单位代码：10293 密　级：公开

南京邮电大学

硕士学位论文

论文题目：基于深度学习的视频人脸表情识别研究

学　　号	1019010406
姓　　名	唐武宾
导　　师	曹雪虹
专业学位类别	工学硕士
类　　型	全日制
专业（领域）	信号与信息处理
论文提交日期	二O二二年四月

Research on video facial expression recognition based on deep learning

Thesis Submitted to Nanjing University of Posts and Telecommunications for the Degree of

Master of Master of Science in Engineering

By

Tang Wubin

Supervisor: Prof. Cao Xuehong

April 2022

摘要

人脸表情识别作为研究热点之一，已经被广泛应用于自动驾驶、网络社交、课堂教学等领域。人脸表情识别涉及多学科的交叉融合，如计算机学、生物学、心理学等，是一个具有价值的研究课题。由于表情存在微妙性，人脸所表现的表情不易区分，因此人脸表情识别仍是一大难题，同时本文主要关注的是视频数据，相比于单帧图像更为复杂。因此，基于以上问题，本文在传统 VGG-16+LSTM 网络框架的基础上进行优化设计，旨在提高视频人脸表情识别的准确性，具体工作内容如下：

（1）针对传统网络提取表情特征不准确的问题，本文提出了特征增强卷积网络（Feature Enhancement Convolutional Networks，FECN），从单帧特征增强和帧间特征增强两个角度进行研究，达到提高视频表情识别准确率的目的。首先，在 VGG-16 中间层外延一个 7×7 卷积层运算用于提取浅层人脸表情特征，并与深层特征融合增加人脸表情空间信息；然后，在 VGG-16 最后层应用扩张率为 2 的空洞卷积，在增加卷积运算感受野的同时降低信息损失；接下来，利用 Squeeze-Excitation 机制给人脸表情特征通道赋予权重，提升人脸表情单帧特征的准确性。最后，引入 Self-attention 机制，根据视频帧之间的相关性给视频帧赋予权重，提升人脸表情多帧特征准确性。该理论思想在 AFEW 数据集、CK+数据集、SFEW 数据集、FER2013 数据集上进行了验证比对，证实了模型优越性。

（2）针对特征增强卷积网络中大卷积块网络参数过大，训练时间长的问题，本文将 Inception 策略应用到特征增强卷积网络中，提出降参型特征增强卷积网络（Reduced Parameter Feature Enhancement Convolutional Networks，RPFECN），达到降低网络计算参量、提高视频表情识别准确率的目的。本文在基础 Inception 模型上又做了两种变形，分别在 AFEW 数据集、CK+数据集、SFEW 数据集、FER2013 数据集上进行实验仿真，证实了 RPFECN 网络不仅可以降低网络参数量同时还能够提高人脸表情识别的准确率。

（3）针对特征增强卷积网络中帧间注意力机制不适合处理长序列且信息提取不丰富的问题，本文采用 transformer 中多头注意力机制代替帧间注意力机制，并针对因注意力权重赋予偏差造成信息丢失问题提出了多头先验注意力机制（Multi-Head Prior Attention Mechanism，MHPAM）。该模型在 AFEW 数据集、CK+数据集上进行了仿真验证，证实了 MHPAM 能够提高人脸表情识别的准确率。

关键词： 人脸表情识别，特征增强，多头注意力机制

Abstract

As one of the research hotspots, facial expression recognition has been widely used in the fields of autonomous driving, social networking, and classroom teaching. Facial expression recognition involves the cross-integration of multiple disciplines, such as computer science, biology, psychology, etc., and is a valuable research topic. Due to the subtlety of expressions, the expressions on faces are not easy to distinguish, so real-time facial expression recognition is still a big problem. At the same time, this paper mainly focuses on video data, which is more complicated than single-frame images. Therefore, based on the above problems, this paper optimizes the design based on the traditional VGG-16+LSTM network framework, aiming to improve the accuracy of video facial expression recognition. The specific work is as follows:

（1）Aiming at the problem of inaccurate expression feature extraction by traditional networks, this paper proposes Feature Enhancement Convolutional Networks (FECNN), which is studied from the perspectives of single-frame feature enhancement and inter-frame feature enhancement to improve the accuracy of video expression recognition. rate purpose. First, a 7×7 convolutional layer operation is extended in the middle layer of VGG-16 to extract shallow facial expression features, and fused with deep features to increase the spatial information of facial expressions; then, the expansion rate is applied in the last layer of VGG-16 It is an atrous convolution of 2, which reduces the information loss while increasing the receptive field of the convolution operation. Next, the Squeeze-Excitation mechanism is used to give weights to the facial expression feature channels to improve the accuracy of facial expression single-frame features. Finally, a Self-attention mechanism is introduced to give weights to video frames according to the correlation between video frames to improve the accuracy of multi-frame features of facial expressions. The theoretical idea has been verified and compared on the AFEW dataset, CK+ dataset, SFEW dataset, and FER2013 dataset, which confirms the superiority of the model.

（2）Aiming at the problem that the inter-frame attention mechanism in the feature-enhanced convolutional network is not suitable for processing long sequences and the information extraction is not rich, this paper adopts the multi-head attention mechanism in the transformer to replace the inter-frame attention mechanism, and aims at the information caused by the deviation of attention weights. The dropout problem proposes the Multi-Head Prior Attention Mechanism (MHPAM). The model is simulated and verified on the AFEW dataset and the CK+ dataset, which confirms that MHPAM can

improve the accuracy of facial expression recognition.

（3）Aiming at the problem that the inter-frame attention mechanism in the feature-enhanced convolutional network is not suitable for processing long sequences and the information extraction is not rich, a multi-head attention mechanism in the transformer is proposed to replace the inter-frame attention mechanism. A new Multi-Head Prior Attention Mechanism (MHPAM) is introduced, which combines the output feature map of the multi-head attention mechanism with the VGG-16 output feature map, and uses the VGG-16 output feature map for feature supplementation and attention. Weight guidance greatly enriches the features. The model is verified by simulation on AFEW dataset and CK+ dataset, which confirms the superiority of the model.

Key words: Facial Expression Recognition， Feature enhancement， Multi-Head Attention Mechanism

目录

专用术语注释表

缩略词说明：

FACS	Facial Action Coding System	面部动作编码系统
LBP	Local Binary Patterns	局部二值模式
SIFT	Scale-Invariant Feature Transform	尺度不变特征变换
LPC	Linear predictive coding	线性预测编码系数
CLQP	Completed Local Quantized Pattern	完备局部量化模型
STCLQP	Spatio-temporal Completed Local Quantized Pattern	时空完成局部量化模式
ELRCN	Long-term recurrent convolutional network	长期循环卷积网络
RPNs	Region Proposal Networks	区域建议网络
AVEC	Audio and video emotional challenges	音视频情感挑战
CNN	Convolutional Neural Network	卷积神经网络
AAM	Active appearance model	主动外观模型
TCDCN	Tasks-Constrained Deep Convolutional Network	任务约束型深卷积网络
MTCNN	Multi-Task CNN	多任务卷积神经网络
GAN	Generative Adversarial Nets	生成对抗神经网络
MLP	MultiLayer Perceptron	多层感知器
SGD	Stochastic gradient descent	随机梯度下降法
RF	Receptive Field	感受野
FM	Feature Map	特征映射图
ReLU	Rectified Linear Unit	线性整流函数
LDL	Label Distribution Learning	标签分布学习
SCN	Self-Cure Network	自愈网络
MSCNN	Multi-Signal Convolutional Neural Network	多信号卷积神经网络
DC	Dilated Convolution	空洞卷积
SENet	Squeeze-and-Excitation Networks，SENet	通道间注意力机制
RNN	Recurrent Neural Networks，RNN	循环神经网络
DC	Depthwise Convolution	深度卷积
PC	Pointwise Convolution	逐点卷积

第一章 绪论

1.1 课题研究背景和意义

人脸表情是人对于外部事物的客观反映，是情感传递的最直接体现[1,2]。人脸表情以非语言的方式传达最真实的情感。在当代，随着电脑硬件水平的提高，人脸表情识别技术已经在多个领域得到应用，如健康[3]、个人辅助机器人[4]和许多其他人机交互系统[5]。对于人脸表情识别，其目标数据主要分为两类：一类是单帧静态图像数据集，一类是动态视频数据集，本文主要关注的是动态视频数据集中的人脸表情识别[6]。

心理学家 Mehrabian 研究发现，7%的信息通过语言进行传递，55%通过面部表情传递[7,8]，从而揭示了人脸表情对于信息传递的重要性。在 20 世纪，Ekman 和 Friesen 基于跨文化研究定义了六种基本表情，分别为愤怒、厌恶、恐惧、快乐、悲伤和惊讶，并建立了第一个系统的包含上千幅不同表情的人脸表情数据库[9]。随后，研究人员又将轻蔑加入到基本表情中，从而将基本表情扩充至七类[10]。近来，研究发现，七种基本表情模型具有文化特异性，但并不具有普遍性[11]。针对日常表情所表现出的复杂性和微妙性，基于基本表情的情感模型相比于其他情感描述模型，如面部动作编码系统（Facial Action Coding System，FACS) [12]和使用情感维度的连续模型[13]会受到一定的局限[14,15,16]，但是由于该套表情模型具有开创性的研究同时对于面部表情有了更为直接的定义，因此为后来的研究者广泛使用。在本文所采用的数据集也是基于基本表情模型的数据集。

