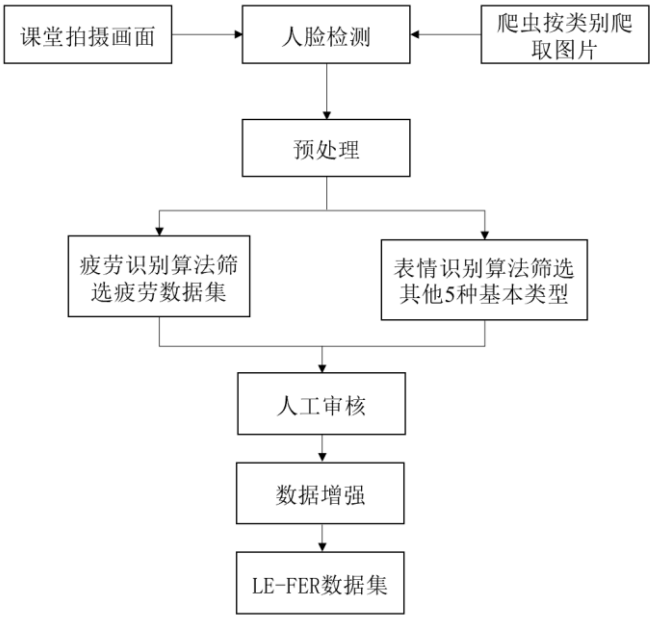


图 3-1 数据集构建过程

Fig. 3-1 Dataset construction process

（1）课堂画面采集

本研究收集数据前已征求教师和学习者同意，并签订承诺书，保证数据仅用于科学研究，不外传不上网不泄露个人隐私。为保证学生表情为自发表情，本研究选取了带有摄像头的教室上课时的监控录像，共收集了 30 段视频，每段视频 45 分钟，采集对象为在校大学生或研究生。由于视频序列过长，包含无用信息，因此首先通过人工观看审核，对视频进行剪切，保留了 50 段每段 15 分钟的有效视频，确保这些有效视频中有我们需要的表情信息，然后对课程视频帧进行抽帧处理，每 10 帧图像抽取一帧进行后续的人脸检测。

（2）网络爬虫采集

网络爬虫[56] 指按照制定的规律自动浏览搜寻互联网资讯的方法，这种规律称为网络爬虫算法。利用 Python 语言开发爬虫程序，在互联网信息中按类别自动化检索符合条件的人脸表情图像，流程如图 3-2 所示。

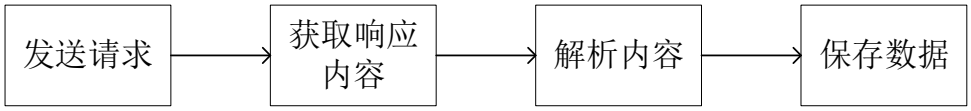


图 3-2 网络爬虫流程

Fig. 3-2 Flow chart of web crawler

3.3.2 人脸检测

采集到包含人脸信息的视频帧后，需要检测人脸，滤除无效背景信息，得到精确的人脸候选框。考虑到在线教育表情识别对于实时性和遮挡鲁棒性有较高要求，本研究选用第二章介绍的 MTCNN 模型进行人脸检测。MTCNN 可以同时完成人脸检测和人脸对齐，检测精度高，能在部分遮挡的情况下检测到人脸；检测速度快，可以实现实时处理。

3.3.3 预处理

对于图像识别任务，要想提高模型识别精度，一方面要不断优化模型设计，另一方面，输入图像质量对于特征提取效果的影响也不容忽视。由于光线差异，姿态变化，角度尺寸不统一等问题，原始人脸图像包含大量噪声，如果直接用作训练数据集，后续算法识别的结果和精度将会大打折扣。因此，如图 3-3 所示，需要对人脸图像进行一系列预处理操作。

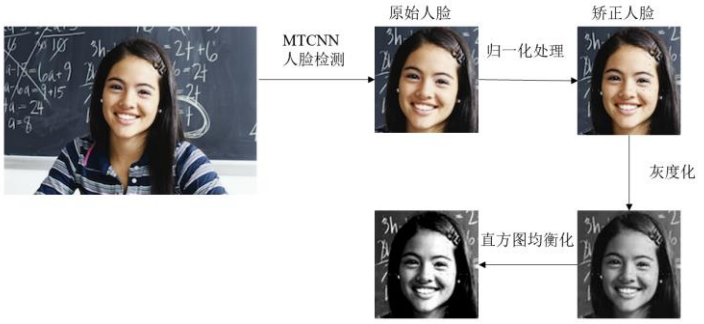


图 3-3 预处理流程图

Fig. 3-3 Pretreatment flow chart

首先，通过光照归一化[24]和姿势归一化[30]得到矫正对齐的人脸。然后通过灰度化处理将图像降至单通道，突出图像细节。最后，通过直方图均衡化，增大图像对比度，突出纹理信息。

3.3.4 数据集标注

本课题定义的学习情绪对应的人脸表情总共 6 种，其中高兴（happy）、悲伤（sad）、惊讶（surprised ）、厌烦（disgusted）、中性(neutral)属于目前表情识

别任务中常见的五种基本表情类型，选择使用目前表现最优的表情识别模型之一——Self-Cure Network （SCN）[19]框架对这几种表情进行标注，疲劳作为学习背景下的特定表情，通过 opencv 疲劳检测算法进行标注。

（1）SCN 表情识别模型

虽然心理学家和人类学家对于人类表情与情绪的关系做了大量研究，但是不容忽视的是，人类情绪复杂多变，同一表情未必表征相同情绪。同时，不同情绪之间可以相互流动转换，这导致各类表情的边界模糊，难以明确定义。这些问题也为表情识别任务带来了难题，当人为进行数据标注时，一些主观差异性难以避免，因此数据标签可能不唯一，数据规模越大，这种不确定性更明显。为了解决表情识别标注困难问题，SCN 表情识别模型被提出，结构如图 3-4 所示，用于降低面部表情不确定性的权重，从而学习更加鲁棒的特征。

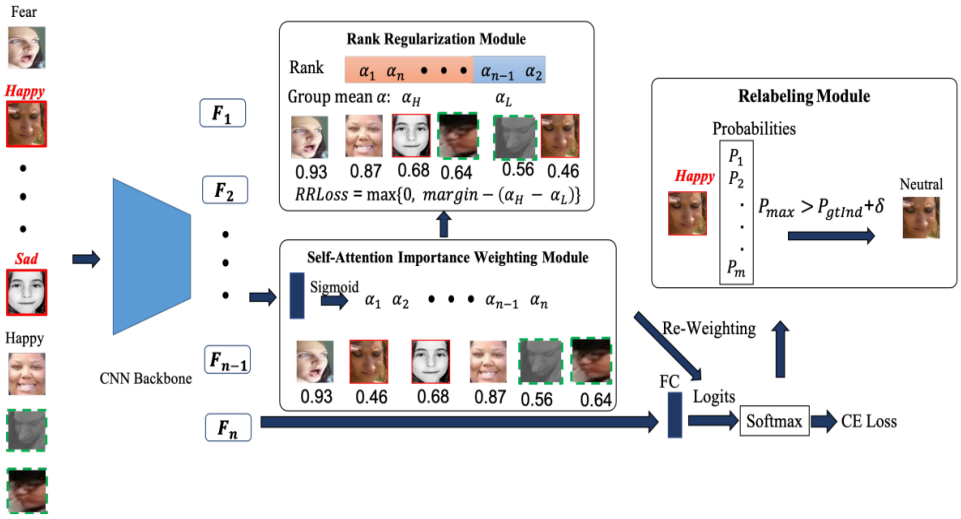


图 3-4 SCN 网络结构图

Fig. 3-4 SCN network structure

该模型包括自注意重要度加权、正则化操作和重标签三个部分。第一个部分学习训练样本中每一张人脸图像对于模型训练的重要性，该样本属于某类表情的概率越高，则对其赋予更高的权重。第二个部分则对这些权重进行排序正则化，以突出确定性高的样本。第三个部分尝试对不确定性的易错样本进行重新归类，并更改其标签。通过大量实验证明，SCN 方法能有效抑制不确定性的表情数据，适用于对大规模数据进行识别和标注。

（2）疲劳检测

疲劳表情则是本文针对课堂背景的特点，加入的特定表情类型，通过基于 Opencv 的疲劳检测算法对其进行标注。其算法流程如图 3-5 所示。

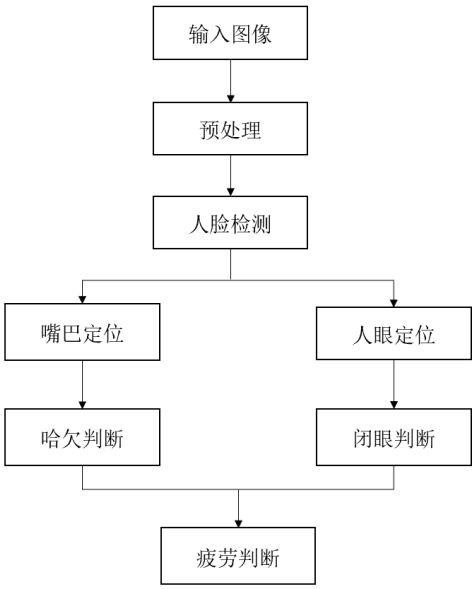


图 3-5 疲劳检测算法流程图

Fig. 3-5 Flow chart of fatigue detection algorithm

（3）数据增强

对于图像识别任务而言，训练样本是决定模型最终性能的关键因素。样本容量过少，模型训练时会出现过拟合现象，导致训练精度虚高，而测试精度相差甚远。同时，样本类别的不平衡也会对识别性能产生较大影响，这点在分类任务中表现得更为明显。然而，原始数据的获取有时需要消耗大量的人力物力。通过数据增强操作，原始数据能变换得到更多数据，同时还可以针对具体类别进行特定数据扩充，可以有效缓解数据量不足和数据不平衡的问题。

数据增强[57]指不增加实际数据，通过对现有数据进行一系列变换操作，得到更多等价的有效数据。常用的数据增强方法大致可以分为几何变换和色彩变换两大类。其中，几何变换包括对图像进行翻转、裁剪等各类操作，色彩变换则是通过添加高斯噪声、颜色扰动等方法来实现。

通过上述数据增强操作，最终得到的 LE-FER 数据集中各类别数量分布如表 3-2 所示，数据集共包含 10000 张学生自发人脸表情图像，对训练集和测试集按 4：1 的比例进行了随机划分。

表 3-2 学生表情数据库

Tab. 3-2 Student expression database

表情类型	训练集	测试集	合计
愉悦	1600	400	2000
悲伤	1200	300	1500
惊讶	1200	300	1500
厌恶	1200	300	1500
疲倦	1600	400	2000
中性	1200	300	1500
合计	8000	2000	10000

3.4 小结

本章主要解决目前学习背景下，人脸表情数据集空缺的问题。首先，通过调研教育领域对于学习情绪的认定结合对在线课堂上学生的访谈调查，确定了愉悦、疲劳、悲伤、惊讶、厌倦、中性六种可以反映常见学习情绪的表情类别。然后，收集了真实课堂环境下学生自发表情数据，通过模型标注和人工审核相结合的方法进行数据标注。最后，通过网络爬虫和数据增强等方式扩充数据集。最终得到了适用于学习情景下表情识别研究的人脸数据集 LE-FER，包含 10000 张人脸图像和 6 类学习表情。

第 4 章 基于多尺度特征结合注意力机制的轻量级人脸表情识别算法研究

4.1 引言

在图像识别领域，深度学习模型往往通过不断加深加宽网络来提高模型识别精度，与此同时，匹配模型运行的硬件配置也越来越高。而在在线教育情境下，利用表情识别研究学生学习情绪，需要处理大量视频数据，硬件成本过高极大地阻碍了该项应用的实际部署推广。因此，针对在线教育情景下的表情识别任务，一方面需要不断深入研究如何提高模型识别精度，另一方面如何压缩模型，降低实际应用部署时的硬件成本也是一个重要的研究方向。

鉴于以上背景，为了满足实际应用时的轻量化要求，考虑解决模型参数量过大导致难以实际应用部署，以及现有模型泛化性能有待提高的问题，本章提出了一种基于多尺度特征结合注意力机制的轻量级人脸表情识别模型 Multi-scale Feature Net（MSFNet）。

4.2 建模分析

4.2.1 Densenet

目前对卷积神经网络进行优化主要有两个方向，一是通过增加网络深度来拟合更加复杂的特征输入，逐层抽象，提高非线性表达能力，经典的网络结构如 ResNet[58]，二是通过增加网络宽度提取更为丰富的特征，经典网络如 Inception[59]。然而网络宽度和深度的增加均会带来模型参数量的大幅度增加。

不同于上述两种模型优化方式，DenseNet[60]网络提出了密集连接思想，即每层网络都与其前面所有层网络进行短路连接。通过这种方式，梯度在训练过程中能进行反向传播，加深了网络，并对原始特征进行复用，从而能在参数量更少的情况下实现更优的性能。

DenseNet 网络结构如如图 4-1 所示，主要由 Dense Block 模块和 Transition Layer 过渡层组成。

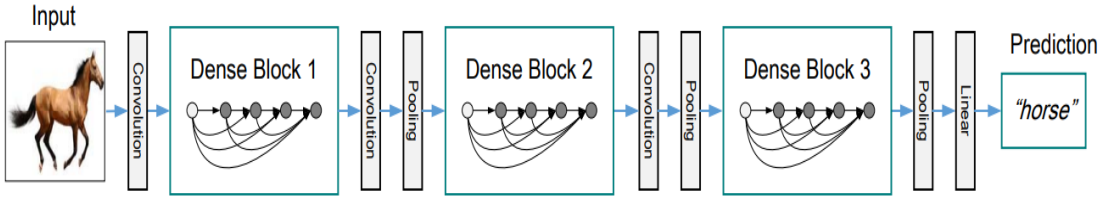


图 4-1DenseNet 网络结构

Fig. 4-1 The network structure of DenseNet

假设CNN包含l层，以x_0为输入，每层网络包含一个非线性变换$H_l(\cdot)$。$H_l(\cdot)$是包含了各项操作的复合函数，比如批量标准化(BN)、激活(ReLU)、池化(Pooling)等。x_l表示第l层的输出。

传统的 CNN 第$(l+1)^{th}$层的输入为第l^{th}层的输出，如（4-1）所示：

$$x_{l+1} = H_{l+1}(x_l) \tag{4-1}$$

为了实现特征复用，如图 4-2 所示，Dense Block 创新性地提出了密集连接模式，第l层的输入为前面所有层的输出，以$x_0,\ldots x_{l-1}$作为输入，输出表示为：

$$x_l = H_l\left([x_0, x_1, ..., x_{l-1}]\right) \tag{4-2}$$

其中$[x_0, x_1, ..., x_{l-1}]$表示前面所有层输出特征图的串联。

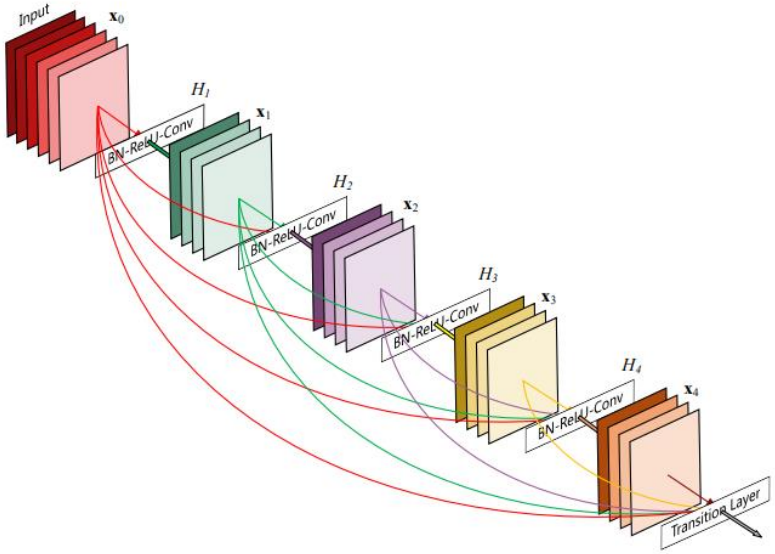


图 4-2 Dense Block

Fig. 4-2 Dense Block

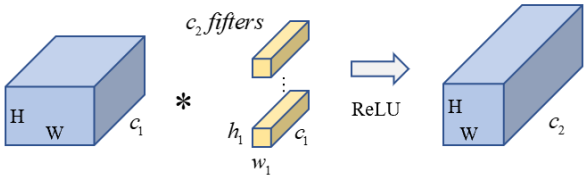
Transition Layer 作为过渡层用来连接相邻的 Dense Block。除了起到连接作用， Transition Layer 还具备压缩网络的功能，其压缩性能由压缩系数 θ（compression rate：$\theta \in (0,1]$）表征。当 $\theta < 1$ 时，对上一个 Dense Block 输出进行压缩，避免密集连接导致后续参数爆炸式增长。