视频人脸表情识别技术的研究发展对于促进人机交互、计算机图形学、非语言交流的发展不言而喻。对于视频人脸表情识别技术的应用，涉及到方方面面，其中突出领域为驾驶，通过对人脸表情的实时监测，可以清楚了解到乘客的心理状况，在出现紧张、害怕等消极情绪后，可以实现自动报警，及时保证乘客的人身安全。在传统机器学习中，特征都是采用手动方式提取，一旦数据量过于庞大，特征提取将是非常复杂的过程，而在深度学习中，主要采用神经网络进行特征提取，避免了手动提取的繁杂过程，同时特征提取效果更好。在本文中，采用的是深度学习进行视频人脸表情识别，主要采用卷积神经网络提取单帧图像人脸表情特征，采用自注意力机制处理多帧图像之间人脸表情的相关性，将多帧进行关联，从而突出多帧之间人脸表情的渐变信息，对于人脸表情特征提取起到了促进作用。同时文中所采用的人脸表情数据集存在光照、背景、翻转等干扰因素，更加贴近现实场景。因此本课题的研究既有理论研究意义，又具有很强的商业价值。

1.2 国内外研究现状

表情识别该领域的兴起源于 20 世纪，科学家根据人脸关键点检测实现表情建模，最后通过建模后的表情模型与原人脸进行比对，从而实现对动态视频数据的表情识别[17]。对于人脸表情识别的研究，起初是采用传统机器学习的方法进行表情识别，而传统机器学习由三部分所组成，分别为表情图像预处理、手动特征获取和表情分类。图像预处理，主要包括人脸对齐、数据增强、人脸归一化，通过人脸对齐自动定位出面部关键特征点，从而利用对齐后的人脸进行表情识别。数据增强主要为了解决数据不充分以及过拟合问题，常见的数据增强方法包括图像旋转、图像裁剪、图像翻转、增加噪声等，这些方法现在都为很多研究者所使用。人脸归一化主要是提高图像质量，减少混合干扰因素。手动特征提取是传统机器学习中最重要的一步，该过程主要是对最终表情识别起作用的特征进行提取压缩，从而进行识别。传统的特征提取方法主要包括，局部二值模式（Local Binary Patterns，LBP）[18]，Gabor 特征[19]，尺度不变特征变换（Scale-Invariant Feature Transform，SIFT）[20]，以及线性预测编码系数（Linear predictive coding，LPC）等[21]。LBP 主要是用来描述图像局部性特征，首先圈定一个 3×3 的方形窗口，将中心值与相邻 8 个像素点比较，如果中心值高于相邻像素点，则标记为 1，反之为 0。最后得到一组 8 位二进制数，此二进制数就是方形窗口的局部特征。张轩阁[22]提出将相邻视频帧之间的光流信息与 LBP-TOP[23]算法所获得的图像时空局部纹理特征相结合来获取面部细节信息，最后通过决策树进行分类，该算法在 CASME II 数据集上取得了不错成绩。基于完备局部量化模型（Completed Local Quantized Pattern，CLQP）方法，Huang[24]将其扩展至三维空间，提出时空完成局部量化模式（Spatio-temporal Completed Local Quantized Pattern，STCLQP）算法用于表情识别，计算方式相较于 LBP-TOP 算法无差，但是网络参数量过大，算法较为复杂。卢官明[25]采用 LBP-TOP 算子获取面部表情特征，然后对 LBP-TOP 特征向量进行降维，该理论的有效性在 CASME II 数据集上得到了验证。Gabor 特征主要就是借助 Gabor 内核对预期输入信号进行加窗处理，然后将加窗部分的信息提取出来，这部分信息就代表了该加窗部分的局部信息。周华平[26]基于 Gabor 特征提出 ENM-Gabor 差分权重特征提取方法，将所提取的 Gabor 特征中眼部、鼻部和嘴部区域通过权重赋值方式凸显对表情识别起关键作用的区域，抑制对表情识别无用或者作用不大区域，该思想的优越性在 JAFFE 数据集上得到了验证。谢慧华[27]基于 Gabor 特征提出 DE-Gabor 特征增强方法，首先基于特征下采样降低 Gabor 特征维度，其次将具有表情信息的特征字典与具有身份鉴别的中性特征字典组合表示，这两者信息的组合既突出了表情特征又突出了每个图像所展现的身份信息，一定程度上进行了特征增强，同时该思想在 BU3DFE 数据集[28]上进行了验证，明显优于其他网络框架。之后，

由于大数据的兴起，机器学习无法满足日益增长的数据，深度学习孕育而生，人脸表情识别也慢慢转向了深度学习，对于深度学习，在设计网络的过程中，需要关注数据集的类型，人脸表情识别的数据集主要分为静态图像表情数据集和动态视频数据集，在基于静态的方法中[29,30,31]，采用网络对单个人脸图像进行特征提取以及表情分类，而基于动态的方法中[32,33,34]需要考虑相邻多帧之间表情的相关性，因此在设计网络过程中，需要体现表情的渐变信息。除去单模态特征提取，也可结合多模态，如文字、音频等，并将提取到的特征进行融合，从而进一步提高表情识别的准确性。

自从 2013 年以来，为了满足研究所需的数据，许多科学家创建出了具有真实价值的数据，如 MMI 数据集[35]、SFEW 数据集[36]等。但是这些数据集的数据不存在任何干扰因素，无法贴合实际，因此后面基于一些国际竞赛，如 FER2013[37]和 EmotiW[38,39,40]，开始从一些现实场景中构建训练以及测试所需的数据集，这些数据集相比于理想情况下的数据集增加了很多干扰因素，如光照、遮挡、翻转等，其中最具代表的是 AFEW 数据集[41]。在大数据的背景下，机器学习已经无法满足在处理效率、提取精度以及智能化的一些需求，而深度学习借助如 GPU 等电脑硬件的发展对于处理大数据表现出了优良的信息处理能力，突出领域为目标检测、图像分类、图像切割等。通过应用深度学习，各领域已经达到了最先进的识别精度，并大大超过了以往的研究成果[42,43,44,45]。

对于人脸表情识别，深度学习都是通过网络自适应学习特征，避免了手动获取特征带来的不便，同时网络对于光照、遮挡、翻转等干扰因素有着更好的鲁棒性。近几年，对于人脸表情识别主要采用卷积神经网络、深层置信网络、循环神经网络等算法。其中卷积神经网络主要适用于单帧图像数据集，同时也可以作为动态视频数据集的单帧图像特征提取，循环神经网络只能适用于动态视频数据集，这是因为该网络主要是将前后多帧之间表情的渐变信息凸显出来，需要多帧之间表情存在相关性，而单帧图像数据集不具备这个条件。Deng 等[46]采用 MIMAMO (Micro-MacroMotion) Net 的两流循环网络将宏观视频表情变化信息和微观视频表情渐变信息结合起来，从而更好的捕获了表情信息，其中 MIMAMO (Micro-MacroMotion) Net 是由双流卷积神经网络和门控循环单元网络所组成。该网络的优越性在视频情感数据集 OMG 数据集和 Aff-Wild 数据集上得到了验证。Khor 等[47]提出了一种丰富的长期循环卷积网络（Long-term recurrent convolutional network，ELRCN），该网络由提取单帧表情特征模块和提取多帧表情特征模块组成，将空间信息和帧与帧之间动态信息结合，促进表情识别的准确性。Li[48]为了克服特征提取的不便以及不准确问题，提出了一种 Faster R-CNN 的算法来提取人脸表情特征，首先通过预处理提高图像质量，同时去除部分干扰因素，其次采用卷积神经网络进行单帧特征提取，之后通过区域建议网络（Region Proposal Networks，RPNs）对特征

进行重组，筛选对最后表情分类重要的特征，最后通过 Faster R-CNN 网络进行表情识别分类，该模型的优越性在多个表情数据集上得到验证。Hu[49]针对视频序列中面部表情识别不准确问题，提出了一个既能实现局部特征增强又能实现全局特征增强的集成网络框架，其中局部特征增强采用人脸检测技术识别人脸边缘性轮廓，通过边缘性轮廓构建人脸注意力区域，从而提高了人脸的分辨率，进而增强了人脸所展现出的表情信息，然后将此表情信息送至 CNN 网络中进行预测。全局特征增强主要是采用经过改进后的 CNN-LSTM 网络框架对未经过处理的人脸图像进行人脸表情提取以及分类，最后通过特征加权将全局特征和局部特征结合，从而得到表情预测结果。该模型在 AFEW 数据集、CK+数据集上具有不错的表现。Ruan[50]为了加强帧与帧之间的相关性同时又能突出特定帧特定信息，提出了一种新颖的特征分解和重建学习（Feature Decomposition and Reconstruction Learning，FDRL）方法，将多帧表情信息提取，其中相似信息共享，非相似信息作为某一特定帧独立信息，将独立信息和共享信息结合即代表了某一特定帧的全局信息，通过该方法兼顾了相似信息以及特定信息。其中 FDRL 主要由两个关键网络组成：特征分解网络（Feature Decomposition Network，FDN）和特征重构网络（Feature Reconstruction Network，FRN）。FDN 主要是用于对单帧图像进行表情特征提取，然后，FRN 对捕获到的特征进行重组，将特征分为共享特征和独立特征，通过特征重组进一步加强相邻帧之间的联系，从而提高了表情识别率。该理论在 CK+数据集、SFEW 数据集等上取得了不错的表现。Wang[51]为了解决网络过深导致梯度消失问题同时进一步提高表情识别的准确率，提出了残差注意力网络，该网络通过残差网络解决网络过深所带来的梯度消失问题，通过堆叠注意力模块对所提取到的特征赋予注意力权重，突出有用特征。整体网络共一百多层，一定程度上避免了由于网络过浅所带来的注意力权重赋予偏差问题。此网络框架在 CIFAR-10 数据集和 CIFAR-100 数据集上取得不错的成绩。Wang 所利用的注意力模块主要是在空间上编码并赋予注意力权重，但是其实对于每个特征通道也可以通过赋予注意力权重来凸显对表情预测起有效作用的信息，抑制对表情预测无效甚至有错误引导的信息，而这一理论思想为 Jie[52]提出，其采用通道赋予注意力权重的方法突出通道对于表情识别所起的作用并命名为 “挤压和激励”（Squeeze-and-Excitation，SE）块，同时将这些 SE 块堆叠在一起可以构建 SENet 架构，这种架构在提升表情识别率的同时还具有很好的泛化作用，能够以较少的计算成本带来显著性的性能提升。之后基于通道以及空间两种注意力权重赋予的基础上，Woo[53]提出了卷积块注意模块（Convolutional Block Attention Module，CBAM），该方法其实就是在空间和通道同时对特征赋予注意力权重，然后将得到的注意力权重图与网络的输出特征图相乘从而获得一个具有区分度的特征图，避免了从某一单维度赋予注意力权重所带来的偏差问题。同时 CBAM 是一种轻量化模块，可以嫁接到任何深度学习网络框架中，所带来的

参数也很少，因此是一个不错的模块。

1.3 课题研究难点

对于视频人脸表情识别所采用的数据集是在真实环境下采集到的视频数据，这些现实场景下的数据有着如光照、遮挡、像素低等干扰因素。在这些具有干扰因素的数据集下提取特征对于机器学习是一件很困难的事，而深度学习对于干扰具有很好的鲁棒性，因此，采用深度学习进行视频人脸表情识别已经成为一种趋势。虽然深度学习对干扰具有很好的鲁棒性，但是相较于机器学习需要更多的数据样本进行训练，而现在所使用的数据集相对而言数据量不足，同时数据本身质量不高，这对于人脸表情鲁棒特征提取产生了一个不小的挑战。

1.4 课题研究工作及章节安排

本文主要基于深度学习搭建视频人脸表情识别框架，希望可以通过此网络框架对具有干扰因素的现实场景下的视频数据集进行特征提取以及表情分类，提高人脸表情的识别准确率。

本文的一开始介绍了深度学习中的一些经典模型框架，其中包括卷积神经网络（Convolutional Neural Network，CNN）和自注意力机制(Self-attention mechanism)。卷积神经网络主要用于对单帧图像表情特征提取，自注意力机制主要用于对多帧之间信息相关性进行记忆学习，从而将表情的渐变信息体现出来，这两者结合可以实现视频人脸表情识别分类。首先，针对传统深度表情识别网络框架提取表情特征不准确的问题，提出特征增强卷积网络，从单帧特征增强和帧间特征增强两个角度进行研究，设计浅层特征增强模块，并引入空洞卷积和 SENet，达到增强单帧人脸表情特征的目的；采用帧间注意力机制实现多帧人脸表情特征增强，两者融合达到提高视频表情识别准确率的目的。其次，针对特征增强卷积网络中大卷积块网络参数过大，训练时间长的问题，将 Inception 策略应用到网络中，提出降参型特征增强卷积网络，达到降低网络计算参量、提高视频表情识别准确率的目的。最后，针对帧间注意力机制处理长序列效果不佳以及信息提取不丰富的问题，本文采用 transformer 中多头注意力机制代替帧间注意力机制，并针对因注意力权重赋予偏差造成信息丢失问题提出了多头先验注意力机制，达到提高视频表情识别准确率的目的。

本文安排主要内容分为六章，各章节安排如下：

第一章介绍了视频表情识别的研究背景、研究意义和国内外的研究现状，提出该研究课题的一些研究难点，最后对本文工作、章节安排进行概括。

第二章首先介绍了使用神经网络进行表情识别的基本流程，然后主要介绍了主流的神经

网络，最后介绍了本文使用到的相关人脸表情数据库。

第三章针对传统深度表情识别网络框架提取表情特征不准确的问题，本文提出了特征增强卷积网络。并详细介绍了空洞卷积，SENet，浅层特征增强模块和帧间注意力机制，对于整体设计进行了详细阐述。对各个模块进行了消融实验，佐证了各个模块的价值，同时分别在多个数据集上进行了仿真对比，证明了该增强卷积网络的优越性。

第四章针对特征增强卷积网络中 7×7 大卷积块网络参数过大，训练时间长的问题，将 Inception 策略应用到网络中，提出降参型特征增强卷积网络，达到降低网络计算参量、提高视频表情识别准确率的目的。该思路在多个数据集上进行了仿真对比，皆证明了该思想的可行性。

第五章针对帧间注意力机制处理长序列效果不佳信息提取不丰富的问题，将 transformer 中多头思想应用到网络中，并针对因注意力权重赋予偏差造成信息丢失问题提出了多头先验注意力机制，该想法在多个数据集上进行了仿真对比，皆证明了该思想的可行性。

第六章对本文的研究工作进行了总结和分析，反思算法研究中存在的不足，并对以后的研究工作进行展望。

第二章 视频表情识别相关技术简介

针对视频进行人脸表情识别，主要分为两个阶段，分别为训练和测试，同时数据集也分为训练集和测试集，首先通过训练集对神经网络进行训练，通过每次传入的视频片段和对应的表情分类标签来更新神经网络中的参数，使得网络在训练结束后能够达到一个较好的拟合效果，然后将测试集输入到已经训练好的神经网络，最终得出人脸表情识别的准确率和 F1 分数。

2.1 基于神经网络的视频表情识别基本流程

对于人脸表情识别，由于没有涉及到音频、文本等其他多模态的信息，因此输入只可能是视频图像。对于视频图像的人脸表情识别，其基本步骤分为四步，分别为：数据收集、数据预处理、特征提取以及表情分类。具体阐述分为：（1）数据集主要分为两类，第一类是在实验室录制的没有任何干扰因素的数据集，该类数据集相对来说特征提取简单，第二类是现实场景下的数据集，该类数据集一般来源于影视作品、网络等，具有如光照、遮挡、像素低等干扰因素，特征提取相对较为困难，本文使用的 AFEW 数据集、CK+数据集就是现实场景下的数据集。（2）数据预处理包括：视频帧提取、人脸对齐、数据增强、人脸归一化，其中视频帧提取就是将采集到的视频提取成视频帧，同时每个视频帧都标记表情标签，在本文中，官方给予的数据集包含了视频以及已经处理过的视频帧，因此不需要进行视频帧提取，直接采用官方提供的视频帧数据集进行表情识别；人脸对齐就是检测人脸图像，复刻人脸关键特征点，通过关键特征点重建人脸，最后利用重建后的人脸进行表情识别；数据增强主要是为了现在主流数据集数据样本不足的问题，通过数据增强扩充所需样本数，提高训练效果；人脸归一化主要为了解决数据中光照、遮挡等干扰因素对于表情识别的影响，其包含几何归一化和灰度归一化，几何归一化分两步：人脸校正和人脸裁剪。而灰度归一化主要是增加图像的对比度，进行光照补偿。（3）根据数据集的类型搭建了神经网络进行特征提取，其中采用卷积神经网络提取单帧图像特征，采用自注意力机制提取多帧之间的表情渐变信息。（4）通过神经网络进行表情分类，通过训练集训练模型参数，最后通过测试集来测试网络的拟合效果，得到一个表情识别率和 F1 分数。对于其基本流程如图 2.1 所示。

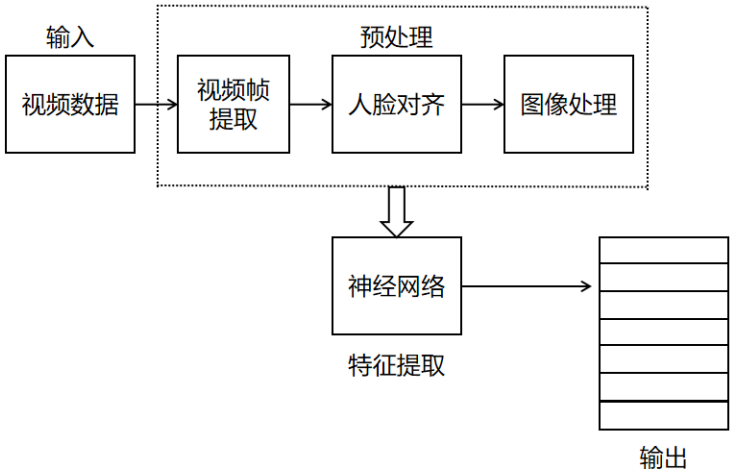


图 2.1 神经网络视频表情识别系统流程图

2.2 数据预处理

2.2.1 人脸对齐

对于现实场景下的数据集，其具有如光照、遮挡、背景等众多干扰因素，这些干扰因素的存在会造成特征提取的不准确。为了解决该问题，需要对人脸图像对齐，使其能够对齐并具有规范性，同时对于部分干扰因素进行预处理，初步筛选并去除无用信息。

表 2.1 常用人脸检测对齐方法对比

类型		关键点	实时	速度	效果
整体	AAM[55]	68	否	一般	差
基于部分	Mot[56]	39/68	否	慢	好
	DRMF[57]	66	否	快	
级联回归	SDM[58]	49	是	快	好
	3000fps[59]	68	是		
	Incremental[60]	49	是		
深度学习	cascaded CNN[61]	5	是	快	好
	MTCNN[62]	5	是		很好