对于表情识别任务而言，模型中每层神经网络提取的有效特征是有限的，在训练过程中随机舍弃一些网络层并不会破坏模型的收敛性。另外，目前表情识别模型往往针对特定数据集表现更优，考虑舍弃全连接方式，能在一定程度上提高模型的泛化性能。

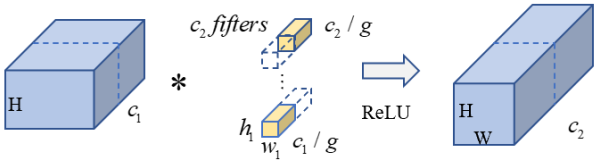
因此，本章借鉴 Densenet 密集连接思想，不再通过加宽加深网络来提升性能，而是从特征角度出发，通过密集连接，使网络中的每一层都可以学习到前面所有层的特征，提高有效特征利用率，在更少的参数量和计算量的条件下实现更高的准确率。

4.2.2 分组卷积

分组卷积（Group Convolution）在 AlexNet[61]网络中第一次被提出，如图 4-3 所示，输入特征图大小为 $C \times H \times W$，卷积核有 N 个。对于普通卷积而言，卷积核以及得到的输出特征图的数量均为 N，卷积核大小为 $C \times K \times K$，总参数量为 $N \times C \times K \times K$。分组卷积对输入进行分组处理后再按组分别进行卷积运算。若分组数为 M，则每组需要输入 C/M 张特征图，输出 N/M 张特征图。卷积核大小为(C /M) × K × K，每组有 N/M 个卷积核，总参数量为 N × (C/ M) × K × K，降低至原来的 1/ M。



(a) 标准卷积



(b) 分组卷积

图 4-3 标准卷积与分组卷积示意图

Fig. 4-3 Schematic diagram of standard convolution and grouping convolution

分组卷积可以看成是稀疏操作，可以在较少参数量的情况下获得更好的效果，相当于正则化操作，同时也保留了一部分通道交互信息，适用于在构建轻量级网络模型时取代普通卷积。

4.2.3 深度可分离卷积

深度可分离卷积[62]对普通卷积做了一些结构上的改变，可以作为一种轻量级的优化来替代标准卷积，分为 Depthwise 卷积和 Pointwise 卷积两步进行。

（1）Depthwise 卷积

与标准卷积操作中各通道共享卷积核不同的是，Depthwise 卷积会对不同通道采用各自特定的卷积核。对于一幅尺寸为 128×128×3 的 RGB 彩色输入图像，如图 4-4 所示，经过 Depthwise 卷积后，得到分别属于三通道的三个特征图。

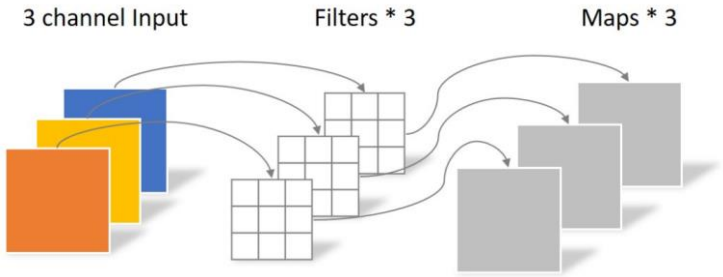


图 4-4 Depthwise 卷积

Fig. 4-4 Depthwise Convolution

（2）Pointwise 卷积

Pointwise 卷积是为了弥补 Depthwise 卷积过程中，各通道间未进行信息交互，导致无法生成扩展特征图的局限性。其运算过程与标准 1×1 卷积类似，会在通道方向上对深度卷积得到的特征图进行加权重组，以实现信息交互。

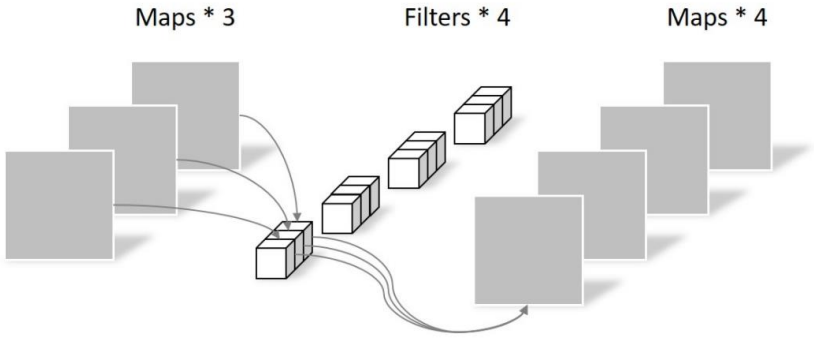


图 4-5 Pointwise 卷积

Fig. 4-5 Pointwise Convolution

表情识别作为对细粒度要求较高的分类任务，如果为了模型轻量级一味地牺牲网络深度，会导致最终分类结果较差。因此本章在搭建模型时，并未直接

舍弃普通卷积以减少网络深度，而是将其替换为深度可分离卷积，尽可能在保证精度的前提下对模型大小进行优化。但如果只用深度卷积，通道间信息交互不够，会影响表情识别最终的分类结果。因此提出“渐进式轻量级”思想，由 1×1 标准卷积向分组卷积，深度可分离卷积逐渐过渡，实现通道间信息传递的逐渐减弱，在压缩模型体积的同时尽可能保留更多有效信息，以保证模型最终实现较高的识别精度。

4.2.4 压缩激励模块

压缩激励模块在 SENet[63]网络中被提出，如图 4-6 所示，主要由两部分组成：

（1）Squeeze 模块。通常使用全局平均池化将原始特征图压缩至一维，原始输入特征图为 C×H×W，可以将其看作 C 张 H×W 的图像，每张图像用其像素的均值表示，得到 C×1×1 大小的向量。

（2）Excitation 模块。

将压缩模块的输出作为输入送入到激励模块，经过两个全连接层，为特征图的每个通道分配对应的重要性权重。然后将输出结果与原始输入相乘，得到最终的输出。

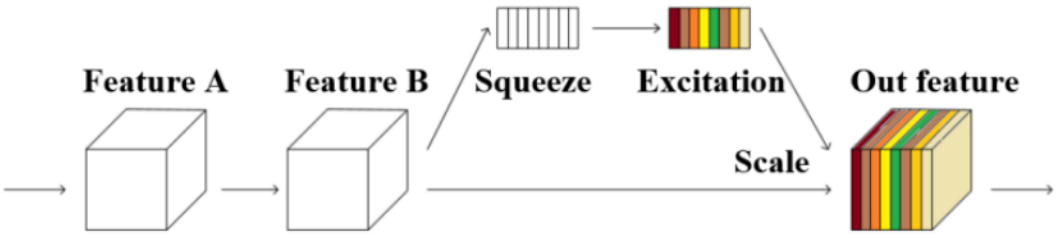


图 4-6 SE 模块

Fig. 4-6 SE module

SE 模块可以作为即用即插模块嵌入到各网络中，表情识别需要模型关注脸部特定特征，利用 SE 模块对特定区域赋予相应的重要性权重，有利于提升模型的分类精度。

4.3 模型设计

4.3.1 MSFNet

本章提出的网络模型借鉴了 Densenet 的密集连接思想，首先，通过特征复用和短路连接，节省网络参数，充分利用原始有效信息的同时缓解梯度消失问

题。然后，为了进一步减少模型参数量，采用了深度可分离卷积和分组卷积。另外，原始的深度可分离卷积通常使用 3×3 单一卷积核，感受野有限，不利于模型分类。本模型采用多分支，卷积核大小不同的深度可分离卷积，并加入通道注意力模块，为每个通道分配重要性权重，以提高模型的分类精度。

模型整体结构图如图 4-7 所示。

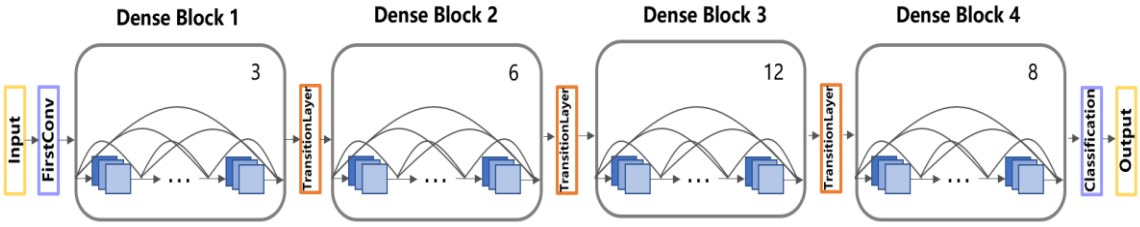


图 4-7 MSFNet 网络架构

Fig. 4-7 MSFNet network architecture

人脸表情图像经过预处理后[64]，首先经过 FirstConv 模块，将图像通道改为自定义通道数，然后进入特征学习模块，该模块包含四个 Denseblock 结构，相邻 Denseblock 之间通过 Transistionlayer 模块连接，最后经过 Classfication 模块，得到最终的分类结果。

4.3.2 Bottlenecklayer

为了对高维数据进行降维，CNN 结构中往往包含下采样层，此时特征图的大小将会改变，公式（4-2）不再适用。为了便于池化层进行下采样操作，本文将网络设计成四个 Denseblock 模块，每个模块由一个过渡层连接。Denseblock 结构可以减少网络的计算负荷，并在过渡层进行向下采样。四个 Denseblock 中依次包含 3，6，12，8 个子结构 Bottlenecklayer，算法流程可由以下公式表示：

$$F(x)=\left[P_w\left(M(x)\right),x\right] \tag{4-3}$$

$$M(x)=\left[M_1(x),M_2(x),...,M_i(x)\right] \tag{4-4}$$

$$M_i(x)=D_{w_i}\left(G_i(x)\right) \tag{4-5}$$

输入图像 x 经过 1×1 卷积，通道注意力机制和分组卷积处理后得到 $G_i(x)$，其中 i 表示的是分支的索引。 $D_w(\cdot)$ 表示深度可分离卷积。以 $G_i(x)$ 为输入，经过深度可分离卷积后得到 $M_i(x)$，$\left[M_1(x),M_2(x),...,M_i(x)\right]$ 表示的是各分支输出经过 concatenate 连接后得到的结果。 P_w 表示逐点卷积，$\left[P_w\left(M(x)\right),x\right]$ 表示对各分支输出进行逐点卷积之后再与原输入进行 concatenate 连接得到最终的输

出。

Bottlenecklayer 模块一共包括四个分支，具体网络结构如图 4-8 所示：

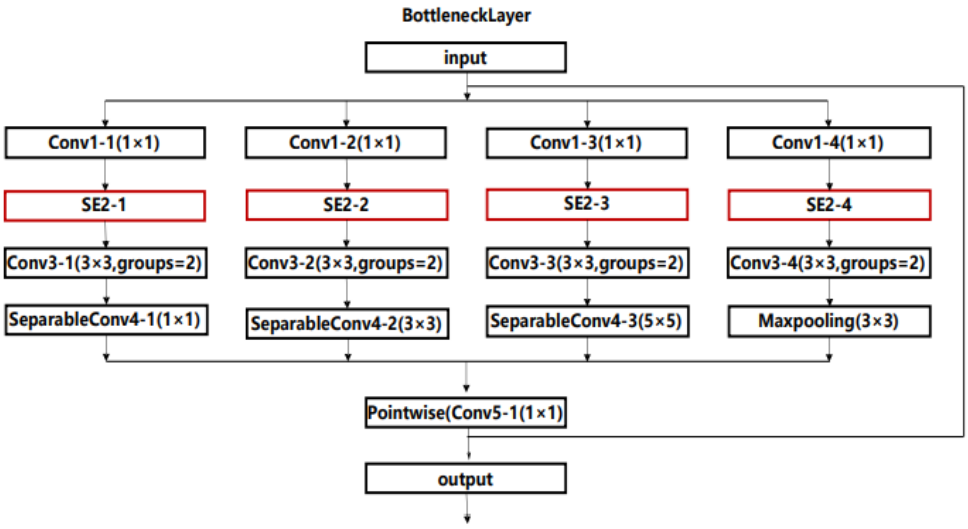


图 4-8 Bottlenecklayer 网络架构

Fig. 4-8 Bottlenecklayer network architecture

各分支[65]的输入图像依次通过 1×1 卷积进行降维，通道注意力模块进行权重分配；然后通过卷积核为 3×3，组数为 2 的分组卷积；接下来，前三个分支依次通过卷积核分别为 1×1，3×3，5×5 的多尺度深度可分离卷积，最后一个分支则通过 3×3 的最大池化层。将四分支输出进行 concatenate 连接，然后进行逐点卷积，最后与原始输入连接得到该模块最终输出。

4.3.3 其他模块

FirstConv 模块[65]: 由 2×2 平均池化层和 3×3 卷积层组成，用来更改经过预处理后的人脸图像的通道数。

Transition Layer 模块: 位于两个 Denseblock 之间，起到连接和压缩的作用，由 1×1 卷积层，批标准化层和 2×2 平均池化层组成。

Classification 模块: 输出分类器采用全连接层分类策略，由批标准化层，平均池化层和线性全连接层组成。

4.3.4 相关运算

（1）1×1 卷积

MSFNet 模型在四个分支中都应用了 1×1 卷积，该卷积有以下作用：

（a）改变卷积核通道数的维度，灵活控制特征图深度。

（b）降低网络参数量和计算量。图 4-9 展示了使用 1×1 卷积对计算量的影响。

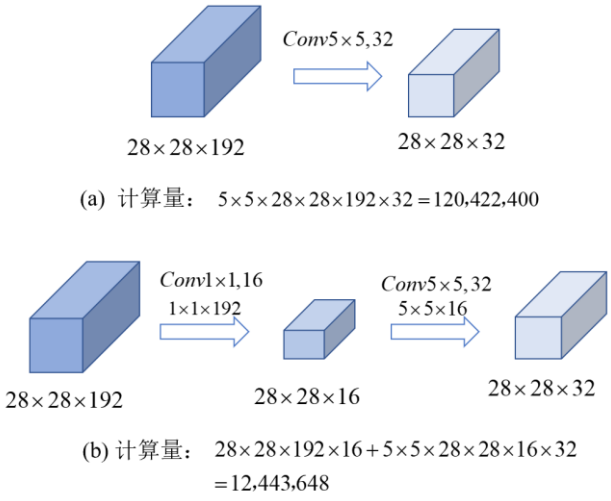


图 4-9 1×1 卷积作用示意图

Fig. 4-9 Schematic diagram of 1×1 convolution

（c）加强各特征图之间信息的交互和融合，即采用 1×1 卷积核将各通道的信息进行各类线性组合，有效加深网络，增加网络非线性特性。

（2）批量归一化（Batch Normalization，BN）

网络在训练过程中，各层参数都需要不断地更新迭代，为了避免各层网络数据在进行更新时分布发生较大改变导致无法拟合原始数据，在网络中引入批量归一化（Batch Normalization，BN）层，对数据进行标准化操作，使其符合正态分布。批量归一化的具体实现过程如下：

（a）计算样本均值，如公式（4-6）所示：

$$\mu_{\mathrm{B}}=\frac{1}{m}\sum_{i=1}^{m}x_i \tag{4-6}$$

其中，$\mathrm{B}=\{x_{1\ldots m}\}$ 为输入批量样本数据，x 为某个样本，m 为样本数量。

（b）计算样本方差，如公式（4-7）所示：

$$\sigma_{\mathrm{B}}^2=\frac{1}{m}\sum_{i=1}^{m}(x_i-\mu_{\mathrm{B}})^2 \tag{4-7}$$