在很多表情识别的相关工作中，人脸对齐是不可或缺的一步，在拿到人脸表情数据集后，首先需要进行人脸对齐，也就是定位出人脸面部关键特征点，去除图像的空白区域，将一些不必要的背景信息筛除，在人脸对齐操作之后的数据集会送到神经网络中进行特征提取以及表情分类。虽然人脸检测是实现特征学习的必须步骤，但是利用面部的局部关键点坐标进一步进行人脸对齐可以显著提高人脸表情识别的性能[63]。在人脸表情领域，人脸对齐是数据预处理部分很重要的一步，因为它可以减少背景、空白区域所带来的影响。

表 2.1 是现阶段用的比较多的一些开源性人脸对齐方法，该表分别从关键点数目、是否

实时、检测速度和效果上进行了对比。主动外观模型（Active Appearance Model ，AAM）是最为经典的人脸对齐方法，它依据整体面部轮廓去除背景信息和空白区域，在定位人脸过程中由于非实时同时速度比较慢，因此效果不佳。MoT（Mixtures of Trees）结构模型和 DRMF（Discriminative Response Map Fitting）都是基于面部局部区域的方法，它主要关注的是面部边缘区域的局部信息来定位人脸，两者虽然非实时的，但是面部定位效果还是不错的。此外还有一些学者通过使用线性回归函数级联的方式将图像中所展示的面部映射至面部特征关键点，从而实现图像与特征关键点一一对应，这种方式的人脸对齐效果是不错的，其中比较典型的是 IntraFace 中实现的监督下降方法（SDM）[64]、人脸对齐 3000 fps 和增量人脸对齐。之后随着深度学习的发展，有些学者将级联型的 CNN 用于人脸对齐，这种方法所需关键点少、能够实时进行定位、速度也非常快，因此效果很好。在这基础上，进一步发展，任务约束型深卷积网络[65]（Tasks-Constrained Deep Convolutional Network，TCDCN）和多任务卷积神经网络（Multi-Task CNN，MTCNN）被用来进行人脸对齐，提高对齐准确率。技术的迭代使得人脸对齐的准确率得到了进一步的提升。

2.2.2 数据增强

对于神经网络的训练需要大量的数据作支撑，但现实情况是现阶段公开的数据集数据相对而言都较少，如 AFEW 数据集，其视频片段总共就 1000 个左右，因此为了扩充数据集，数据增强是关键的一步。数据增强主要分为两类：实时数据增强[66]和离线数据增强[67]。

通常实时数据增强会内嵌至深度学习软件包中，其目的在于抑制网络过拟合现象的产生。在输入样本后，网络会对输入样本的边角处或者中心处进行裁剪，同时会对裁剪完毕的样本水平翻转，从而进一步扩充样本数，这样操作后得到的数据将是原始数据的数倍。由于数据在网络中进行了扩充，因此在训练结束后的测试会有两种预测模式：分别为使用图像的中心点进行预测[68,69]和对所有数据增强区域[70]的预测结果求均值。

离线数据增强就是在输入样本之前对数据进行手动扩充，从而扩充数据的数量，增加数据的多样性。常见的离线数据增强的方法包括图像翻转、图像裁剪、增加噪声、对比度变换、色彩抖动等。对于增加噪声，常见的噪声类型分为斑点噪声、散斑噪声[71]和高斯噪声[72]。对比度变换通过改变每个像素的饱和度和值[73]来扩增数据。基于离线数据增强，可以增加更多的样本，促进网络的训练，使得网络具有更好的拟合效果。除了手动扩充数据，Yu[74]研究发现可以采用仿射变换矩阵来自动对图像转换方位。与此同时，有学者发现也可以采用神经网络对数据进行样本扩充，比如 Abbesneja[75]发明了一个基于 3D 卷积神经网络的数据扩充网络，

经过该网络的数据会自动生成具有不同饱和度的人脸图像。最近较为火热的生成对抗网络（Generative Adversarial Nets，GAN）也可以实现数据增强，在表情标签一致的情况下生成多种姿态的人脸表情图像，从而实现数据的扩充。

2.3 深度学习神经网络

在设计网络的时候，主要基于深度学习网络框架自适应提取特征，因此，在本章节中，我们将会介绍一些较为经典的深度学习神经网络。

2.3.1 卷积神经网络

卷积神经网络（Convolutional Neural Networks，CNN）是深度学习里第一个问世的网络框架，它的理论机制源于动物感知外部事物。1959 年末，Hubel 对猫的视觉皮层细胞进行研究，提出了感受野的概念，即存在部分细胞对于输入事物的空间信息很敏感。之后，到 80 年代，Fukushima 在感受野的理论基础上提出了神经认知机（Neocognitron）的概念，该概念指出脑部会将视觉模式分解成多个子视觉模式，从而每个子视觉模式会并行处理输入信号的特征，最后将处理完毕的信号整合输出。该理论概念可以作为卷积神经网络的前身。20 世纪 90 年代，LeCun 等人提出了第一个已经训练完毕并具有一定效果的卷积神经网络框架，并取名为 LeNet-5，之后，一系列卷积神经网络应运而生，其中包括 AlexNet[76]，VGGNet[77]，GoogleNet[78]，ResNet 等基础模型，现阶段大部分研究都是基于这些基础模型产生。

卷积神经网络现阶段已经在多个领域得到了应用，这其中也包括了人脸表情识别领域。在 21 世纪初，有研究者发现使用卷积神经网络作为人脸表情识别的主干网络能够较好的提取特征，并且对于多变表情的识别性能明显优于多层感知器（MultiLayer Perceptron，MLP），因此，在这之后，很多研究者都采用卷积神经网络来提取表情特征并进行表情分类。

卷积神经网络的核心组成模块包括卷积层、激活层、池化层、全连接层。一般来说，卷积层和池化层是联合使用，一个或者多个卷积层之后必定会是池化层，同时为了增加网络的非线性能力，对于每个层都会添加激活层。在进行网络参数更新时，卷积神经网络一般采用随机梯度下降法（Stochastic gradient descent，SGD），通过不停的迭代训练使得网络的损失函数降到最低，在损失函数降低的过程中同步更新网络参数，从而使得网络具有很好的拟合效果。卷积神经网络的前置区域一般由卷积层和池化层组成，最后由全连接层和分类器组成，通过前置网络对输入图像进行特征提取，最后通过分类器进行分类，从而确定输入图像的标

签。

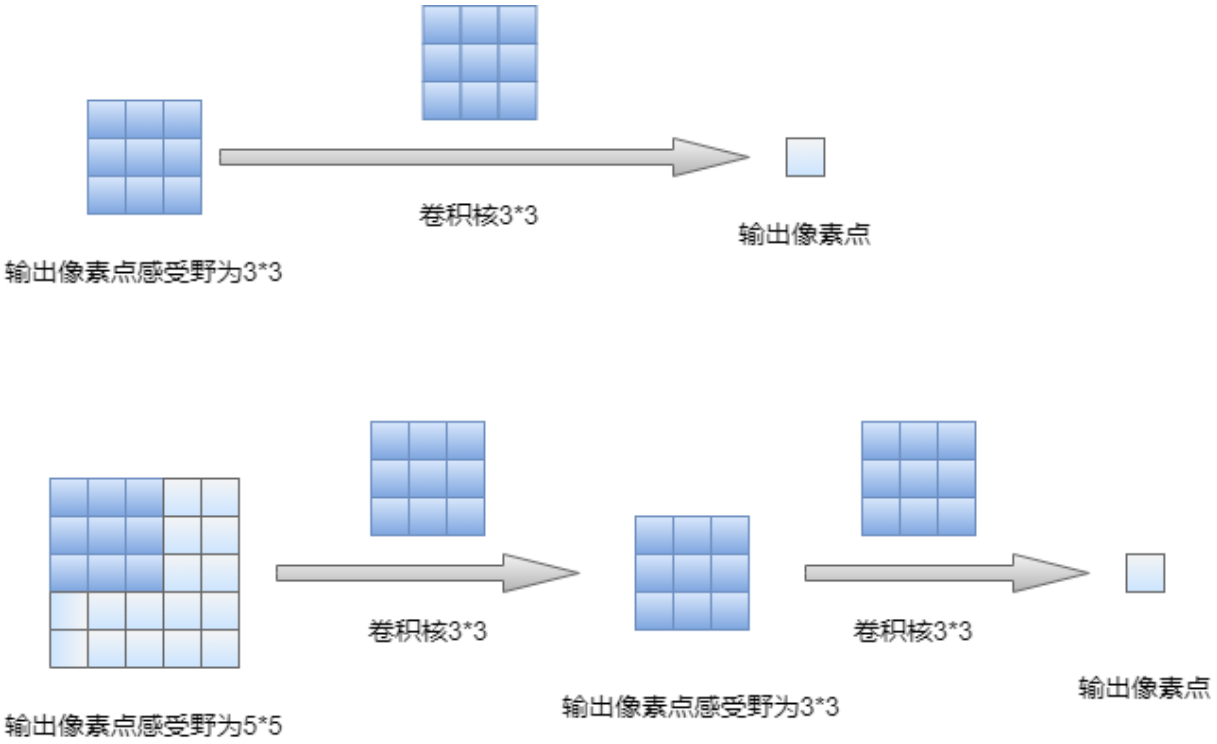


图 2.2 不同卷积核下的感受野表示

卷积层主要是采用卷积核平移方式进行卷积运算，针对特定输入图像会形成多个特征映射图（Feature Map，FP），每个特征映射图由多个像素点组成，每个像素点代表输入某一区域特征，一般而言，该像素点为该对应区域特征加权求和而来。对于卷积核的选择，一般主要是 4 种，分别为 1×1、3×3、5×5 和 7×7，其中 1×1、3×3、5×5 为小卷积核，7×7 为大卷积核，同时针对整个卷积神经网络卷积层的数量确定网络属于深层卷积神经网络还是浅层卷积神经网络，其中浅层卷积神经网络一般层数较少，侧重于图像的浅层特征，如图像的边缘信息等，深层卷积神经网络一般层数较多，提取的是图像深层特征，每个特征所带有的信息较为丰富，每个深层特征都由多个浅层特征所组成，包含了输入图像各区域特征。相比于浅层特征，由于深层特征处于网络的后端，因此感受野（Receptive Field，RF）会更大，每个像素点所蕴含的信息更为丰富，所映射的特征区域也更广。假设第一个卷积层中的卷积核为 3×3，那么它对应输入的感受野区域就为 3×3，同理，第二个卷积层中的卷积核也为 3×3，那么输出特征映射的像素点感受野也为 3×3，对应到初始输入就是 5×5 的感受野区域，等价于一个 5×5 的卷积操作。详情见图 2.2。