（c）使数据分布遵循标准正态分布，如公式（4-8）所示：

$$\hat{x}_i=\frac{x_i-\mu_{\mathrm{B}}}{\sqrt{\sigma_{\mathrm{B}}^2+\varepsilon}} \tag{4-8}$$

（d）通过尺度缩放和偏移操作，保证恒等变换，如公式（4-9）所示：

$$y_i=\gamma\hat{x}_i+\beta\equiv BN_{\gamma,\beta}(x_i) \tag{4-9}$$

其中，参数 γ 表示输入数据的方差，β 表示输入数据的偏移量，加入批量归一

化层后，这两个参数只与当前层的信息有关。

（3）ReLU 激活函数

对于 CNN 而言，在缺少激活函数的情况下，网络输出只能是输入的线性组合，此时网络表征数据的能力有限，为了增加网络的非线性，本章提出的网络中加入了 ReLU 激活函数。

如图4-10所示，ReLU[66]的梯度只可以取两个值：0或1，当输入小于0时，梯度为 0；当输入大于 0 时，梯度为 1。优点在于 ReLU 函数的梯度连乘不会收敛到 0 ，连乘的结果也只可以取 0 或 1 ，如果值为 1 ，梯度保持值不变进行前向传播；如果值为 0 ，梯度从该位置停止前向传播。ReLU 函数数学表示为：

$$f(y)=\max(0,y) \tag{4-10}$$

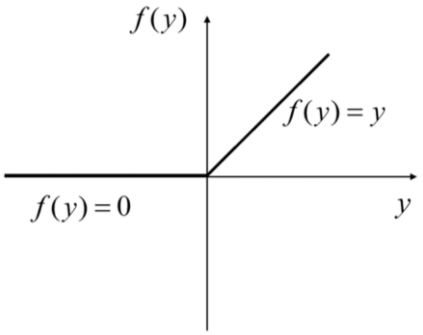


图 4-10 ReLU 函数

Fig. 4-10 ReLU function

4.4 模型训练

4.4.1 环境设置

网络训练和测试的环境配置如表 4-1 所示：

表 4-1 实验环境配置

Tab. 4-1 Experimental environment configuration

实验环境		配置说明
硬件环境	CPU	Intel(R)Core(TM)i5-3470@3.2GHz
	GPU	NVIDIA GeForce RTX 2080，8GB
软件环境	操作系统	Ubuntu 18.04
	Python	3.6
	PyTorch	1.4

4.4.2 参数设置

网络训练的各参数设置如下：

（1）构建的学生自发人脸表情数据集 LE-FER 中训练集和测试集的比例设置为 4：1；

（2）网络训练时，epoch 设置为 350，batch size 设置为 32；

（3）在不同数据集上进行对照实验时，输入图像大小和表情类别根据数据集而定；

（4）初始学习率设置为 0.01，从第 50 个 epoch 开始，每 5 个 epoch 的学习率降为上一个 epoch 的 90%；

（5）激活函数为 ReLU，优化器为 SGD。

4.4.3 数据集

为了提高模型的认可度，检验模型的鲁棒性，本章模型在 LE-FER 数据集上训练和测试的同时，也在目前表情识别领域各项国际比赛认可的开源数据集上进行了训练和验证。用到了如下数据集：

（1）Real-world Affective Faces Database (RAF-DB)

RAF-DB[67]数据集包含约 30000 张从网络上获取的人脸图像，分为两个子集，分别包含 7 类基本表情和 12 类复合表情。同时，该数据集还对年龄范围、性别等信息进行了标注，示例图如图 4-11 所示。

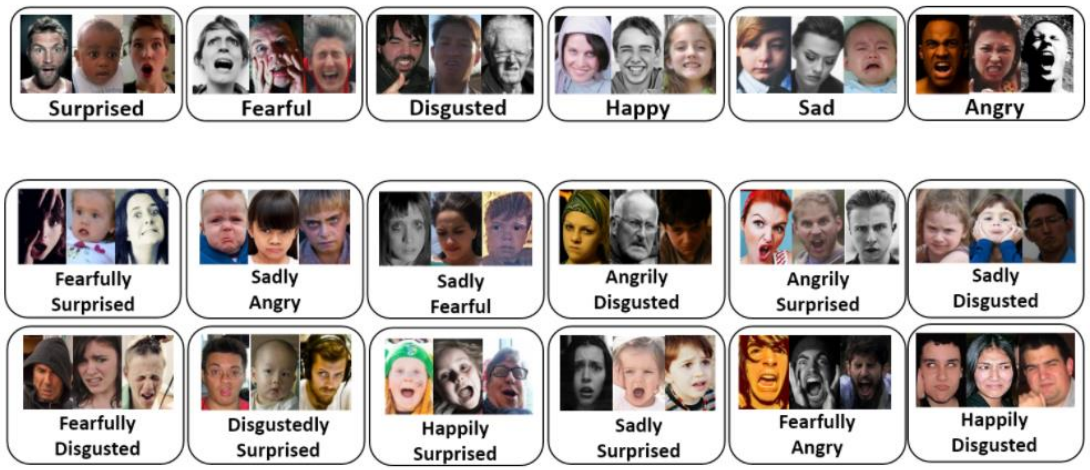


图 4-11 RAF-DB 数据集示例图

Fig. 4-11 Sample diagram of the RAF-DB dataset

（2）CK+

CK+[68]数据集内容为视频帧序列，包含 123 个对象，593 段视频帧序列，每段视频帧序列的最后一帧被标注了面部运动单元标签，其中有 327 段视频帧序列包含了表情标签，示例图如图 4-12 所示。

图 4-12 CK+数据集示例图

Fig. 4-12 Sample diagram of the CK+ dataset

（3）FER2013 和 FERPlus

FER2013[69]为 Kaggle 在 2013 年所举办的人脸表情识别竞赛数据集，示例图如图 4-13 所示，共包含 35887 张人脸图像。表情类别涵盖了快乐、惊讶、愤怒、厌恶、恐惧、悲伤、中性七类基本表情。

FER2013 数据集标签由人工标注得到，由于不同标注人员存在主观性差异，标签准确率较低，通过众包方法对其进行重新标注，同时新加入了“蔑视”、“未知”和“非人脸”类别，得到 FERPlus 数据集。

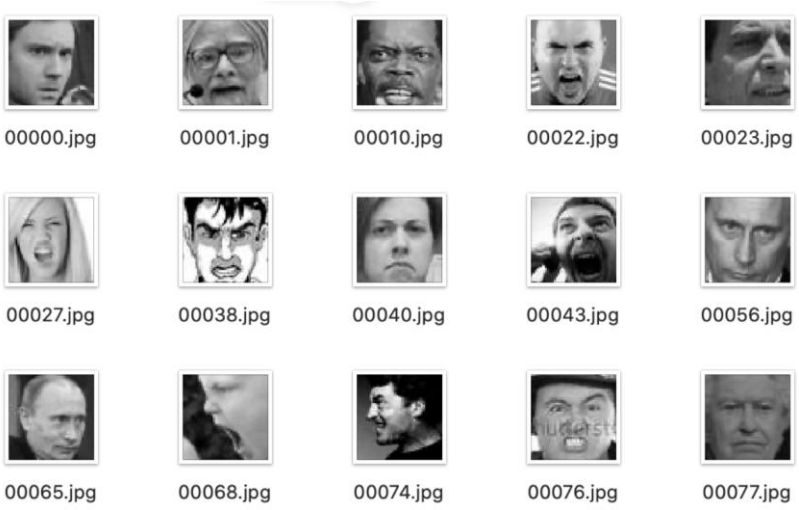


图 4-13 FER2013 数据集示例图

Fig. 4-13 Sample diagram of the FER2013 dataset

4.5 实验结果及分析

（1）深度可分离卷积有效性实验

本章研究主要考虑解决模型参数量过大导致难以实际应用部署以及现有模型泛化性能有待提高的问题。首先借鉴 Densenet 网络的思想，在表情识别任务中引入密集连接网络，在降低参数量的同时，提升泛化性能，保证较高的识别精度。面向在线教育应用的表情识别技术，在实际应用部署时，需要同时处理

大批量视频数据，因此需要进一步降低参数量来控制硬件成本，鉴于此在网络中引入深度可分离卷积。Model1 将深度可分离卷积加入到 BottleneckLayer 中，取代原 Densenet 结构中的普通卷积，卷积核设为 1×1，步长设为 1，实验结果如表 4-2 所示。

表 4-2 RAF-DB 数据集上深度可分离卷积有效性实验

Tab. 4-2 Validity experiment of depth-separable convolution on RAF-DB dataset

模型	参数量（M）	准确率（%）
Densenet	0.34	81.26
Model1	**0.22**	**81.32**

由表4-2可知，加入深度可分离卷积后，准确率虽只是略有提升，但参数量降低了很多，这显示了深度可分离卷积在表情识别任务上的有效性。采用密集连接的 CNN 网络通过特征重用，能够在保证识别精度的前提下，大幅度降低模型参数量，但由于密集连接卷积层之间的运算比普通卷积更为复杂，因此需要更多的浮点计算。为了降低模型计算量，利用深度可分离卷积来简化运算。

（2）多尺度卷积有效性实验

原深度卷积一般采用 3×3 卷积，网络在进行特征提取时，感受野有限，给模型后续的识别分类带来负面影响。为了提升卷积操作所提取特征的丰富性，考虑以不同尺度的卷积核代替原始的单一尺度卷积核。Model2 在 Model1 的基础上，改变了卷积核的组合，其中不同卷积核占比数量默认为平均分配。实验结果如表 4-3 所示。

表 4-3 RAF-DB 数据集上多尺度卷积有效性实验

Tab. 4-3 Effectiveness experiments of multi-scale convolution on RAF-DB dataset

模型 l	卷积核组合	参数量（M）	准确率（%）
Model1	3×3	0.22	81.32
Model2-1	3×3，5×5	0.22	81.42
Model2-2	1×1，3×3，5×5	0.22	81.48
Model2-3	1×1，3×3，5×5，7×7	0.24	81.36
Model2-4	**1×1，3×3，5×5，max-pooling**	**0.22**	**81.52**

由表 4-3 可知，Model2-4 表现最优。Model2-4 采用了三种不同尺度的卷积核，同时加入了最大池化层，这是因为最大池化层能减小神经网络各层由于参数误差导致的均值偏差，从而保留更多的人脸纹理信息，适用于表情识别这类对细节要求较高的分类任务。

（3）分组卷积有效性实验

引入深度可分离卷积后，模型的参数量和计算量均有所降低，但其中的深度卷积只收集了每个通道的特征，通道间信息交互不够会在一定程度上影响模型精度。为了增强通道间的交互性，同时避免参数量的大幅增加，考虑在深度卷积前加入分组卷积，Model3 在 Model2-4 的基础上，在四个分支中依次加入分

组卷积。实验结果如表 4-4 所示。

表 4-4 RAF-DB 数据集上分组卷积有效性实验

Tab. 4-4 Effectiveness experiments of group convolution on RAF-DB dataset

模型	参数量（M）	准确率（%）
Model2-4	0.22	81.52
Model3	**0.25**	**83.05**

从表 4-4 可以看出，加入分组卷积，使通道间能进行信息交互，避免丢失有用信息，准确率也因此得到了较大提升。事实上，输入数据依次通过了 1×1 卷积，分组卷积和多尺度深度可分离卷积，这一顺序构成了“渐进式轻量级”结构，实现通道间信息交互的逐渐减弱，在压缩模型体积的同时，尽可能保留更多的有效信息。

（4）SE 有效性实验

识别人脸表情需要关注人脸面部特定区域的特征，在模型中引入注意力机制，有助于特定信息的高效传播。Model4 在 Model3 的基础上加入了通道注意力模块 SE，并改变 SE 位置进行了对比实验。其中前三个模型将 SE 模块加在分支外不同位置，Model4-4 在每个分支中分别加入 1×1 卷积层和 SE 模块，实验结果如表 4-5 所示。

表 4-5 模型不同位置加入 SE 模块在 RAF-DB 数据集上的对比实验结果

Tab. 4-5 Comparative experimental results of adding SE modules in different positions of the model on RAF-DB dataset

模型	SE 位置	参数量（M）	准确率（%）
Model3	-	0.25	83.05
Model4-1	Transitionlayer 中平均池化层后	0.30	82.98
Model4-2	Pointwise 前	0.28	83.30
Model4-3	Pointwise 后	0.28	83.41
Model4-4	**四分支 1×1 卷积后**	**0.33**	**84.58**

由表可知，Model4-4 表现最优。这是因为 SE 模块的作用是给各通道分配各自的重要性权重，输入图像在经过 1×1 卷积层降维后通过 SE 模块时通道数较多，效果最好。

（5）与其他模型对比实验

Model4-4 为本文提出的最终模型，将其命名为 MSFNet（Multi-scale feature net），将本文提出的模型与目前表情识别任务中准确率较高的网络模型[70]以及目前主流的轻量级模型[71-73]进行对比实验，表 4-6 和表 4-7 展示了对比结果。

表 4-6 RAF-DB 数据集上 MSFNet 与高精度模型对比实验

Tab. 4-6 Comparison experiment between MSFNet and high-precision model on RAF-DB dataset

模型	参数量（M）	准确率（%）
VGG16[70]	138	88.98
Resnet18	11.17	87.78
MSFNet	**0.33**	**84.58**

表 4-7 RAF-DB 数据集上 MSFNet 与轻量级模型对比实验

Tab. 4-7 Comparison experiment between MSFNet and lightweight model on RAF-DB dataset

模型	参数量（M）	准确率（%）
Mobilenetv2[71]	3.88	80.66
Shufflenet[72]	1.86	81.36
Efficientnetv1-b3[73]	1.62	81.29
Densenet	0.34	81.26
MSFNet	**0.33**	**84.58**

由表4-6可知，对比准确率较高的模型，本章提出的模型虽然准确率略有下降，但参数量对比 VGG16 下降了 99.76%，对比 Resnet18 下降了 97.05%，更有利于模型的实际应用部署。由表 4-7 可知，对比目前主流的轻量级模型，MSFNet 与 Mobilenet，Shufflenet，Densenet 相比，参数量分别下降了 91.49%，82.26%，2.94%，准确率分别提高了 3.92%，3.22%，3.32%。对比实验结果说明，MSFNet 能在压缩模型的同时，达到较高准确率。

图 4-14 直观展示了 MSFNet 与各轻量级模型准确率及参数量对比关系。

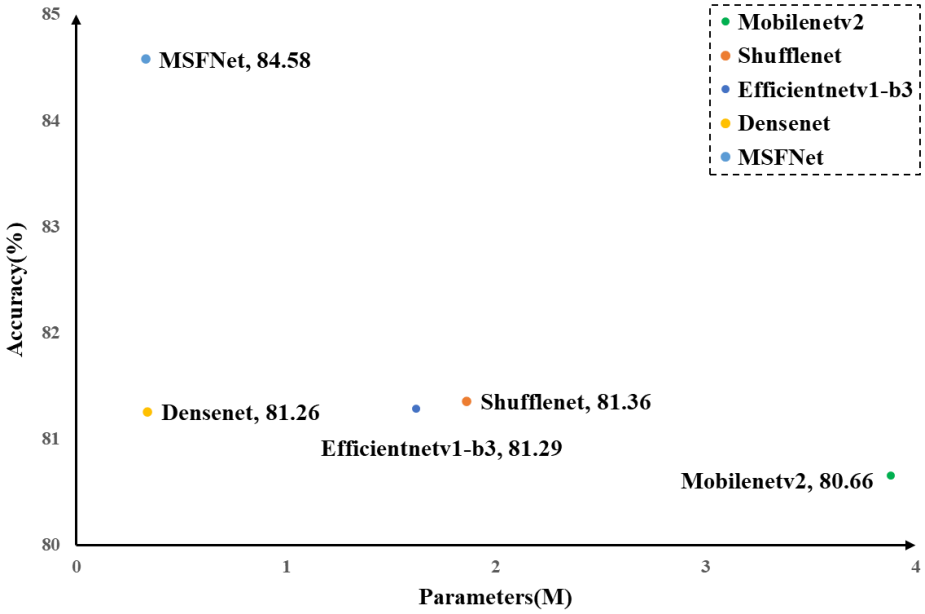


图 4-14 MSFNet 与轻量级模型在 RAF-DB 数据集上的对比实验

Fig. 4-14 Comparative experiments of MSFNet and lightweight model on RAF-DB dataset

（6）模型泛化性实验

为了测试模型泛化性能，将 MSFNet 在 FER2013，FERPlus，CK+数据集上进行了进一步验证。实验结果如表 4-8 所示[74]。

表 4-8 MSFNet 在不同数据集上的泛化性实验

Tab. 4-8 Generalization experiments of MSFNet on different datasets

数据集	模型	参数量（M）	准确率(%)
CK+	Mobilenetv2	3.88	92.21
	Densenet	0.34	87.37
	MSFNet	**0.33**	**90.10**
FER2013	Resnet18	11.17	71.38
	Efficientnetv2[75]	20.19	65.09
	Inception	0.39	71.60
	MSFNet	**0.33**	**71.89**
FERPlus	VGG19	20.04	86.50
	Resnet18	11,17	83.35
	Efficientnetv1-b3	1.62	79.80
	MSFNet	**0.33**	**82.67**

图 4-15 为不同数据集上各模型准确率散点图。

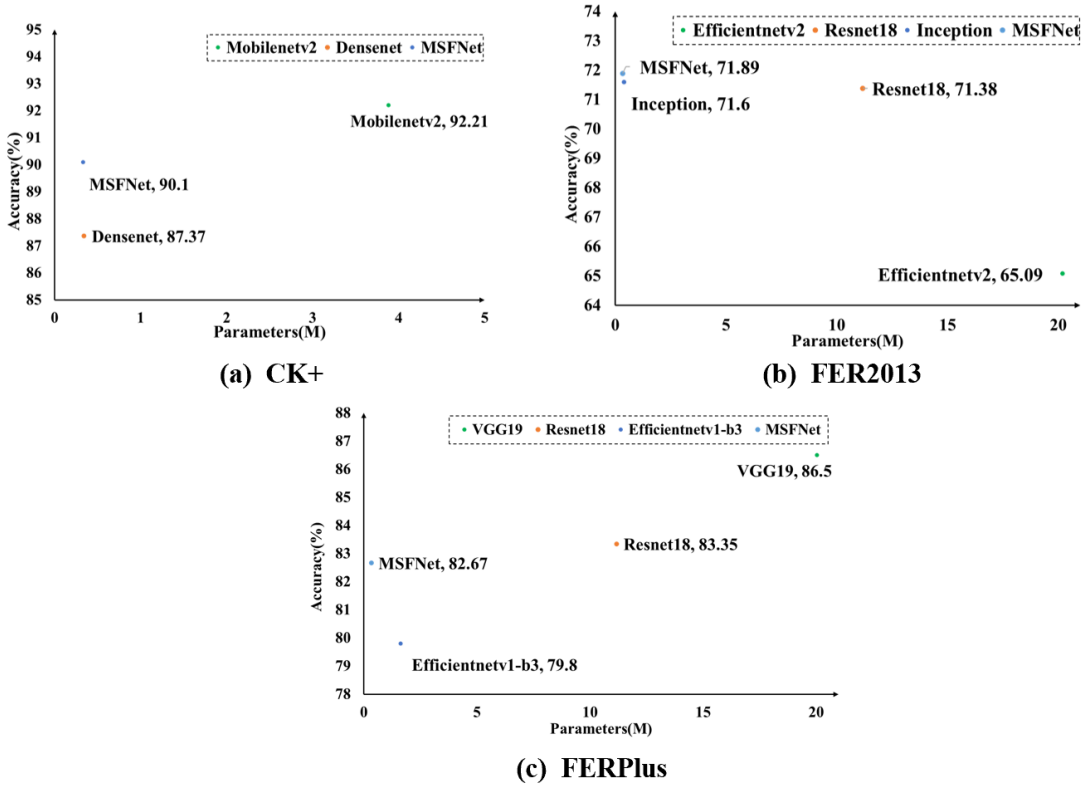


图 4-15 不同数据集上各模型准确率对比散点图

Fig. 4-15 Scatter diagram of comparison of accuracy rates of different models on different datasets

由图 4-15 和表 4-8 可知，在 CK+数据集上，MSFNet 准确率仅次于 Mobilenet，但 MSFNet 参数量仅占 Mobilenet 的 8.5%；在 FER2013 数据集上，MSFNet 准确率高于其他轻量级模型的同时，还高于参数量远大于其参数量的 Rsenet18；在 FERPlus 数据集上，MSFNet 准确率也仅次于大体积模型 VGG19 和 Resnet18。

由此说明，本章提出的轻量级模型 MSFNet 在保证较高准确率的同时，压缩了模型大小，降低了模型参数量，同时具有较好的泛化性能。

（7）LE-FER 数据集对比实验

在 LE-FER 数据集上，将 MSFNet 与目前性能较优的表情识别模型进行对比

实验，实验结果如表 4-9 所示。

表 4-9 LE-FER 数据集对比实验

Tab. 4-9 Comparative experiment on LE-FER dataset

模型	参数量（M）	准确率（%）
Mobilenet	3.88	82.66
Shufflenet	1.86	84.36
Efficientnetv1-b3	1.62	85.14
Densenet	0.34	87.77
MSFNet	**0.33**	**89.12**

由实验结果可知，MSFNet 在基于学习环境构建的数据集 LE-FER 上表现优于其他模型，说明该模型对于学习背景适用性强，在实际应用部署时，能保证识别精度，同时降低硬件成本。

4.6 小结

近几年基于深度学习的表情识别模型在公共数据集上取得了不错的表现，但是为了提高识别精度，现有的方法通常将模型设计得复杂庞大，导致实际应用部署时对硬件成本要求过高。目前工业界追求的重点逐渐从准确率向速度和模型大小倾斜，实现模型的轻量级、降低系统在实际应用中的成本也是未来值得考虑的问题。在线教育情境下，利用表情识别分析学生学习情绪时，学生人数多，课程时间长，如何把握识别精度和硬件成本的平衡直接决定了最终能否投入实际应用。

为此，本章提出了一种基于密集连接思想，融合多尺度特征和注意力机制的表情识别模型 MSFNet，该模型的主要优势在于：

（1）引入了密集连接卷积神经网络，通过特征复用充分利用原始有效信息，且舍弃全连接层有效地降低了网络参数量，提升了网络的泛化性能。

（2）提出了“渐进式轻量级”思想，模型在进行特征学习的过程中，输入图像依次通过1×1 普通卷积，分组卷积和深度可分离卷积，实现通道间信息交互量的逐渐降低，在压缩模型和保留有效信息间取得了较好的平衡。

（3）提出了多尺度深度可分离卷积网络。深度可分离卷积本身能极大程度地降低网络的计算量和参数量，同时，将不同尺度的卷积核进行合理组合，网络在进行特征学习时，更加丰富的感受野能提取到更多尺度的特征，有效地提高了模型的分类性能。

（4）在模型中引入了通道注意力机制，并深入探究了多尺度模块与注意力机制的结合方法，证实了两者结合能进一步提高表情识别精度。

将本章提出的模型分别在 LE-FER，CK+，RAF-DB，FER，FERPlus 数据

集上与其他模型进行比较，实验结果表明，MSFNet 能在压缩模型的同时保证较高的识别精度，这对于在线教育需要处理大规模视频数据而言，能有效降低硬件成本。同时，模型在不同环境下构建的数据集上都取得了较优的结果，具有良好的泛化性能。

第 5 章 基于多层级特征结合判别机制的快速轻量级人脸表情识别算法研究

5.1 引言

线下教育由于地点固定，学生数量也随之有限，教师可以根据管理要求限制上课人数，来保证良好的课堂秩序。在线教育不受上课地点限制，可能出现大规模学生同时上课的情况，教师要及时监控到每个学生的学习状态，除了对模型参数量，计算量有要求外，对模型的推理速度也有较高要求。

针对面向在线教育应用时，需要实时分析大规模学生学习情绪的要求，本章在沿用第 4 章轻量化思想的基础上，对加快模型推理速度进行了深入研究，构造了基于多层级特征结合判别机制的快速轻量级人脸表情识别模型 Multi-ghostnet。

5.2 建模分析

5.2.1 Ghostnet

目前对于 CNN 进行轻量级的改进，往往会用到分组卷积，通道混合等机制，来实现使用较少的浮点运算构建有效的网络结构。但如图 5-1 所示，网络结构中 1×1 卷积层依旧会消耗大量内存和计算量。

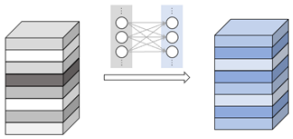


图 5-1 普通1×1 卷积

Fig. 5-1 Ordinary 1×1 convolution

事实上，深度学习网络在进行特征学习的过程中，为了保证学习到的信息足够充分，往往会产生大量冗余的特征图。图 5-2 为 Resnet50 经过第一个残差块后得到的特征图可视化结果，相似特征图用相同颜色框标注。Ghostnet[75]中指出，这些相似特征图之间可以进行信息共享，将原始特征图通过简单线性变换得到相似特征图，这一过程比普通卷积节省了大量的内存和计算量。

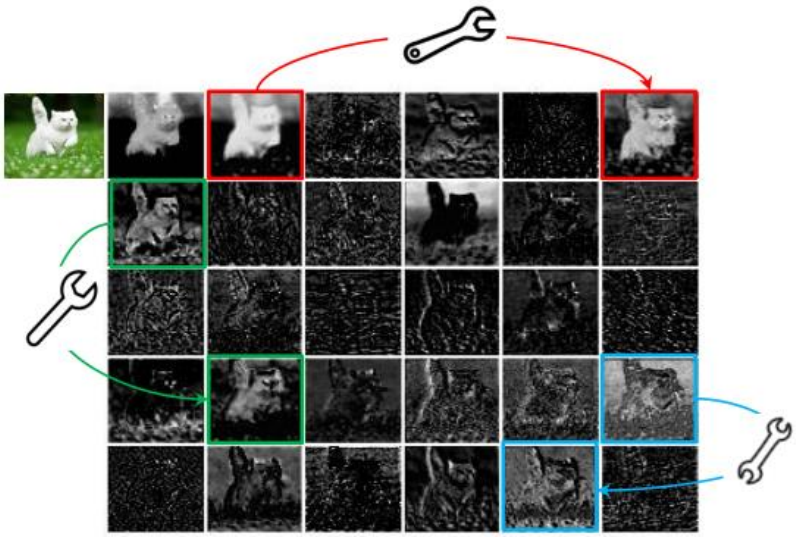


图 5-2 ResNet-50 残差块特征图可视化

Fig. 5-2 Visualization of resnet-50 residual block feature map

如图 5-3 Ghost Module 结构图所示，蓝色表示普通卷积，其生成的原始特征图具有较小的尺寸，绿色表示简单线性变换，得到相似特征图。

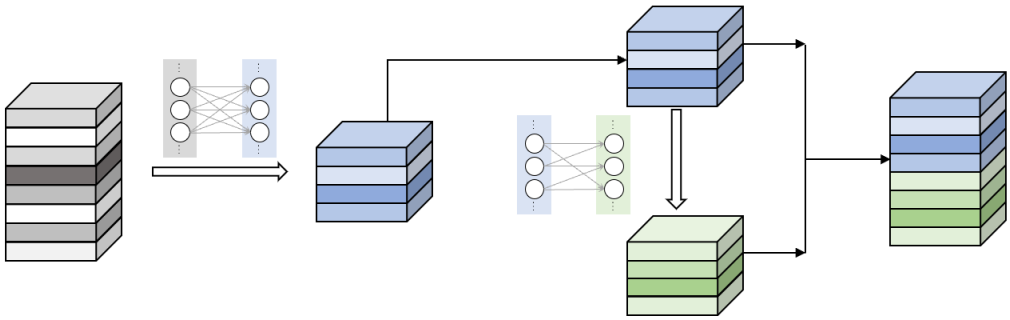


图 5-3 Ghost Module 结构图

Fig. 5-3 Ghost Module structure diagram

基于 Ghost Module 提出了 Ghost bottleneck，适用于小型 CNN 网络，如图 5-4 所示，Ghost bottleneck 主要由前后两个 Ghost 模块组成，分别用来增加和减少通道数，并在两个 Ghost 模块之间加入了残差连接。

基于 Ghost bottleneck 进一步提出了 GhostNet，如图 5-5 所示，GhostNet 堆叠了若干 Ghost bottleneck，其中，部分 Ghost bottleneck 之间加入了 SE 模块。

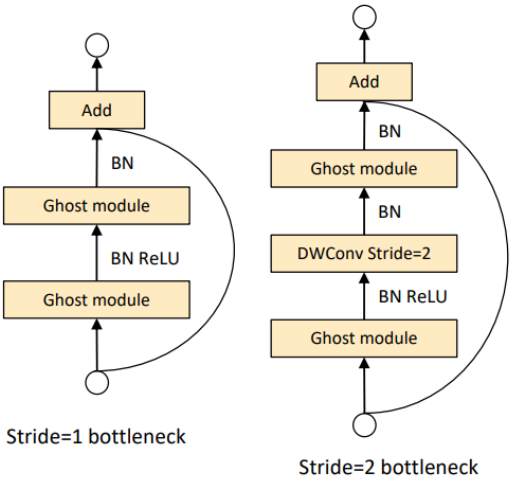


图 5-4 Ghost Bottleneck 结构图

Fig. 5-4 Ghost Bottleneck structure diagram

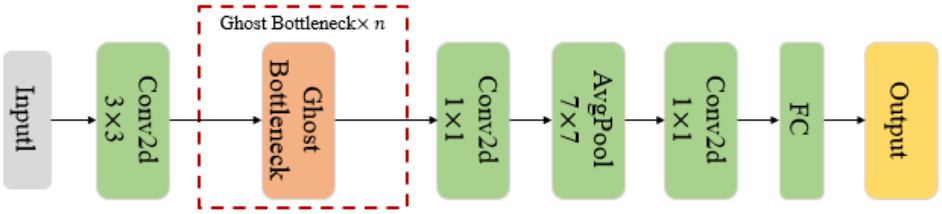


图 5-5 Ghostnet 结构图

Fig. 5-5 Ghostnet structure diagram

这种相似特征图思想可以很大程度上节省网络计算量，加快推理速度。因此，本章考虑将该思想运用到面向在线教育应用的人脸表情识别研究中。需要注意的是， Ghost Module 只将特征图进行二等分，并未继续对该思想进行深入探索，针对特征图的重要性只进行了两个等级的划分，并未充分挖掘出特征图之间相似信息共享的潜力。

5.3 模型设计

5.3.1 Multi-ghost Moudle

本章提出了一种更加具有普适性的网络结构 Multi-ghost Moudle，通过多个层级的 ghost 特征图划分，在特征图之间实现更加细致的变换操作。

Multi-ghost Moudle 的设计如图 5-6 所示，将特征图分为 4 个等级。由于第一组输出的特征图要作为后续 3 层相似特征图的基础，故其用普通卷积实现以保证模型性能，在图中表现为蓝色卷积块。另外 3 种相似特征图用较为简单的线性变换，以减少模型的参数量和计算量。利用这种结构，使相似特征图之间的线性变换依次叠加，在极大程度上加深网络，有效减少因网络中包含过多的