对于一个 5×5 的卷积核，其感受野区域也就是 5×5，如果采用 3×3 的卷积核，想要实现相同的感受野，那就需要两个 3×3 卷积核级联才能实现，这种操作会加深网络的深度，但是网络中的参数量确实是降低了。比如一层 5×5 卷积需要训练的参数量是 25（5×5），而两

个 3×3 卷积级联需要训练的参数量是 18（2×3×3），训练参数量相比于 5×5 卷积得到了降低，在实际操作过程中，卷积核一般是多通道的，如 512，1024 等，因此如果考虑到通道那参数量将会成几何倍数降低，这使得小卷积核为广大研究者所热衷。对于小卷积核，是否就可以完全替代大卷积核呢？那不会，小卷积核的频繁使用会造成网络层数的增加，进而导致梯度消失以及梯度爆炸的产生。因此对于使用小卷积核一定要根据具体情况，不能盲目采用小卷积核堆积。

在卷积神经网络发展的初期，网络主要采用线性函数叠加，但是实际上这种线性函数的叠加是无法提升网络的学习能力的，进而会导致网络参数无法得到更新。为了解决该问题，引入了激活函数的理论思想，在卷积层后会加入激活层用以提高网络的非线性能力，使得网络具有很好的泛化能力。常用的激活函数有 Sigmoid 函数、Tanh 函数、线性整流函数（Rectified Linear Unit，ReLU）等。

Sigmoid 激活函数如式 2-1 所示：

$$h(z)=\frac{1}{1+e^{-z}} \tag{2-1}$$

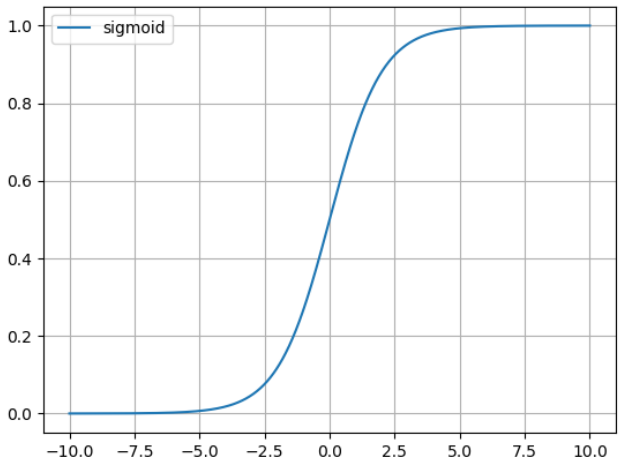


图 2.3 Sigmoid 函数曲线图

Sigmoid 激活函数的形状如图 2.3 所示，该函数输出大小介于（0，1）之间，是一个抛物线型递增函数。在神经网络中，该激活函数可以将输入值通过非线性变换得到概率输出，从而实现分类识别。对于该激活函数，其存在一个问题，那就是随着网络深度的加深，Sigmoid 函数会趋于平缓，反向传播算法所计算的梯度变化几乎为零，从而浅层网络参数无法得到更新，最后导致梯度消失的问题。

Tanh 激活函数如式 2-2 所示：

$$h(z)=\frac{e^{z}-e^{-z}}{e^{z}+e^{-z}} \tag{2-2}$$

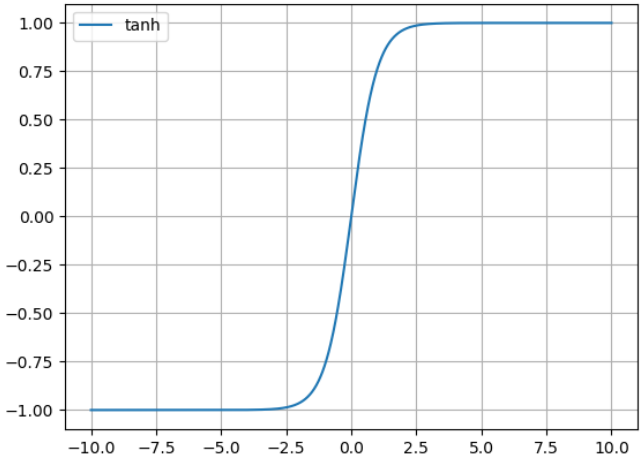


图 2.4 Tanh 函数曲线图

Tanh 激活函数的形状如图 2.4 所示，函数输出结果介于（-1，1）之间，同时整个函数的均值为 0，是一个抛物线型递增函数。对于该激活函数的应用绝大多数是在循环神经网络中，最常见的是 LSTM。该函数和 Sigmoid 激活函数一样，存在梯度消失的风险。

ReLU 激活函数如式 2-3 所示：

$$h(z)=\max(0,z) \tag{2-3}$$

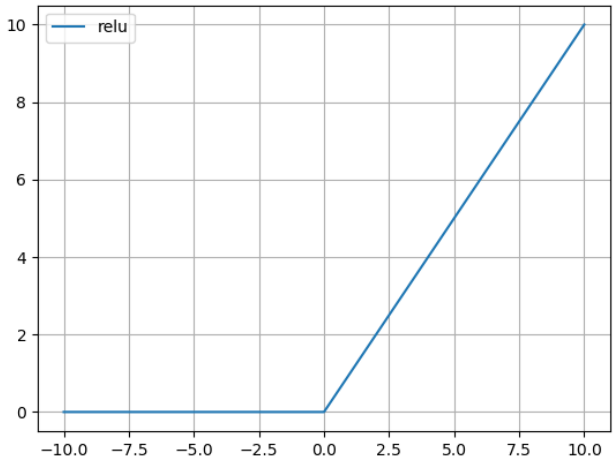


图 2.5 ReLU 函数曲线图

ReLU 激活函数的形状如图 2.5 所示, 该激活函数将输入值非正部分进行截断，保留非负输入值，使得网络中部分神经元失活。在网络反向传播更新参数时，ReLU 激活函数能够使梯度变化保持为 1，从而使得网络中的每个参数都能得到更新，降低了梯度消失的风险。因此现阶段 ReLU 激活函数被广泛使用。

池化层（pooling layer）一般是跟卷积层和激活层联合使用，对于池化层的使用不会改变原有的二维特征映射，其输出特征图对应着输入特征图，因此池化层其实就是一个特征二次提取且不改变通道数的过程，其通过降低特征映射的分辨率来获得具有空间不变性的特征[60]。在实践过程中，池化层会通过移动步长和改变信息窗口控制输出特征图的大小。现阶段主要使用的池化层分为两类，分别为平均池化（mean pooling）和最大池化（max pooling），平均池化就是将指定窗口内的像素点相加然后除以该窗口大小，从而得到一个实数，该实数就代表了这一窗口区域的特征，最大池化则是将指定窗口内的像素点进行比较，选择其中最大的像素点，该像素点就代表了这一窗口区域的特征。除了这两类，还有随机池化（stachastic pooling）等，但是较前两类使用的较少。对于平均池化和全局池化，两者使用场景存在差异化，其中最大池化适合用于特征图中特征分布较为稀疏的情况，而平均池化则不然。

输入经过卷积层、池化层提取特征后，所输出的特征图分辨率会较低，这是由于卷积和池化的每次操作会造成特征图的缩小。虽然分辨率降低了，但是每个像素点所代表的语义信息会更丰富，最后会通过全连接层进行特征映射，从而完成分类任务。全连接层就跟它的命名一样，会将前一层所有的神经元信息按照线性排列进行整合，从而获取前一层所有的语义信息。为了增加网络的非线性能力，全连接层后面也会跟激活层，一般来说激活函数为 ReLU 激活函数。在该激活函数之后便是网络的输出，对于分类任务而言一般采用 softmax 逻辑回归（softmax regression）进行分类，也就是常说的 softmax 层（softmax layer），对于二分分类问题，也可以采用 Sigmoid 函数进行分类。

对于卷积神经网络的参数更新主要采用反向传播算法，当数据集数据较少同时网络较为复杂的时候容易出现过拟合现象，过拟合现象的产生会导致网络训练效果很好而测试集测试结果不佳。因此为了避免过拟合现象的产生，研究者们采用了很多方法，其中最常用的主要是三个，分别为增加训练集、L2 正则化以及 Dropout 技术。增加训练集就是通过数据增强的方法增加样本数，L2 正则化就是对损失函数添加参数惩罚项，从而可以抑制网络的过拟合现象。L2 正则化如式 2-4 所示：

$$\Omega(\theta)=\frac{1}{2n}\sum_{i}\theta_i^2 \tag{2-4}$$

其中 θ_i 是网络中第 i 个待更新的参数，n 是训练集中的样本数。

Dropout 技术是现阶段抑制过拟合最广泛的技术，通过设置 Dropout 参数可以有概率性的使得某些神经元失效，从而避免了网络参数的更新过分依赖某几个神经元，使得网络具有很好的泛化能力。在本文中，主要使用 Dropout 技术抑制过拟合。