简单变换而造成的性能下降。同时，不同级别的特征图在最后进行统一融合，适用于人脸表情识别这种对细粒度要求较高的分类任务。

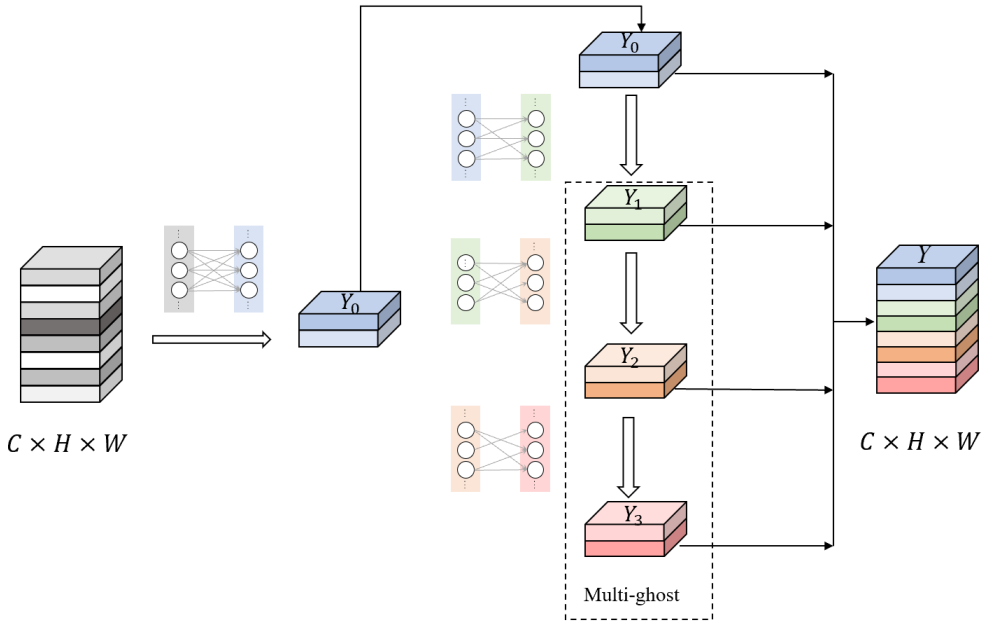


图 5-6 Multi-ghost Moudle 结构图

Fig. 5-6 Multi-ghost Moudle structure diagram

具体来说，首先使用一次标准卷积生成l_0个原始特征图$Y_0 \in \mathbb{R}^{h' \times w' \times n}$：

$$Y_0 = X * f' \tag{5-1}$$

其中 $f' \in \mathbb{R}^{c \times k \times k \times m}$ 是使用的卷积核，$l_0 < n$。然后对 Y_0 中的原始特征进行一系列简单线性变换，得到Y_1个相似特征图，线性运算的数学表达如公式（5-2）所示：

$$y_{i,j} = \Phi_{i,j}\left(y_i^1\right), \forall i = 1,\ldots,l_0, j = 1,\ldots,q_1 \tag{5-2}$$

其中 y_i^1 是 Y_1 中第 i 个原始特征图，上述函数中的 $\Phi_{i,j}$ 是第 j 个线性运算，用于生成第 j 个相似特征图 y_{ij}，也就是说，y_i^1可以包含若干个相似特征图$\left\{y_{ij}\right\}_{j=1}^{q_1}$。通过使用简单线性变换，得到$l_1 = l_0 \cdot q_1$个特征图$Y_1 = \left[y_{11}, y_{12}, \ldots, y_{l_0 q_1}\right]$作为$l_1$的输出。

上述推理过程总共重复三次，每次都以上一层级的输出作为下一层级的输入，最终得到$n = l_0 + l_1 + l_2 + l_3$ 个特征图$Y = \left[Y_0, Y_1, Y_2, Y_3\right]$ 作为 Multi-ghost Moudle 的输出。基于第 4 章的研究，本章对于 Multi-ghost Moudle 中取代原始卷积的简单线性变换组合的选择，沿用了第 4 章多尺度深度可分离卷积思想，通过实验探究，最终组合为：1×1，5×5，3×3，maxpooling。

5.3.2 Ghost Selection Moudle

本模型在两个 Multi-ghost 模块中间加入了一个 Ghost Selection 模块，该模块具体结构如图 5-7 所示。将第一个 Multi-ghost 模块的高维度输出进行融合，由此得到各层级相似特征图之间的潜在联系。随后根据融合后的特征图分别对各层级相似特征图进行对应重要性的学习，因此得到每个相似特征图对于全体特征图的重要性判断。最后将得到的各重要性施加在对应特征图上，并与输入进行相加。该鉴别机制可以针对输入特征图的改变动态调整各相似特征图的重要性，由此可以显著提升整个 Multi-ghost Block 的性能。

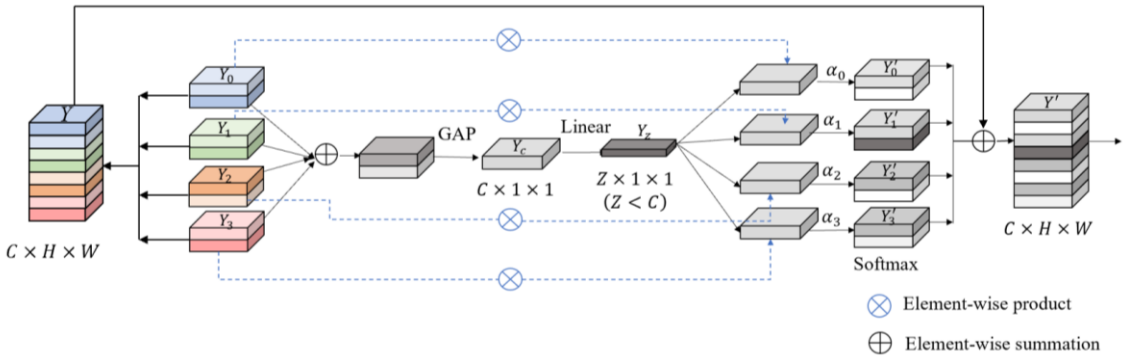


图 5-7 Ghost Selection 结构图

Fig. 5-7 Ghost Selection structure diagram

输入图像经过第一个 Multi-ghost Moudle 后，得到四个层级不同尺度的输出特征图。

首先计算每个层级卷积核的重要性权重。将四层级的特征图按元素求和，如公式（5-3）所示：

$$Y = Y_0 + Y_1 + Y_2 + Y_3 \tag{5-3}$$

特征图Y的维度为维度为$C \times H \times W$，其中 H 是高度（Height），W 是宽度（width），C 是通道数（channel），可以将其看作 C 张 H×W 的图像，如公式（5-4）所示，通过全局平均池化（Global Average Pooling），每张图像用其像素的均值表示，得到 C×1×1 大小的向量。

$$Y_c = \mathcal{F}_{gp}(Y) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} Y(i, j) \tag{5-4}$$

Y_c通过全连接层得到紧凑特征图Y_z：

$$Y_z = \mathcal{F}_{fc}(Y_c) = \delta\left(\mathcal{F}_{bn}(W_c)\right) \tag{5-5}$$

$$W_c = d \times C \tag{5-6}$$

$$d = \max\left(\frac{C}{r}, L\right) \tag{5-7}$$

其中，δ 为 Relu 激活函数，$\mathcal{F}_{bn}$ 表示批量归一化（BN），W_c 为 Y_c 的维度，d 表示全连接后的特征维度，L 为通道最小值，设置为 32，r 为压缩因子。

通过 softmax 运算给不同层级特征图分配权重 $\alpha_n\left(n \in [0,3]\right)$，$n$ 表示各个层级，然后与原始四层级特征图对应进行相乘运算，得到新的特征图 $Y_n^{'}$，将四层级特征图通过 concatenate 连接后，与原始特征图 Y 进行元素级相加运算，得到最终特征图 Y'，维度与原始特征图保持一致。

5.3.3 MG-Block

进一步，搭建 MG-Block，如图 5-8 所示，MG-Block 主要由 2 个 Multi-ghost Module 和 1 个 Ghost Selection Module 组成，前后两个 Multi-ghost Module 分别用来增加和减少通道数，并在两个 Multi-ghost Module 之间加入了残差连接。

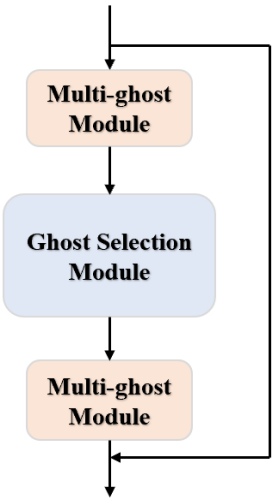


图 5-8 MG-Block 结构图

Fig. 5-8 MG-Block structure diagram

中间的 Ghost Selection 模块在通道扩张到较大维度的情况下，对多层级的相似特征图进行动态鉴别，以在每个 MG-Block 中实现特征的动态融合。

5.3.4 Multi-ghostnet

Multi-ghostnet 如图 5-9 所示：

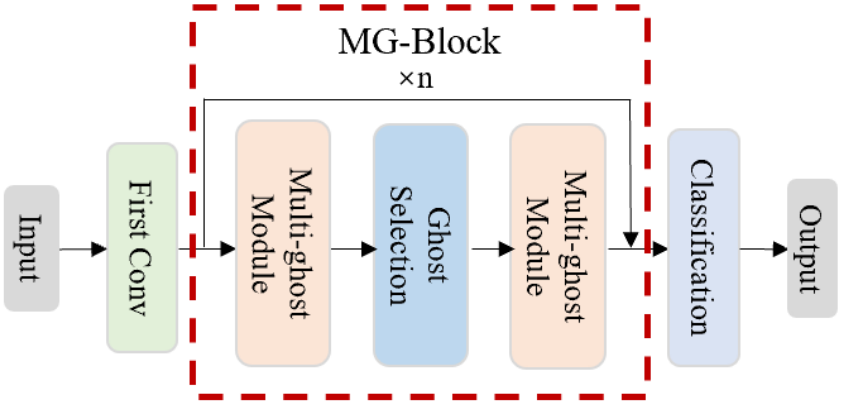


图 5-9 Multi-ghostnet 结构图

Fig. 5-9 Multi-ghostnet structure diagram

基于 MG-Block，将 Ghostnet 中的 Ghost-Bottleneck 替换为 MG-Block，由于 MG-Block 中包含 Ghost Selection Module，可以实现对特征图重要性的动态鉴别，因此不再保留原 Ghostnet 中的 SE 模块，进一步降低参数量和计算量。

5.3.5 相关运算

（1）全局平均池化（Global Average Pooling）

全局平均池化[76]（GAP）在 2013 年被首次提出，随即风靡各种卷积神经网络。

一般情况下，卷积层用于提取二维数据如图片、视频等的特征，后续针对于具体任务（分类、回归、图像分割等），卷积层会用到不同类型的网络。拿分类问题举例，最简单的方式就是将卷积网络提取出的特征（特征图）输入到 softmax 全连接层，最终对应不同的类别。这里的特征图是二维多通道的数据结构，具有空间上的信息。如图 5-10 所示，在 GAP 被提出之前，常用的方式是将特征图直接拉平成一维向量（图 a），但是 GAP 不同，是将每个通道的二维图像做平均，最后每个通道对应一个均值（图 b）。

可以看到，GAP 的设计简单直接，具备以下优点：

（a）抑制过拟合，降低参数量。直接拉平做全连接层的方式依然保留了大量的空间信息，假设特征图是 32 个通道的 10×10 图像，那么拉平就得到了 32×10×10 的向量，如果最后一层对应两类标签，那么这一层就需要 3200×2 的权重矩阵。而 GAP 不同，将空间上的信息直接用均值代替，32 个通道 GAP 之后得到的向量都是 32×1×1 尺寸的向量，那么最后一层只需要 32×2 的权重矩阵。相比之下 GAP 网络参数会更少，而全连接更容易在大量保留的空间信息上产生过拟合现象。

（b）输入尺寸更加灵活。特征图经过 GAP 处理后的神经网络，参数不再与输入图像的尺寸有关，也就是输入图像的长宽可以不固定。

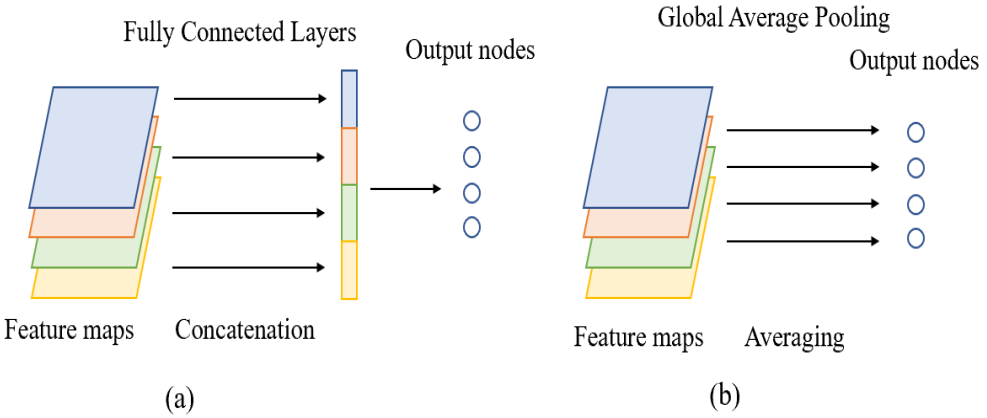


图 5-10 全连接层与全局平均池化层结构图

Fig. 5-10 Structure diagram of fully connected layer and global average pooling layer

（2）softmax 分类器

Softmax[77]分类器用来解决多分类问题，类标签 y 可以取 k 个不同的值。用假设函数分别计算输入 x 针对每一个可能的类别 j 的概率值 $p(y=j\,|\,x)$。假设函数 $h_\theta(x)$ 数学表示如下：

$$h_\theta\left(x^{(i)}\right)=\begin{bmatrix} p\left(y^{(i)}=1|x^{(i)};\theta\right) \\ p\left(y^{(i)}=2|x^{(i)};\theta\right) \\ \vdots \\ p\left(y^{(i)}=k|x^{(i)};\theta\right) \end{bmatrix}=\frac{1}{\sum_{j=1}^{k}e^{\theta_j^T x^{(i)}}}\begin{bmatrix} e^{\theta_1^T x^{(i)}} \\ e^{\theta_2^T x^{(i)}} \\ \vdots \\ e^{\theta_k^T x^{(i)}} \end{bmatrix} \tag{5-8}$$

概率 p 公式表示为：

$$p\left(y^{(i)}=j|x^{(i)};\theta\right)=\frac{e^{\theta_j^T x^{(i)}}}{\sum_{j=1}^{k}e^{\theta_i^T x^{(i)}}} \tag{5-9}$$

其中，$\theta_1,\theta_2,\ldots,\theta_k\in\mathcal{R}^{n+1}$ 是模型对应参数，$\frac{1}{\sum_{j=1}^{k}e^{\theta_j^T x^{(i)}}}$ 对概率分布进行归一化处理。

5.4 模型训练

5.4.1 环境配置

模型训练时的环境配置见表 4-1。

5.4.2 参数设置

网络训练的各参数设置如下：

（1）构建的学生自发人脸表情数据集 LE-FER 中训练集和测试集的比例设置为 4：1；

（2）网络训练时，epoch 设置为 350，batch size 设置为 32；

（3）在不同数据集上进行对照实验时，输入图像大小和表情类别根据数据集而定；

（4）初始学习率设置为 0.08，从第 50 个 epoch 开始，每 5 个 epoch 的学习率降为上一个 epoch 的 90%；

（5）激活函数为 ReLU，优化器为 SGD。

5.4.3 数据集

为保证模型的认可度，训练过程中，除了在自构建的学习情绪数据集 LE-FER 上进行了实验，同样也在目前表情识别领域常用的开源数据集 RAF-DB，CK+，FER2013，FERPlus 上进行了对比实验。

5.5 实验结果及分析

（1）学习率探究实验

在训练深度学习模型时，学习率是一个十分重要的参数，其对模型性能的影响主要取决于训练开始时初始学习率的大小和训练过程中学习率的变化策略。其中，对于初始学习率而言，过大会导致模型始终徘徊在最优值附近，无法收敛，过小会导致网络收敛得非常慢，增大找到最优值的时间，同时容易在局部极值点收敛，无法找到最优解。

本章首先采用原始 Ghostnet 网络进行最优初始学习率的探究实验。模型训练时，epoch 设置为 350，学习率变化策略设置为从第 50 个 epoch 开始，每隔 5 个 epoch，学习率衰减至上一个 epoch 学习率的 90%，实验结果如表 5-1 所示。

表 5-1 RAF-DB 数据集上初始学习率探究实验

Tab. 5-1 Experiment of initial learning rate on RAF-DB dataset

模型	初始学习率	准确率（%）
Model 1-1	0.12	80.70
Model 1-2	0.10	80.93
Model 1-3	**0.08**	**81.23**
Model 1-4	0.06	81.10
Model 1-5	0.04	80.77