在卷积运算过程中，网络参数的更新会实时共享，因此需要训练的参数得到了降低，而参数的降低会减少网络的复杂性。除此之外，1×1 卷积降维、池化层的使用、Inception 思想的引入也极大的降低了网络参数量。现阶段越来越多的数据集趋于现实场景，特征提取较为复杂，为了进一步的提取特征，在设计网络的时候通常会考虑较为复杂的网络结构，采用更深的网络层数，需要训练学习的参数也相应的会增多，因此参数的降低尤为必要。总的来说，CNN 的感受野机制、权值共享和池化等降维操作使得模型有更少的连接和参数，从而更易于训练。

2.3.2 自注意力机制

自注意力机制（Self-Attention Mechanism），最早被用于视觉图像领域，该思想主要利用了人眼看事物的机制，人眼在感知外部事物时，不会将整体看全，而是根据自我需求关注那些需要特别关注的区域。对于人工智能而言，这些特别需要关注的区域就是注意力需要集中的区域，因此会将此区域特征凸显，而对于那些不重要或者无用区域信息会进行抑制。

自注意力机制的基本原理如下：首先将输入分解成多个键值对，如图 2.6 所示。键值对包括三个元素：键（Key，K）、值（Value，V）、查询（Query，Q）。K、V、Q 皆为初始输入特征矩阵 X 与一组 W_K、W_V、W_Q 权重矩阵相乘得到，从而得到了一个特征映射，公式如 2-5，2-6，2-7 所示。其中这一组权重矩阵一开始被设置为一个初始值，之后通过网络反向传播算法进行更新，从而使得 K、V、Q 得到了更新。其次计算键和值之间的相关性，该相关性通过 K 乘以 Q 的转置得到，然后计算键和值之间的注意力权重，该注意力权重通过将得到的注意力值进行放缩，使其值介于（0，1）之间获得，该步操作通过将相关性做 softmax 函数运算即可得到，最后将得到的注意力权重与对应的值加权求和，从而获取了具有不同注意力权重的特征图。

$$Q = W_q X \tag{2-5}$$

$$K = W_k X \tag{2-6}$$

$$V = W_v X \tag{2-7}$$

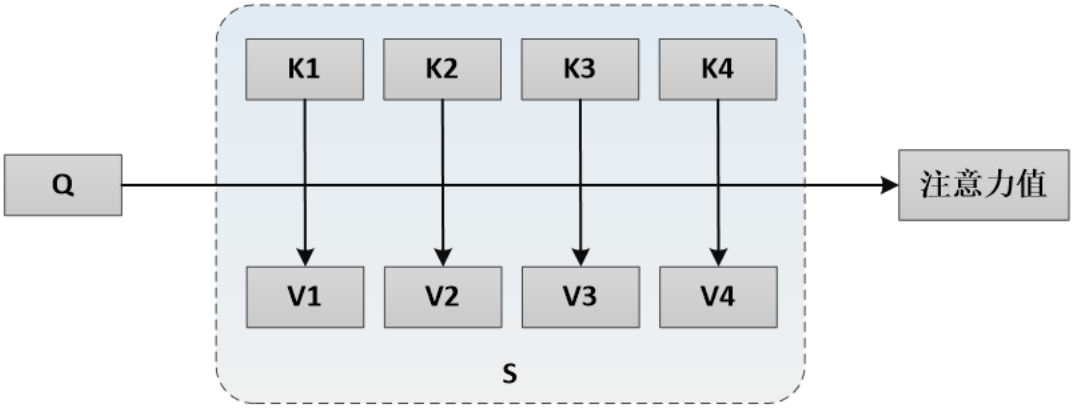


图 2.6 自注意力机制

自注意力机制和注意力机制不一样的在于注意力机制是用于目标（Target，T）的两个不同的 Q 之间，而自注意力机制是用于目标的单个 Q 内部元素之间，也可以理解为自注意力机制是注意力机制的一种特例情况。自注意力机制的公式如式 2-8 所示。

$$Attention(Q,K,V) = soft\max\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2-8}$$

其中除以维度 d_k 的开方是由于 Q 和 K 转置的内积过大，通过除以维度 d_k 的开方可以将其放缩，以便下一步的操作。

2.4 数据集

2.4.1 AFEW 数据集

数据集 AFEW，该数据集为动态视频数据集，同时作为竞赛级数据集，相比于实验室录制的数据集，其增加了一些干扰因素，其中干扰因素主要包括遮挡，像素点过低，背景变化等，因为这些干扰因素的存在，其更具现实性。同时该数据集已经将视频分帧进行切割从而形成了一个视频帧数据集，每个视频帧都有一个表情标签。其标签如图 2.7 所示，由于某些表情不具有分辨性，因此特地加入中性标签。其中 AFEW 数据分为三个数据分区：Train（773 个样本），Val（383 个样本）和 Test（653 个样本）。与此同时，作为竞赛级数据集，测试集并不对外开放，因此在本文中验证集即为测试集。

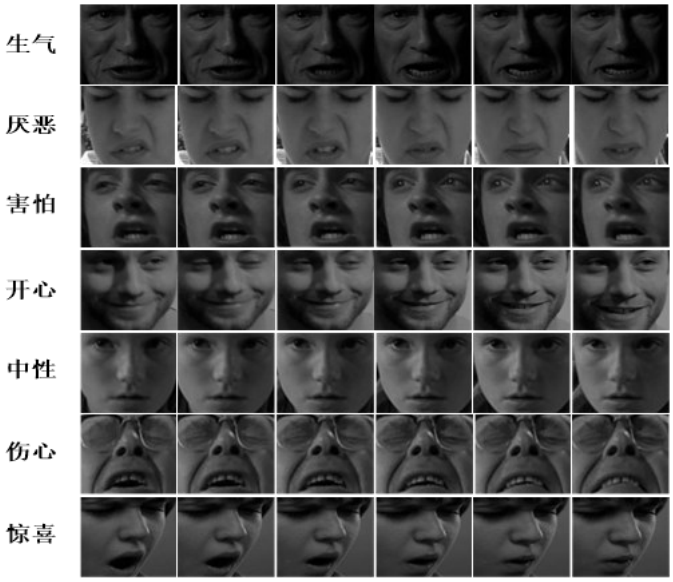


图 2.7 AFEW 数据集

2.4.2 CK+数据集

CK+数据集是表情识别常用的数据集之一，包含来自 123 个受试者的 593 个视频序列。序列的持续时间从 10 到 60 帧不等，其为动态视频数据集，它通过对视频进行截帧操作形成视频帧数据集，对于每个视频帧已经预先标明了表情标签，但是该数据集没有区分训练集、验证集和测试集，需要实验者自行划分。在本文中采用 5 折交叉验证方法划分数据集，其中部分数据如图 2.8 所示。

图 2.8 CK+数据集

2.4.3 SFEW 数据集

SFEW 数据集是由 AFEW 数据集中部分样本组成，样本与样本之间不存在相关性，属于静态图像数据集，该数据集已经被官方进行了划分，分别为训练集、测试集、验证集，其中 Train（958 个样本），Val（436 个样本）和 Test（372 个样本），同时每张图片都具有表情标签，

标签总共分为七种，但是由于是竞赛级数据集，因此测试集不对外公开，在本文中将验证集作为测试集。该数据集部分数据如图 2.9 所示。

图 2.9 SFEW 数据集

2.4.4 FER2013 数据集

FER2013 数据集是由大量无关人脸图片组成的静态图像数据集，其中按照一定比例分为训练集、测试集、验证集，每个图片都具有表情标签，标签总共为七类。其包含 28709 个训练图像，3589 个验证图像和 3589 个测试图像。该数据集为静态图像数据集。该数据集部分数据如图 2.10 所示。

图 2.10 FER2013 数据集

2.5 本章小结

在本章主要阐述了基于神经网络的视频表情识别基本流程，然后对于基本流程中的各个模块进行了详细介绍，包括数据增强、卷积神经网络、自注意力机制等，接下来介绍了一些常用的也是本文所用的人脸表情数据集，其中包括动态视频数据集：AFEW 数据集、CK+数据集，静态图像数据集：SFEW 数据集、FER2013 数据集。

第三章 基于特征增强卷积网络的视频表情识别研究

对于深度学习，现在使用的比机器学习更为普遍，这是因为其有着更好的特征加工能力同时能够适应深层次特征提取网络，因此成为了国内外学者研究的主流方向之一。对于表情识别领域而言，现在的研究者也开始采用深度学习进行特征提取以及表情分类。Moez Baccouche 第一个将卷积神经网络应用于特征提取[79]。2015 年 EmotiW 表情识别竞赛冠军获得者[80]采用深层次卷积级联进行特征提取从而将深层特征有效提取出来，最终建立特征与表情识别之间的相关性。由于卷积神经网络存在的一些优越性，Pooya Khorrami 考虑使用深度学习对视频进行情感识别[81]，但是卷积神经网络无法将相邻两帧之间的信息差异性关联起来，因此他们又采用了循环神经网络，该网络可以将相邻两帧之间的信息相关性提取出来，从来建立连续特征变换差异性，这一应用相比于单独使用卷积神经网络产生了更好的效果。