由表 5-1 可知，初始学习率的设置会对后续模型的表现产生较大影响，最优初始学习率为 0.08，后续实验均将初始学习率设置为 0.08。

（2）多层级特征图有效性实验

一般卷积神经网络各层特征图其实都存在不同程度的冗余，原始 Ghostnet 首先通过普通卷积获得一层原始特征图，然后令该原始特征图通过一个 Ghost Module 得到一层相似特征图，相当于只对特征图的重要性进行了两个等级的划分。事实上，第一层相似特征图中依旧包含冗余信息，这种划分方式并未充分挖掘出特征图之间相似信息的潜力。因此本章进一步推进相似特征图思想，将特征图分为不同等级，层层递进，实验结果如表 5-2 所示。不同网络等级中，由于第一级输出的特征图要作为后续 3 级相似特征图的基础，故其用普通卷积实现，来保证该网络模块的信息量，而后续等级的相似特征图采用线性变换获得。为保证对比实验的公平性，此时的线性变换采用原始 Ghostnet 中 3×3 深度可分离卷积。

表 5-2RAF-DB 数据集上多层级特征图有效性实验

Tab. 5-2 Effectiveness experiment of multi-level 特征图 on RAF-DB dataset

模型	特征图层级数	准确率（%）	参数量（M）	Flops（M）	推理速度（ms）
Model 2-1	2	81.23	2.70	145.91	13.01
Model 2-2	**4**	**81.62**	**2.50**	**123.10**	**12.06**
Model 2-3	6	80.97	2.31	102.97	10.31

由表 5-2 可知，将特征图分为 4 个等级效果最好，利用这种结构，可以使得各级相似特征图之间的线性变换进行依次叠加，在极大程度上加深网络，有效减少因网络中包含过多的简单变换而造成的性能下降。

（3） 多尺度卷积核实验

基于第 4 章的研究可知，利用不同尺度的卷积核能使模型获取到多尺度的特征，提高信息的丰富性。同时，使用最大池化层能减小网络在训练过程中由于参数误差造成的均值偏移，保留更多的人脸纹理信息，适用于表情分类任务。因此，本章设计模型时，选择引入多尺度卷积核，并沿用了第 4 章得到的最优卷积核组合，即同时使用 1×1 卷积，3×3 卷积，5×5 卷积和最大池化层。

由于本模型使用到的多层特征图之间具有层层递进的关系，通过实验发现，

多尺度卷积核的使用顺序会对实验结果产生影响，实验结果如表 5-3 所示。

表 5-3 RAF-DB 数据集上多尺度卷积核顺序探究实验

Tab. 5-3 Multi-scale convolution kernel sequential exploration experiment on RAF-DB dataset

模型	卷积核排列	准确率（%）	参数量（M）	Flops（M）	推理速度（ms）
Model 3-1	1×1,3×3,5×5, MaxPooling	82.27	2.51	125.84	12.32
Model 3-2	1×1,Maxpooling, 5×5,3×3,	81.94	2.51	125.84	13.45
Model 3-3	**1×1,5×5,3×3 MaxPooling**	**82.37**	**2.51**	**125.84**	**15.17**

由表 5-3 可知，将各尺度卷积核按 1×1，5×5，3×3，MaxPooling 的顺序排列表现最优，按该排列方式搭建的网络模块即为本章提出的 Multi-ghost Module.

（4）MG-Block 探究实验

进一步，对 Module 3-3 进行改进，在相邻的两个 Multi-ghost Module 之间引入 Ghost Selection 模块，同时建立 Ghost Selection 模块上下两个 Multi-ghost Module 的残差连接。需要注意的是，Module3-3 还保留着 Ghostnet 原始设计中的若干 SE 模块，但由于 Ghost Selection 模块已经具备使模型动态识别特征图重要性的特点，因此不再需要 SE 模块，可以进一步降低模型参数量和计算量。实验结果如表 5-4 所示。

表 5-4 RAF-DB 数据集上 Ghost Selection 模块探究实验

Tab. 5-4 Ghost Selection module exploration experiment on RAF-DB dataset

模型	准确率（%）	参数量（M）	Flops（M）	推理速度（ms）
Model 3-3	82.37	2.51	125.84	15.17
Multi-ghostnet	**84.06**	**1.33**	**153.14**	**25.75**

由表 5-4 可知，使用 MG-Block 搭建 Multi-ghostnet，Flops 和推理速度虽然有所增加，但准确率显著升高，参数量缩减为原来的 53%，模型综合表现更优秀。模型推理速度虽提升到 25.75ms，但考虑到目前摄像头每秒传输帧数一般为 30 帧，而模型识别速度可以达到 38.83 帧/秒，依旧可以实现实时识别。

（5）Multi-ghostnet 与其他模型对比实验

将 Multi-ghostnet 与目前表情识别领域推理速度较快的模型 Resnet50、EfficientNetv1-b3、Efficientnetv2、Ghostnet 以及第 4 章提出的轻量级模型 MSFNet 进行对比实验，表 5-5 列出了参数量，计算量，推理速度多个维度的对比结果。

表 5-5 RAF-DB 数据集上模型多维度性能对比实验

Tab. 5-5 Comparison experiment of multi-dimensional performance of model on RAF-DB dataset

数据集	模型	准确率（%）	参数量（M）	Flops（M）	推理速度（ms）
RAF-DB	Resnet50	78.00	23.52	125.84	10.88
	EfficientNetv1-b3	81.29	1.62	27.59	32.51
	Efficientnetv2	77.70	20.19	2873.04	25.41
	Ghostnet	81.23	2.70	145.91	13.01
	MSFNet	84.58	0.33	688.62	121.04
	Multi-ghostnet	**84.06**	**1.33**	**153.14**	**25.75**

图 5-11 直观展示了各模型参数量，推理速度和准确率的关系。由图表结果可以看出，本章提出的模型 Multi-ghostnet 在准确率，参数量，计算量，推理速度上取得了一个平衡较优的结果，其综合性能优于其他对比模型。

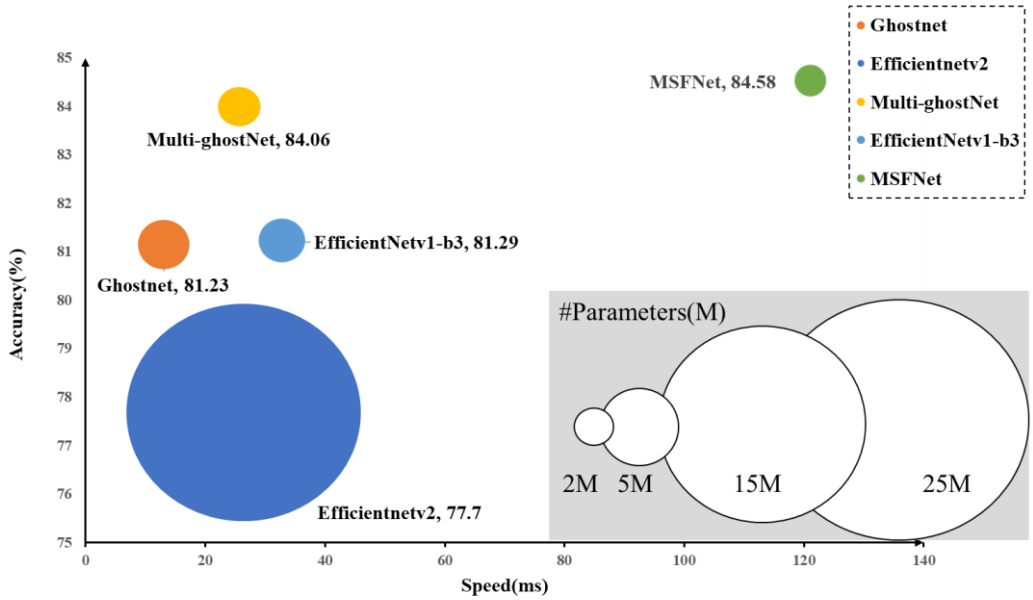


图 5-11 RAF-DB 数据集上模型多维度性能对比图

Fig. 5-11 Comparison diagram of multi-dimensional performance of model on RAF-DB dataset

（6）常用开源表情识别数据集对比实验。

为了验证模型的泛化性能，在其他常用开源表情识别数据集 CK+，FER2013，FERPlus 上进行了对比实验，实验结果分别如表 5-6，表 5-7，表 5-8 所示。

表 5-6 CK+数据集对比实验

Tab. 5-6 Comparative experiment on CK+ dataset

数据集	模型	准确率（%）
CK+	Resnet50	92.83
	EfficientNetv1-b3	94.55
	Efficientnetv2	97.27
	Ghostnet	94.04
	MSFNet	90.10
	Multi-ghostnet	**96.57**

表 5-7 FER2013 数据集对比实验

Tab. 5-7 Comparative experiment on FER2013 dataset

数据集	模型	准确率（%）
FER2013	Resnet50	73.08
	EfficientNetv1-b3	65.09
	Efficientnetv2	71.74
	Ghostnet	70.91
	MSFNet	71.89
	Multi-ghostnet	**71.13**

表 5-8 FERPlus 数据集对比实验

Tab. 5-8 Comparative experiment on FERPlus dataset

数据集	模型	准确率（%）
FERPlus	Resnet50	82.84
	EfficientNetv1-b3	79.80
	Efficientnetv2	83.45
	Ghostnet	82.45
	MSFNet	82.67
	Multi-ghostnet	**82.66**

图 5-12 直观展示了不同模型在各数据集上的准确率对比。分析对比结果可知，目前的表情识别模型中，有的精度高，推理速度快但参数量大，而参数量小的轻量级模型，即使能实现较低参数量和较高准确率，也无法保证推理速度。Multi-ghostnet 能同时实现高准确率，高推理速度，低参数量和低运算量，在压缩模型的同时保证模型综合性能，并且在基于不同场景构建的数据集上均表现较优，模型鲁棒性较好。

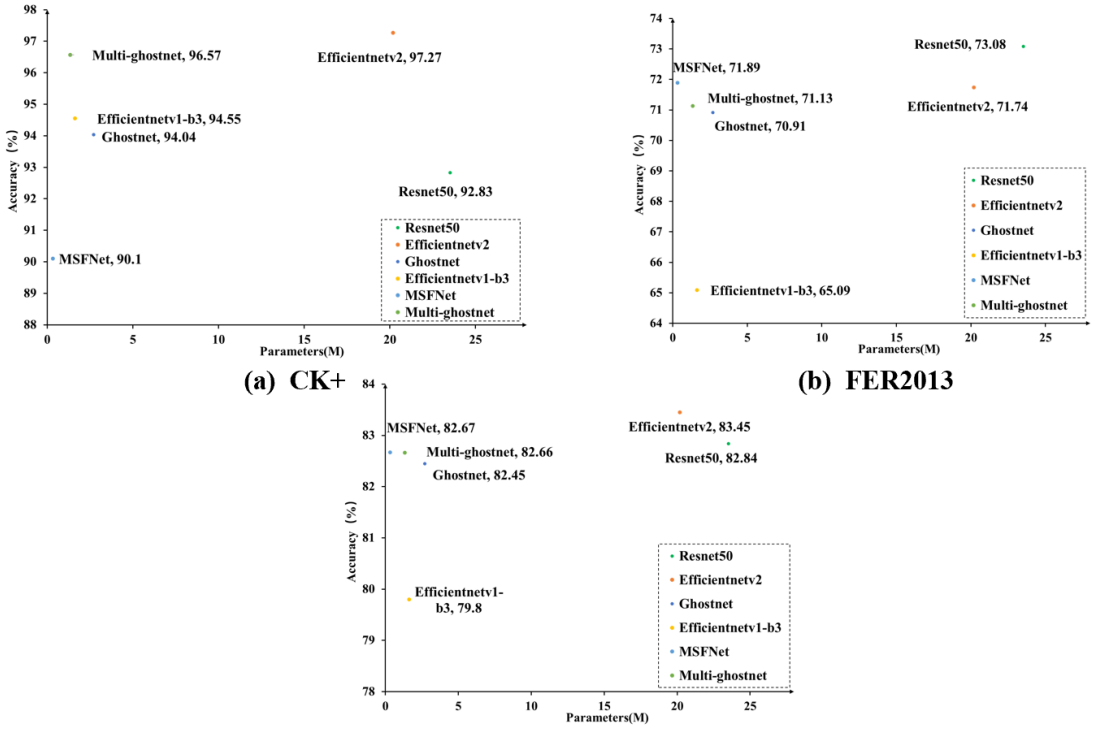


图 5-12 不同模型各数据集准确率对比图

Fig. 5-12 Comparison of accuracy of datasets of different models

（7）LE-FER 数据集对比实验

在 LE-FER 数据集上，将 Multi-ghostnet 与目前性能较优的表情识别模型进行了对比实验，实验结果如表 5-9 所示。

表 5-9 LE-FER 数据集对比实验

Tab. 5-9 Comparative experiment on LE-FER dataset

数据集	模型	准确率（%）
LE-FER	Resnet50	82.09
	EfficientNetv1-b3	85.14
	Efficientnetv2	82.32
	Ghostnet	86.12
	MSFNet	89.12
	Multi-ghostnet	**88.60**

由实验结果可知， Multi-ghostnet 在基于学习环境构建的数据集 LE-FER 上表现最优，适用于在线教育环境下学习情绪的研究。Multi-ghostnet 在识别精度，模型大小和推理速度上实现了较好的平衡，应用于在线教育背景下，通过表情识别对学生学习情绪进行研究时，能够既保证识别准确率，又保证识别速度，有利于教师对教学进度进行及时地调整。

5.6 小结

通过表情识别对在线教育学生进行学习情绪研究时，由于上课人数不受地点限制，可能出现大规模学生同时上课的情况，教师若想实时监测到所有学生的表情变化，对于模型的推理速度要求很高。

鉴于以上问题，本章构建了基于多层级特征结合判别机制的快速轻量级人脸表情识别模型 Multi-ghostnet。本模型主要特点有：

（1）根据特征图之间的信息共享特性，利用简单线性运算取代原始卷积来获取相似特征图，以此节省参数量和计算量；

（2）通过多个层级的特征图划分，在特征图之间实现线性变换依次叠加。在极大程度上加深网络，保证较小参数量和较快推理速度的同时，有效减少因网络中包含过多的简单变换而造成的性能下降。

（3）基于第 4 章的研究，引入多尺度深度可分离卷积，深度可分离卷积本身能极大程度地降低网络的计算量和参数量，同时，将不同尺度的卷积核进行合理组合，网络在进行特征学习时，更加丰富的感受野能提取到更多尺度的特征，有效地提高了模型的分类性能。

（4）搭建动态判别模型，实现对特征图重要性的动态鉴别，并加入残差连接，提高了模型精度。

本章将 Multi-ghostnet 与目前表现较优的表情识别模型进行了对比实验。在保证较高准确率的前提下， Multi-ghostnet 兼顾了模型参数量，计算量和模型推理速度，综合性能较优。模型有效降低了实际部署时的硬件成本，实时识别速度达到25ms/帧，适用于在线教育环境下，对大规模同时上课的学生进行实时学习情绪的识别和分析，从而辅助教师把握课程节奏，提高课堂教学质量。

第 6 章 学习情绪分析系统的设计与实现

6.1 引言

学习者的面部表情传达了他们对教学的直接感受和理解程度。教师可以通过关注学生表情了解到学生的学习状态，从而设计更加合理化的教学内容，减轻学生在学习过程中产生的挫败感，提高教学质量。目前针对学习情境下的表情识别研究，主要存在数据集缺乏，模型复杂导致实际部署时硬件成本高，模型推理速度较慢无法实现实时识别的问题。本文第 3 章构建了学习情绪人脸表情数据集 LE-FER，弥补了目前针对学习背景的人脸表情数据集的空缺。第 4 章提出了基于多尺度特征结合注意力机制的轻量级表情识别网络，同时实现了低参数量和高精度，降低了硬件成本。第 5 章又在其基础上对模型推理速度进行研究，提出了基于多层级特征图融合判别机制的快速表情识别模型，实现了高精度，高推理速度，低参数量，低计算量的平衡，解决了大规模学生上课时实时识别的问题。

上述工作为面向在线教育应用的学习情绪分析任务奠定了良好的基础。为了进一步将其投入实际应用，本章搭建了学习情绪分析系统，将第 5 章提出的快速轻量级表情识别模型部署至平台上，教师端通过访问该平台，可以直观地获取学生的表情和学习状态。本章主要介绍了搭建该系统时的平台选择和功能设计过程，并对各功能模块进行了展示。