Zhang F 等人提出用生成对抗网络（Generative Adversarial Network，GAN）来自动生成具有表情的人脸图像[82]，从而扩大训练集，这是在图像预处理部分进行特征丰富。Chen S 等人提出标签分布学习（ Label Distribution Learning，LDL）[83]技术，将同一类别标签特征集中，减小类间间距。Wang K 等人提出自愈网络（Self-Cure Network，SCN）[84]来抑制表情判别不确定性，该网络主要运用自注意力机制来对样本进行预处理从而加强样本的表情特点。这些方法都是从图像本身出发进行特征增强，而忽略了网络带来的特征丢失问题。因此 Huai-Qian Khor 提出了一种丰富长期循环卷积网络（Enriched Long-term Recurrent Convolutional Network，ELRCN）用于细微表情识别，其主要通过通道级堆叠和特征级堆叠来增强对于人脸面部情感特征的提取。对于 ELRCN 网络，其就是采用 CNN-LSTM 进行特征提取以及表情识别，从未考虑特征之间的相关性以及侧重性，从而造成特征无区别度，同时还存在信息丢失的问题。之后陈乐等在 ELRCN 网络基础上提出了端到端增强特征神经网络[85]，该网络从视频多帧角度出发，通过双 LSTM 级联实现信息回溯来进行相邻帧信息的关联，但是忽略了视频单帧也存在信息丢失的现象。Zhang K 等提出了一种多信号卷积神经网络（Multi-Signal Convolutional Neural Network，MSCNN）[86]从静止帧中提取“空间特征”，该网络主要在单帧上进行特征加强，其利用监督学习不同损失函数达到类内差异缩小，类间差异增大效果，但是其只是在反向传播更新参数时利用了多种损失函数，一开始的信息损失仍然存在，只不过是将没有损失的信息采用多种损失函数组合凸显出来，并没有做到信息的保护，同时没有兼顾到视频多帧存在相关性的特点。

以上这些方法没有做到单帧和帧间信息的共同保护而只是在某一方面进行了特征加强，为了解决这些问题，本章从单帧和多帧两个角度着手进行特征增强，其中单帧采用的是浅层

特征与深层特征融合，浅层特征即在 VGG-16 网络中间层外延卷积模块，从而提取浅层的特征，深层特征即在 VGG-16 网络最后融合空洞卷积[87]（Dilated Convolution，DC)和 SENet（Squeeze-and-Excitation Networks，SENet）；多帧采用的是帧间注意力机制来提取帧与帧之间的相关性，从而将对于最终表情识别作用较大的帧凸显，将作用不大的帧抑制。该方法在 AFEW 动态视频（Acted Facial Expressions in the wild）数据集、CK+动态视频数据集、SFEW 静态图像数据集、FER2013 静态图像数据集上得到了有效的验证。

3.1 基于特征增强卷积网络的视频表情识别网络框架

本章设计的表情识别模型如图 3.1 所示，共包括两个部分，分别为单帧特征增强网络和帧间特征增强网络，其中 FC 为全连接层。单帧特征增强网络作为动态视频数据集的单帧特征提取，帧间特征增强网络作为动态视频数据集的相邻帧表情渐变信息提取，对于模型中输出特征向量维度大小会在 3.3 节实验证明 2048 维效果是最好的。在整个网络框架中，单帧特征增强网络分为深层特征增强和浅层特征增强，网络结构如图 3.2 所示，帧间特征增强网络为帧间注意力机制。

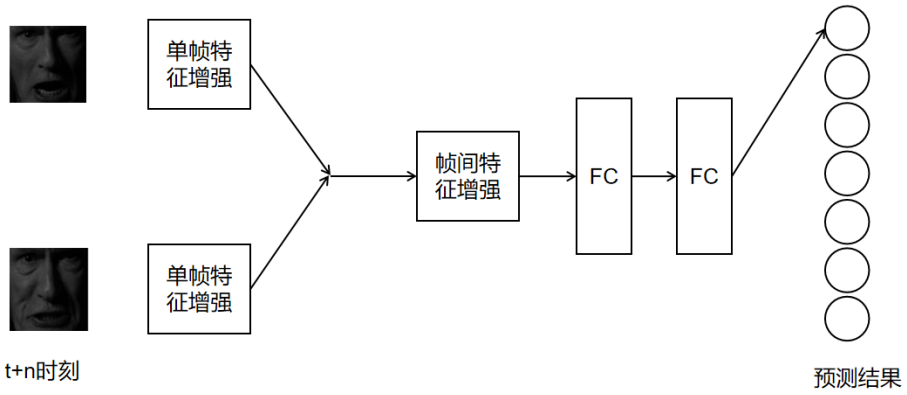


图 3.1 特征增强卷积网络的表情识别框架

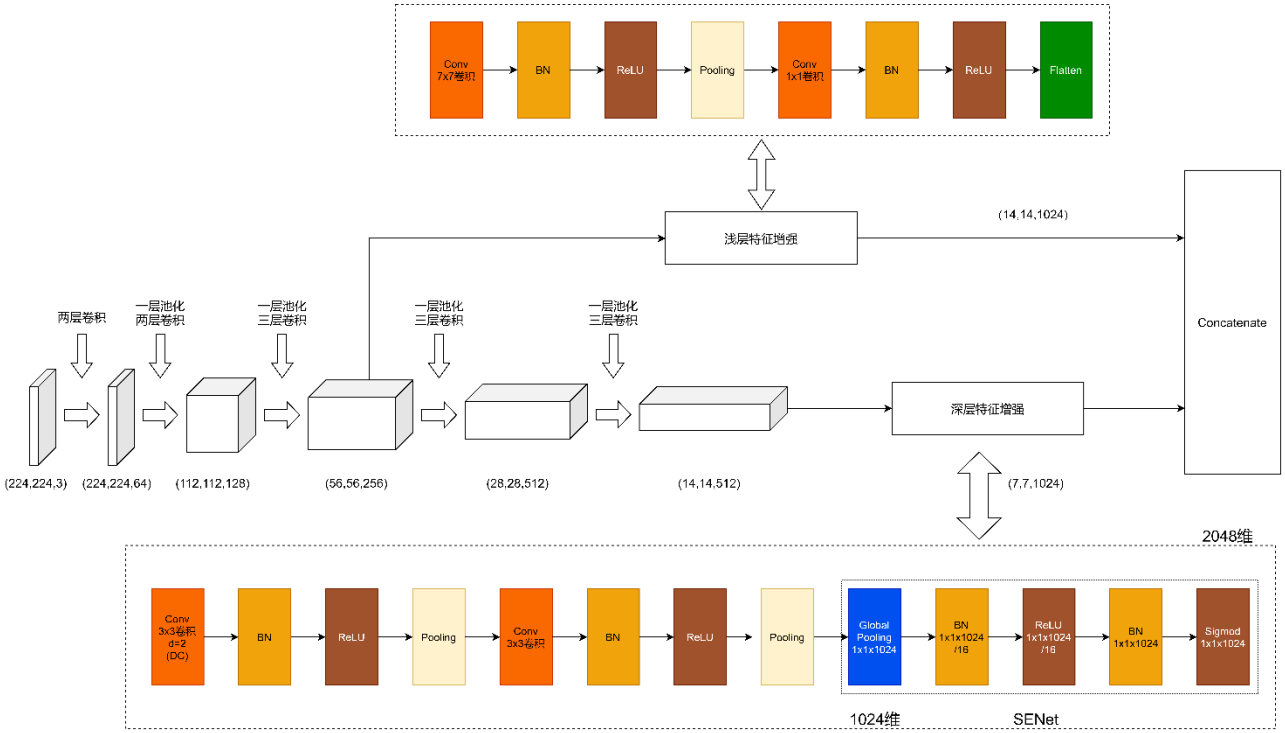
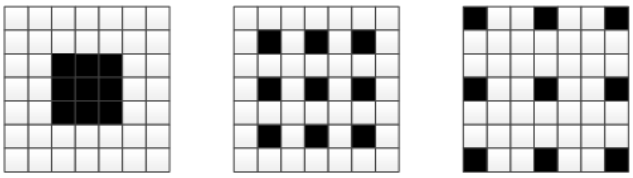


图 3.2 单帧特征增强网络

3.1.1 单帧特征增强

在单帧特征增强部分，主要应用了三种技术的融合，分别为空洞卷积（Dilated Convolution，DC)， SENet 以及浅层特征提取模块。

空洞卷积（Dilated Convolution，DC)，又称扩张卷积，最初是在算法“小波分解小波”中开发的，其基本原理就是在普通卷积的基础上，引入了一个扩张率（dilation rate，d)的超参数，该超参数定义了相邻卷积核各值的间距。卷积核大小为 3×3，扩张率为 d 的空洞卷积如图所示。从图 3.3（a）可以看出普通卷积是空洞卷积的一个特例，即为扩张率 1 的空洞卷积，以图 3.3（b）为例，扩张率为 2，让原本 3×3 的卷积核，在参数不变的前提下感受野增加到了 5×5。不仅如此，空洞卷积的应用也避免了 polling 所带来的下采样问题，polling 每次操作都会造成一半信息的丢失，这种无条件的一半信息的丢失会直接导致重要特征的损失，而使用空洞卷积代替 polling 操作就能避免这种问题。以图 3.3（c）为例，其主要是三个相邻像素点保留一个，而一般而言相邻像素点对于最终表情识别的作用是无差的，因此保留其一即可，从而可以保证强弱信息的结合。因此空洞卷积的应用有效解决了标准卷积所带来的内部数据结构损失以及空间层级化信息丢失的问题。本章对于空洞卷积的应用位置以及扩张率的大小尝试过多种可能性，最终在 VGG-16 网络最后一层应用扩张率为 2 的空洞卷积达到最好的识别率。