6.2 开发工具选择

本章系统在开发时选用 PyQt5 框架[78-83]。PyQt5 由 Phil Thompson 开发，是一个创建 GUI 应用程序的工具包，包含 600 多个类，6000 多个方法和函数，能跨平台在多个主流操作系统上运行，作为 Python 编程语言和 Qt 库的成功融合，PyQt5 是 Qt 库里最强大的库之一。

6.3 功能模块设计与开发

本文设计的学习情绪分析系统核心功能为学生个人学习情绪的分析和群体学习情绪的分析。同时考虑到在线教育学生人数较多，为便于教师管理，增加了人脸注册和学生签到等功能模块，图 6-1 展示了系统的各功能模块和对应算法。

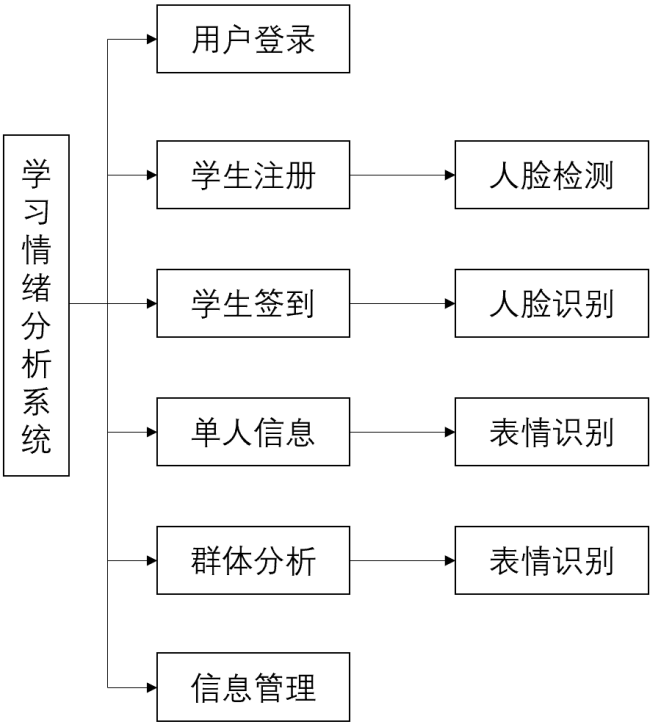


图 6-1 学习情绪分析系统功能模块图

Fig. 6-1 Functional module diagram of learning emotion analysis system

6.3.1 用户登录模块

由于学生人脸信息具有隐私性，本系统首先考虑信息的使用安全，设置了用户注册功能。用户在首次使用时需要先进行注册，后续登录则需要输入相互匹配的用户名和密码。

6.3.2 学生注册模块

学生首次进入课堂学习之前，需要进行人脸注册，注册界面如图6-2所示。注册成功后，该学生人脸图像将加入数据库，如图6-3所示。同时，系统设置了静默活体检测，使用照片无法进行注册，注册失败界面如图 6-4 所示。

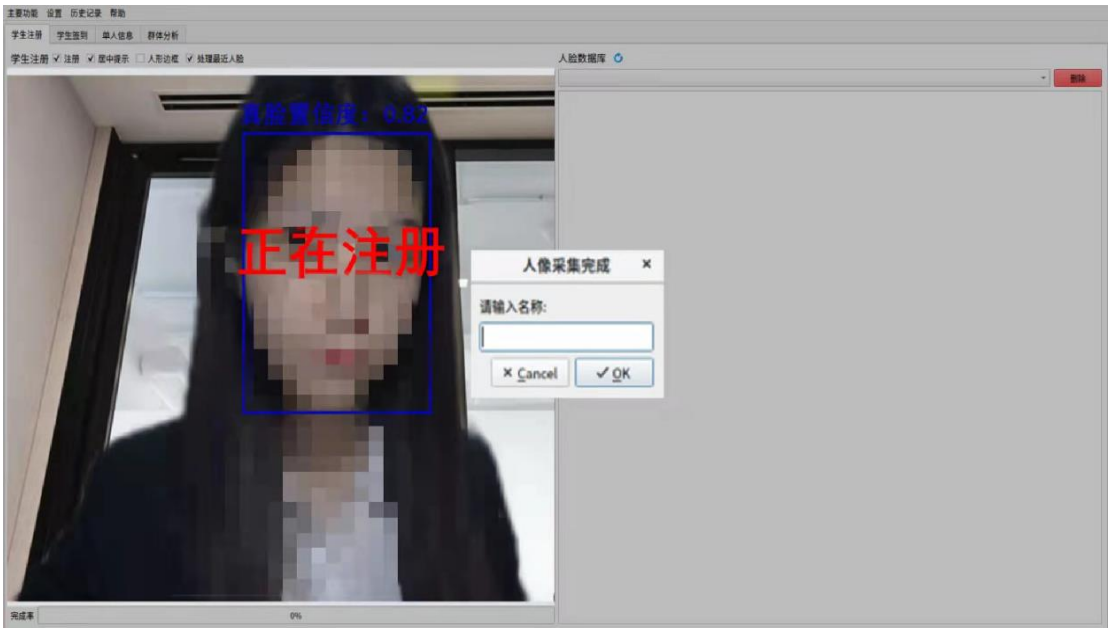


图 6-2 学生注册界面

Fig. 6-2 Student registration page

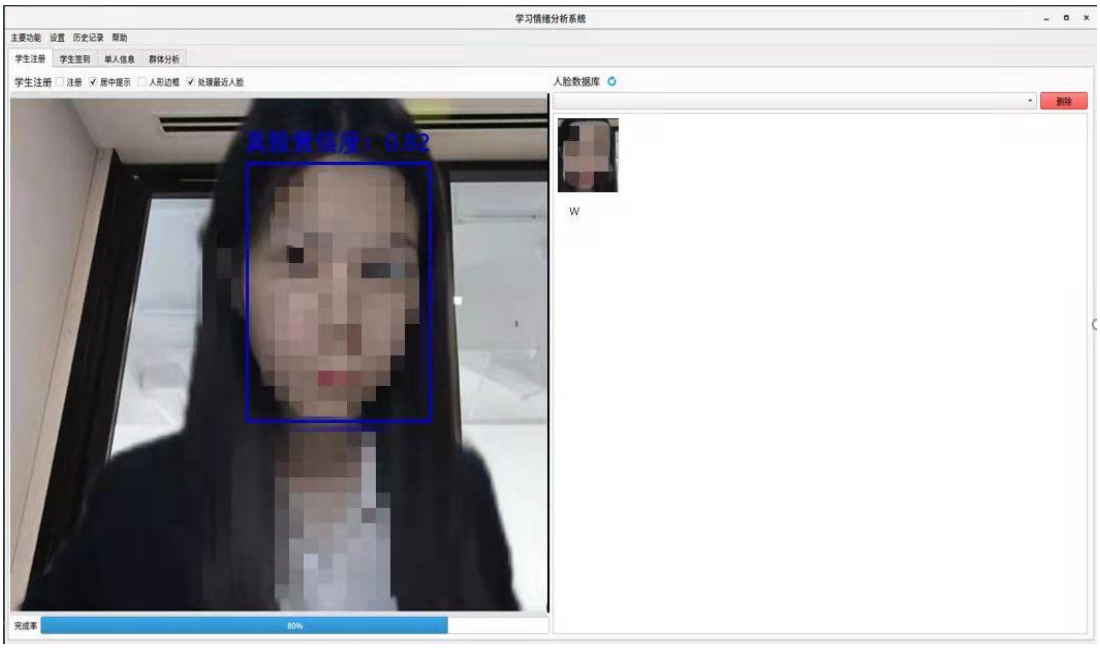


图 6-3 注册成功界面

Fig. 6-3 Successful registration page

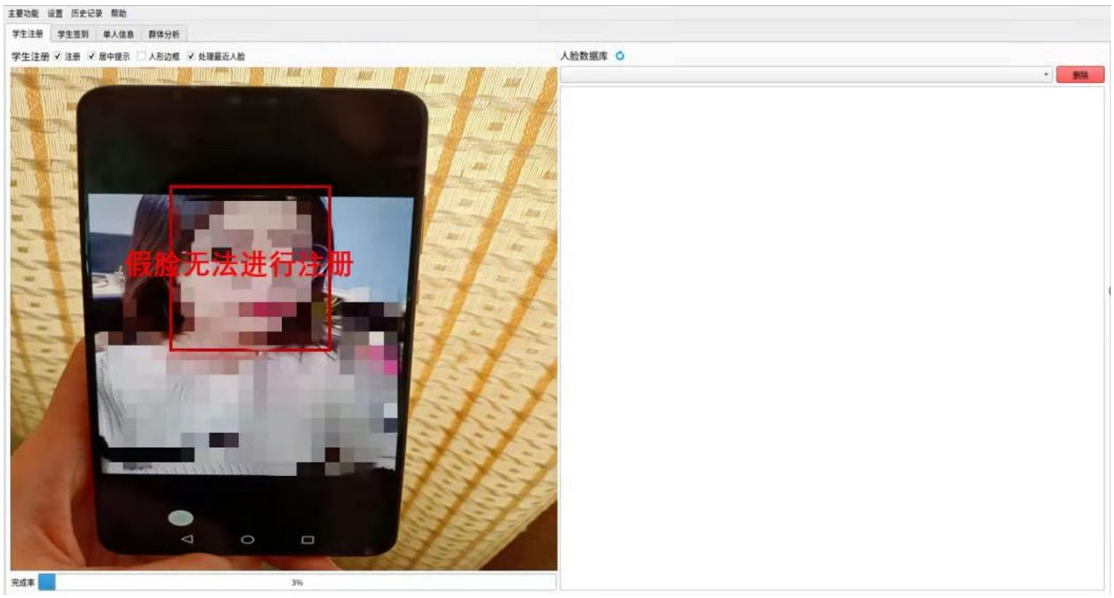


图 6-4 注册失败界面

Fig. 6-4 Registration failure page

6.3.3 学生签到模块

基于学生注册信息得到初始人脸数据库后，学生再次进入课堂需要进行签到，签到界面如图 6-5 所示，教师点击不同摄像头，可以看到不同学生的签到情况。

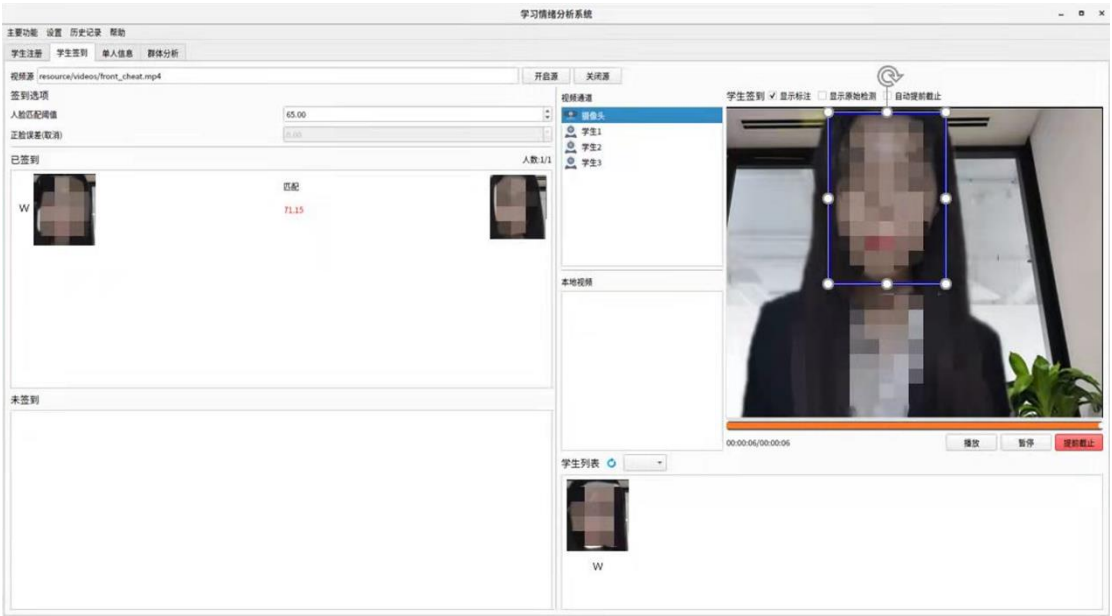


图 6-5 学生签到界面

Fig. 6-5 Student check-in interface

6.3.4 单人信息模块

教师端选择调用某一个摄像头可以得到该学生的单人信息，如图6-6所示，包含摄像头实时画面，学习情绪的实时监测结果，以及单人学习情绪随时间变化的曲线。

图 6-6 单人学习情绪分析界面

Fig. 6-6 Single-player learning emotion analysis screen

6.3.5 群体分析模块

本系统为方便教师了解整堂课所有学生的学习状态，设计了群体分析模块，该模块包含所有学生的实时画面，同时给出了占比人数最多的学习情绪随时间变化的曲线图，并统计了随时间累计各学习情绪占比情况。图6-7给出了同时调用两个摄像头的示范效果。

图 6-7 群体学习情绪分析界面

Fig. 6-7 Group learning emotion analysis interface

6.4 平台性能分析

本章平台开发基于 PyQt5 框架，对第 5 章提出的快速轻量级人脸表情识别模型进行部署应用，设计并实现了学习情绪分析系统。经过多次现场试运行，系统每 10 帧抽取一帧画面进行识别，连续测试了数百帧，识别速度达到 25ms/帧，能实现实时高效识别。同时，经过 PyInstaller 打包成.exe 格式后，该系统可以灵活应用于各主流操作系统，实用性和扩展性较高。

6.5 小结

本章基于第 5 章训练好的表情识别模型，搭建了学习情绪分析系统，并将模型部署到了平台上。首先介绍了平台开发工具，并介绍了系统主要包含的用户登录，学生注册，学生签到，单人学习情绪分析，群体分析等几个功能模块，最后对系统测试性能和推广性能进行了综合分析。教师端通过访问该平台，可以直观地获取学生的表情和学习状态，将面向在线教育的人脸表情识别技术推向了实际应用。

总结与展望

随着教育部关于“教育信息化”,“智慧教育”政策的不断推进，利用人工智能等现代化技术辅助教学受到了越来越多的关注，疫情冲击下，在线教育关注度大幅度上升。目前，针对在线教育，师生存在交互困难是亟待解决的问题。利用表情识别技术分析学生学习情绪，能帮助教师及时掌握学生学习状态，避免学生在出现疲倦厌烦等消极情绪后，无法及时缓解，从而影响课堂效率，有利于教师及时调整课堂安排，加强师生互动，提高教学质量。

目前面向在线教育应用的表情识别技术，还面临数据集缺乏；模型复杂，对硬件成本要求高；模型推理速度慢，难以实现实时识别等问题，为了解决上述问题，本文主要做了以下工作:

（1）深入研究了人脸检测和表情识别流程中各关键步骤涉及到的相关技术，并对卷积神经网络的相关概念进行了系统研究。首先，研究了人脸检测算法基本流程，并重点介绍了 MTCNN 算法。然后，重点研究了表情识别的算法流程，详细介绍了人脸图像预处理过程中涉及到的各项技术，以及特征提取和表情分类过程中常用到的传统方法和深度学习方法。最后，对卷积神经网络的相关概念和运算做了详细介绍，为后续算法设计奠定了理论基础。

（2）针对目前学习背景下人脸表情数据集空缺的问题，构建了学习情境下的学生自发学习情绪人脸表情数据集。首先，通过调研教育领域对于学习情绪的认定结合对在线课堂上学生的访谈调查，确定了愉悦、疲劳、悲伤、惊讶、厌倦、中性六种可以反映常见学习情绪的表情类别。然后，收集了真实课堂环境下学生自发表情数据，通过模型标注和人工审核相结合的方法进行数据标注。最后，通过网络爬虫和数据增强等方式扩充数据集。最终得到了适用于学习情景下表情识别研究的人脸数据集 LE-FER，包含 10000 张人脸图像和 6 类学习表情。

（3）针对实际应用部署时对模型轻量化和泛化性的要求，设计了一种基于多尺度特征结合注意力机制的轻量级人脸表情识别模型 Multi-scale Feature Net（MSFNet）。首先，该模型借鉴了密集连接的思想实现了特征图的重用，并利用不同尺度的卷积核使模型获取到多尺度的特征，提高了识别精度。其次，提出了一种“渐进式”轻量级结构，实现通道间信息交互的逐渐递减，在尽可能保证精度的前提下对模型的大小进行优化。最后，在模型中引入注意力机制，有助于特定通道信息的高效传播。该模型的参数量仅 0.33M，在 LE-FER 数据集上实现了 89.12%的准确率，在各开源数据集上均表现了较好的性能，实现轻量

化设计的同时保证了较高的识别精度和良好的泛化性。

（4）针对在线教育系统应用时对于模型识别实时性的要求，设计了一种基于多层级特征结合动态判别机制的快速轻量级人脸表情识别模型 Multi-ghostnet。首先，该模型利用相似特征图之间存在信息共享的特性，提出用多层级的简单线性变换取代普通卷积来获得相似特征图，以此实现模型的压缩。多层级变换使得特征图之间的线性变换依次叠加，显著加深了网络，并有效减少网络中直接进行粗糙的单层变换而造成的性能下降。其次，提出自适应动态判别机制，对各层级特征图重要性进行权重分配，最大程度地利用有效信息。最后，引入残差连接，从而保留较多的原始人脸信息，提高模型精度。将 Multi-ghostnet 与目前表现较优的表情识别模型进行了对比实验。在保证较高准确率的前提下，Multi-ghostnet 兼顾了模型参数量，计算量和模型推理速度，综合性能较优。模型有效降低了实际部署时的硬件成本，同时实时识别速度达到25ms/帧，适用于在线教育环境下，对大规模同时上课的学生进行实时学习情绪的识别和分析，从而辅助教师把握课程节奏，提高课堂教学质量。

（5）设计并实现了一种针对在线教育应用的学习情绪分析系统。基于已经训练好的表情识别模型，采用 PyQt5 框架搭建了学习情绪分析系统，将表情识别模型部署至平台上，该系统主要包括用户登录，学生注册，学生签到，单人学习情绪分析，群体分析等几个功能模块。现场试运行表明，系统每 10 帧抽取一帧画面进行识别，连续测试了数百帧，能顺利实现实时高效识别。同时，经过 PyInstaller 打包成.exe 格式后，该系统可以灵活应用于各主流操作系统，实用性和扩展性较高。教师端通过访问该平台，可以直观地获取学生的表情和学习状态，将面向在线教育的人脸表情识别技术推向实际应用。

本文未来的工作展望：

（1）在学习情绪数据集方面，首先，需要对数据样本进行进一步的扩充，可以对学习情绪分类做进一步的细分；其次，需要增加不同角度，不同光线下的人脸数据，以提高模型训练时的鲁棒性；最后，在进行数据标注时，可以选择更多性能优秀的模型进行辅助标注，对人脸表情图像进行人工审核分类时，需要充分结合表情识别领域以及教育心理领域的相关专业知识，以保证数据集的可靠性。

（2）在算法设计方面，首先，本文主要基于静态图像进行表情识别，后续可以考虑融入视频时序信息；其次，表情具有伪装性，有时无法体现真实情绪，而微表情难以控制和隐瞒，后续可以考虑对微表情识别做进一步的研究；最后，表情识别算法对人脸角度要求较高，而学生在学习过程中难免出现头部偏移摄像头，姿势改变等情况，后续应当对模型鲁棒性做进一步优化，并考虑加入头

部姿势分析，眼动分析等模块对学生学习情绪进行综合分析。

（3）在系统设计方面，首先可以丰富功能模块，结合教学内容增加一些师生互动的功能，进一步解决在线教育师生交互困难的问题；同时美化界面布局，优化用户体验。

参考文献

[1] 艾瑞咨询. 中国在线教育行业研究报告 2020 年[R]. 2020.

[2] Mehrabian A. Silent message[M]. Belmont, CA: Wadswworth, 1971.

[3] AlZoubi O, Calvo R A, Stevens R H. Classification of EEG for emotion recognition: an adaptive approach[C]// Proceedings of the 22nd Australasian Joint Conference on Artificial Intelligence. Springer, Berlin, Heidelberg, 2009: 52-61.

[4] Picard R W, Vyzas E, Healey J. Toward machine emotional intelligence: analysis of affective physiological state[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001, 23(10): 1175-1191.

[5] Belle A, Hargraves R H, Najarian K. An automated optimal engagement and attention detection system using electrocardiogram[J]. Computational and mathematical methods in medicine, 2012, 2012:528781-528781.

[6] Lee H, Kim Y, Park C. Classification of human attention to multimedia lecture[C]//2018 International Conference on Information Networking (ICOIN). IEEE, 2018: 914-916.

[7] D'Mello S, Graesser A. Automatic detection of learner's affect from gross body language[J]. Applied Artificial Intelligence, 2009, 23(2): 123-150.

[8] Bearden T S, Cassisi J E, White J N. Electrophysiological correlates of vigilance during a continuous performance test in healthy adults[J]. Applied psychophysiology and biofeedback, 2004, 29(3): 175-188.

[9] 吴沧海,熊焕亮,王映龙.远程学习中学习状态判断的情感计算研究[J].软件导刊(教育技术),2013,12(07):24-27.

[10] 卢希. 学习者在线学习状态检测工具的设计与实现[D]. 武汉: 华中科技大学,2016.

[11] 易佳玥. 基于眼动检测的在线学习状态实时评估系统的研究与应用[D]. 上海: 上海交通大学，2016.

[12] 熊碧辉,周后盘,黄经州,阮益权,周里程.一种融合视线检测的注意力检测方法[J].软件导刊,2018,17(07):31-36.

[13] Jewitt C, Kress G, Tsatsarelis J O & C. Exploring learning through visual, actional and linguistic communication: the multimodal environment of a science classroom[J]. Educational Review, 2001, 53(1):5-18.

[14] Gu W, Cheng X, Venkatesh Y V, et al. Facial expression recognition using radial encoding of local Gabor features and classifier synthesis[J]. Pattern Recognition, 2012, 45(1):80-91.

[15] Zhao S, Cai H, Liu H, Zhang J, Chen S. Feature selection mechanism in cnns for facial

expression recognition[C]. BMVC, 2018.

[16] Zeng J, Shan S, Chen X. Facial expression recognition with inconsistently annotated datasets[C]//Proceedings of the European conference on computer vision (ECCV), 2018, 222-237.

[17] Yang H, Ciftci U, Yin L. Facial expression recognition by de-expression residue learning[J]. International Journal on Computer Science & Engineering, 2018, 2(5):2220-2224.

[18] Kuo C M, Lai S H, Sarkis M. A compact deep learning model for robust facial expression recognition[C]. Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE. 2018.

[19] Wang K, Peng X, Yang J, et al. Suppressing uncertainties for large-scale facial expression recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 6897-6906.

[20] Vulpe-Grigoraşi A, Grigore O. Convolutional Neural Network Hyperparameters optimization for Facial Emotion Recognition[C]//2021 12th International Symposium on Advanced Topics in Electrical Engineering (ATEE). IEEE, 2021: 1-5.

[21] Savchenko A V. Facial expression and attributes recognition based on multi-task learning of lightweight neural networks[C]//2021 IEEE 19th International Symposium on Intelligent Systems and Informatics (SISY). IEEE, 2021: 119-124.

[22] Farzaneh A H, Qi X. Facial expression recognition in the wild via deep attentive center loss[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2021: 2402-2411.

[23] Zhang K , Zhang Z , Li Z , et al. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks[J]. IEEE Signal Processing Letters, 2016, 23(10):1499-1503.

[24] Shin M, Kim M, Kwon D S. Baseline CNN structure analysis for facial expression recognition[C]//25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN). IEEE, 2016: 724-729.

[25] Yu Z, Zhang C. Image based static facial expression recognition with multiple deep network learning[C]//Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. 2015:435-442.

[26] Pitaloka D A, Wulandari A, Basaruddin T, et al. Enhancing CNN with preprocessing stage in automatic emotion recognition[J]. Procedia computer science, 2017, 116: 523-529.

[27] Ebrahimi S, Michalski V, Konda K, et al. Recurrent neural networks for emotion recognition in video[C]//Proceedings of the 2015 ACM on International Conference on Multimodal

Interaction. 2015: 467-474.

[28] Bargal S A, Barsoum E, Ferrer C C, et al. Emotion recognition in tne wild from videos using images[C]//Proceedings of the 18th ACM International Conference on Multimodal Interaction. 2016: 433-436.

[29] Hassner T, Hared S, Pay E, et al. Effective face frontalization in unconstrained images[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 4295-4304.

[30] Huang R, Zhang S, Li T, et al. Beyond face rotation: global and local perception gan for photorealistic and identity preserving frontal view synthesis[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 2439-2448.

[31] Yin X, Yu X, Sohn, et al. Towards large-pose face frontalization in the wild[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 3990-3999.

[32] Tran L, Yin X, Liu X. Disentangled representation Learning gan for pose-invariant face recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017:1415-1424.

[33] Ahonen T, Hadid A. Face recognition with local binary patterns[C]//European Conference on Computer Vision. Springer, Berlin, Heidelberg, 2004: 469-481.

[34] Viola P A, Jones M J. Rapid object detection using a boosted cascade of simple features[C]// Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on. IEEE, 2001.

[35] Lyons M J, Akamatsu S, Kamachi M G, et al. Coding facial expressions with Gabor wavelets[C]// Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition. IEEE, 1998:200-205.

[36] Cai J, Huang P. Research of a real-time feature point tracking method based on the combination of improved SURF and P-KLT algorithm[J]. Acta Aeronautica et Astronautica Sinica, 2013, 34(5):1204-1214.

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

[38] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. The journal of machine learning research, 2014, 15(1): 1929-1958.

[39] Sohail A S M, Bhattacharya P. Classification of facial expressions using K-nearest neighbor classifier[C]//International Conference on Computer Vision/Computer Graphics Collaboration Techniques and Applications. Springer, Berlin, Heidelberg, 2007: 555-566.

[40] Wang X H, Liu A, Zhang S Q. New facial expression recognition based on FSVM and KNN[J]. Optik, 2015, 126(21): 3132-3134.

[41] Valstar M, Patras I, Pantic M. Facial action unit recognition using temporal templates[C]//RO-MAN 2004. 13th IEEE International Workshop on Robot and Human Interactive Communication (IEEE Catalog No. 04TH8759). IEEE, 2004: 253-258.

[42] Moghaddam B, Jebara T, Pentland A. Bayesian face recognition[J]. Pattern recognition, 2000, 33(11): 1771-1782.

[43] Mao Q, Rao Q, Yu Y, et al. Hierarchical Bayesian theme models for multipose facial expression recognition[J]. IEEE Transactions on Multimedia, 2016, 19(4): 861-873.

[44] Surace L, Patacchiola M, Battini Sönmez E, et al. Emotion recognition in the wild using deep neural networks and Bayesian classifiers[C]//Proceedings of the 19th ACM International Conference on Multimodal Interaction. 2017: 593-597.

[45] 周飞燕, 金林鹏, 董军. 卷积神经网络研究综述[J]. 计算机学报, 2017, 40(6):1229-1251.

[46] Fukushima K, Miyake S. A self-organizing neural network model for a mechanism of visual pattern recognition[M]//Competition and Cooperation in Neural Nets. Springer, Berlin, Heidelberg, 1982: 267-285.

[47] 李彦冬, 郝宗波, 雷航. 卷积神经网络研究综述[J]. 计算机应用, 2016, 36(9):2508-2515.

[48] Ekman P, Friesen W V. Constants across cultures in the face and emotion[J]. Journal of personality and social psychology, 1971, 17(2): 124.

[49] 张娜, 乔德聪. 基于深度学习的在线学习评论情感分析研究[J]. 河南城建学院学报, 2020, 29(4): 63-71.

[50] 何秀玲, 高倩, 李洋洋, 等. 基于深度学习模型的自发学习表情识别方法研究[J]. 计算机应用与软件, 2019, 36(3): 180-186.

[51] 唐康．人脸检测和表情识别研究及其在课堂教学评价中的应用[D]．重庆：重庆师范大学，2019.

[52] Whitehill U J, Serpell Z, Lin Y C, et a1. The faces of engagement: automatic recognition of student engagement from facial expressions[J]. IEEE Transactions on Affective Computing, 2014, 5(1): 86-98.

[53] Sharma P, Joshi S, Gautam S, et al. Student engagement detection using emotion analysis, eye tracking and head movement with machine learning[J]. arXiv preprint arXiv:1909.12913, 2019.

[54] Tonguç G, Ozkara B O. Automatic recognition of student emotions from facial expressions during a lecture[J]. Computers & Education, 2020, 148: 103797.

[55] Lehman B A, Zapata-Rivera D. Student emotions in conversation-based assessments[J]. IEEE Transactions on Learning Technologies, 2018, 11(1): 41-53.

[56] 孙立伟, 何国辉，吴礼发. 网络爬虫技术的研究[J].电脑知识与技术, 2010, 6(15):4112-4115.

[57] 高友文, 周本君, 胡晓飞. 基于数据增强的卷积神经网络图像识别研究[J]. 计算机技术与发展, 2018, 28(08):62-65.

[58] 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.

[59] Cooijmans T, Ballas N, Laurent C, et al. Recurrent batch normalization[J]. arXiv preprint arXiv:1603.09025, 2016.

[60] 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.

[61] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Advances in neural information processing systems, 2012, 25.

[62] Chollet F. Xception: Deep learning with depthwise separable convolutions[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 1251-1258.

[63] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7132-7141.

[64] Wu Y, Jia K, Sun Z. Facial Expression Recognition Based on Multi-scale Feature Fusion Convolutional Neural Network and Attention Mechanism[C]//Chinese Conference on Pattern Recognition and Computer Vision (PRCV). Springer, Cham, 2021: 324-335.

[65] 北京工业大学. 基于多尺度密集连接深度可分离网络的人脸表情识别方法: CN202110948629.9[P]. 2021-11-09.

[66] Nair V, Hinton G E. Rectified linear units improve restricted boltzmann machines[C]//Icml. 2010.

[67] Li S, Deng W. Reliable crowdsourcing end deep locality-preserving learning for unconstrained facial expression recognition[J]. IEEE Transactions on Image Processing, 2018, 28(1): 356-370.

[68] Lucey P, Cohn J F, Kanade T, et al. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression[C]//2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. IEEE, 2010: 94-101.

[69] Barsoum E, Zhang C, Ferrer C C, et al. Training deep networks for facial recognition with crowd-sourced label distribution[C]//Proceedings of the International Conference on Multirnodal Interaction. 2016: 279-283.

[70] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image

recognition[J]. arXiv preprint 2014.

[71] 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. 2018: 4510-4520.

[72] Ma N, Zhang X, Zheng H T, et al. Shufflenet v2: Practical guidelines for efficient cnn architecture design[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 116-131.

[73] Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks[C]//International Conference on Machine Learning. PMLR, 2019: 6105-6114.

[74] Tan M, Le Q. Efficientnetv2: Smaller models and faster training[C]//International Conference on Machine Learning. PMLR, 2021: 10096-10106.

[75] Han K, Wang Y, Tian Q, et al. Ghostnet: More features from cheap operations[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 1580-1589.

[76] Lin M, Chen Q, Yan S. Network in network[J]. arXiv preprint arXiv:1312.4400, 2013.

[77] Joulin A, Cissé M, Grangier D, et al. Efficient softmax approximation for gpus[C]//International Conference on Machine Learning. PMLR, 2017: 1302-1310.

[78] Summerfield M. Rapid GUI Programming with Python and Qt: The Definitive Guide to PyQt Programming (thesisback)[M]. Pearson Education, 2007.

[79] Harwani B M. Introduction to Python programming and developing GUI applications with PyQt, 1st Edition[J]. 2011, 32(12): 1088.

[80] 邱霞, 段渭军, 黄亮, 等. 基于 PyQt 无线传感器网络监控软件开发[J]. 现代电子技术, 2014, 37(16): 65-67.

[81] 何月顺, 杜萍, 丁秋林. 基于 Python 的电子邮件系统的研究与应用[J]. 现代图书情报技术, 2004 (4): 72-74.

[82] 孙强, 李建华, 李生红. 基于 Python 的文本分类系统开发研究[J]. 计算机应用与软件, 2011, 28(3): 13-14.

[83] 李琳. 基于 Python 的网络爬虫系统的设计与实现[J]. 信息通信, 2017 (9): 26-27.