（a）d=1　　（b）d=2　　（c）d=3

图 3.3 空洞卷积

在实践过程中，卷积神经网络认为每一个像素点对于最终表情识别所起的作用都是一模一样的，然而从人眼角度来说，有些像素点所起的作用应该权重更多，而不是平均分配，因此在 CNN 的基础上融合了 SENet，该模块通过网络自动训练学习从而获取到每个特征通道的重要程度，然后依照这个重要程度去提升有用的深层特征并抑制对当前任务用处不大的深层特征。其结构如图 3.4 所示，其总共通过三个操作来重新标定 CNN 所输出的通道特征。首先是 Squeeze 操作，将每一个特征通道的里的像素点相加，然后除以特征通道大小，从而产生一个实数，该实数代表这一通道全部信息，其公式如 3-1 所示。其次是 Excitation 操作，在该步骤采用网络自学习机制生成参数 w，该值代表了每个通道对于最终表情识别所产生的影响因子，该值大小介于（0，1）之间，其公式如 3-2 所示。最后是 Reweight 操作，该操作就是将各个通道的权重加权乘到相应特征通道上，从而每个特征通道对于最终的表情识别结果做出了不同的贡献，实现了通道特征的重标定，公式如 3-3 所示。

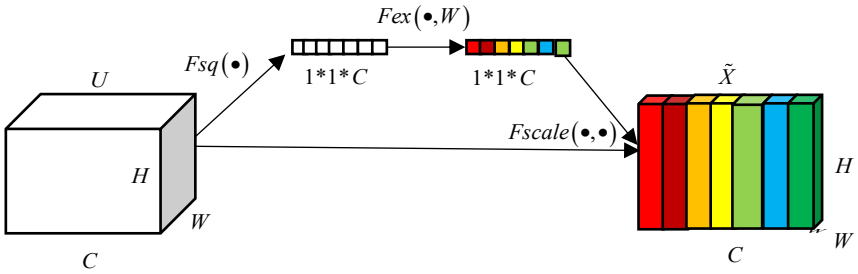


图 3.4 SENet

$$Z_c = F_{sq}(U_c) = \frac{1}{H \bullet W}\sum_{i=1}^{H}\sum_{j=1}^{W} U_c(i,j) \tag{3-1}$$

$$s = F_{ex}(z,W) = \sigma\left(g(z,W)\right) = \sigma\left(W_2\delta(W_1 z)\right) \tag{3-2}$$

$$X_c = F_{scale(U_c,S_c)} = S_c \bullet U_c \tag{3-3}$$

由于层数的提高必然会导致部分有用信息丢失，因此在注重深层特征的同时也需要关注浅层特征。在本章中，所采用的 CNN 模块为 VGG-16，此网络总共 16 层，虽然从层数而言并不是很深，但是在 VGG-16 最后一层出来的其实就是深层特征，而部分浅层特征已经被丢失，为了弥补这一缺陷，在 VGG-16 中某一层外延支路进行浅层特征提取，在尝试多次之后，最终选取 VGG-16 中间层外延支路。同时在实验过程中进行了多种尝试，最终确定以两层卷

积层（第一层为 7×7 的卷积层，第二层为 1×1 的卷积层）的级联最佳，其结构如图 3.5 所示。其第一层是一个 7×7 的卷积层，该卷积层保留更多的表情特征同时也保证了网络的感受野。批量归一化层是为了防止梯度爆炸和梯度消失。激活层使用 ReLU 非线性函数。第二层为 1×1 的卷积层，其是为了降维同时提高网络的表达能力，从而方便后面的特征融合。该浅层特征增强与深层特征增强互相融合，从而促进特征的进一步突出，为识别率的提高做出了卓越贡献。

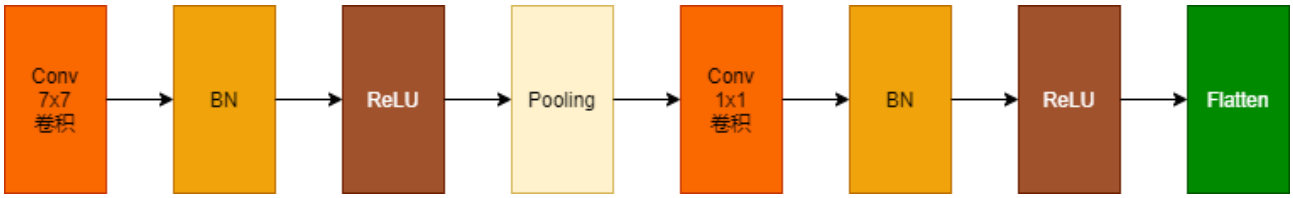


图 3.5 浅层特征增强

3.1.2 帧间特征增强

CNN 对于处理单帧图像十分有效，但是对于视频而言，相邻帧之间存在运动信息，同时由于表情是一个渐变的过程，前一帧的表情将会直接影响到下一帧的表情，因此单独的 CNN 将不适合处理这种关系，而循环神经网络（Recurrent Neural Networks，RNN）则更为有效。RNN 模型的循环特性可以使信息在网络中留存一段时间，从而可以建立相邻帧的表情变化关系。现阶段，主要采用 LSTM 关联多帧之间表情渐变信息。但是由于 LSTM 每个细胞中都有 4 个全连接层(MLP)，在时间跨度很大的情况下，会导致运算时间成几何级数上升。因此本章提出采用帧间注意力机制来代替 LSTM，其也能够将相邻帧的信息关联起来，该思想由 Fajtl J[88]首次提出，他指出使用自注意力机制来处理相邻帧，并给每一帧赋予不同的权重，从而将对于表情识别起关键作用的帧凸显出来，将对于表情识别误导性极高的帧抑制起来，促进最终表情识别率的准确性，其操作性跟基于通道间注意力机制类似。

3.2 评分标准

F1 评分是一种用来评判网络好坏的标准。它是通过准确率（precision）和召回率(recall)的数学组合形成的数学表达式，介于 0 到 1 之间：

$$F_1 = 2 \bullet \frac{precision \bullet recall}{precision + recall} \tag{3-4}$$

$$precision = \frac{TP}{TP + EP} \tag{3-5}$$

$$recall = \frac{TP}{TP + FN} \tag{3-6}$$

其中 TP（True Positive）为正确预测的数目，FP（False Positive）为将其他预测产生错误的类预测为本类的数目，FN（False Negative）为本应预测正确但错误的本类数目。

准确率（Accuray）：通常也作为网络好坏评判的标准：

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3-7}$$

其中 TN（True Negative）为将其他类预测为正确的数目。

3.3 实验结果与分析

在本章中，根据帧间注意力机制的结构特点，其要求 CNN 网络需同时输出多张连续人脸特征，因此模型每次输入 n 张人脸图像，每张人脸图像能够共享 CNN 网络权重进行特征提取。由于实验室条件有限，又为了避免内存溢出的问题，最终 n 值设置为 10，一次传入 10 张连续人脸图像。又为了增加相邻帧的特征共享，因此在实验过程中同一个视频中的相邻子视频段存在 5 帧的重合。VGG16 采用预训练权重 VGG-16-FACE 模型，初始学习率（Learning rate）为 1e-4，并且在训练过程中一直迭代下降。优化算法为随机梯度下降法（Stochastic Gradient Descent，SGD），动量（Momentum）参数值设置为 0.9。对于输入，采用了数据集自带的图像集，其中每张图像大小为 224×224。网络训练完毕，将测试集送入模型中，最后得出测试结果。实验代码使用 Keras 在 Ubuntu 16.4 系统下完成，主机配备 2 块 NVIDIA GTX 1080Ti。表 3.1 为本章特征增强卷积网络模型的参数设置。表 3.2 为本章特征增强卷积网络模型的训练参数信息。

表 3.1 特征增强卷积网络模型的参数设置

网络层	输出大小	参数
Conv_block_1	n×224×224×64	2层3×3卷积
Max_pooling_1	n×112×112×64	-
Conv_block_2	n×112×112×128	2层3×3卷积
Max_pooling_2	n×56×56×128	-
Conv_block_3	n×56×56×256	3层3×3卷积
Max_pooling_3	n×28×28×256	-
Conv_block_4	n×28×28×512	3层3×3卷积
Max_pooling_4	n×14×14×512	-
Conv_block_5	n×14×14×512	3层3×3卷积
Max_pooling_5	n×7×7×512	-
深层特征增强	n×7×7×1024	2层3×3卷积，2层池化
浅层特征增强	n×14×14×1024	1层7×7卷积，1层1×1卷积，1层池化
Concatenate	n×2048	-

表 3.2 特征增强卷积网络模型的训练参数设置

动量（Momentum）	0.9
初始学习率（Learning rate）	1e-4
优化算法	随机梯度下降法
epothes	50
Dropout	0.5
batchsize	10

对于各个数据集测试结果如表 3.4、表 3.8、表 3.9、表 3.10 所示，很明显可以看出通过特征增强 F1 分数和表情识别准确率都具有一定的提高，同时对于单帧特征增强和帧间特征增强进行了消融实验，证明了两者对网络性能的提高皆具有一定的作用。

3.3.1 AFEW 数据集和 CK+数据集

表 3.3 展示了特征增强卷积网络对于输出特征向量维度大小设置比较情况，通过该表确定输出特征向量维度设置为多少更合适。模型主要通过 1×1 卷积以及删减 VGG-16 卷积层来

改变通道的大小，表中 3072 维特征向量是采用 CNN 的输出 1024 维特征向量与浅层特征模块输出的 2048 维特征向量融合得出。表中 2048 维特征向量是采用 CNN 的输出 1024 维特征向量和浅层特征增强模块的输出 1024 维特征向量融合得出，表中 1536 维特征向量是采用 CNN 的输出 1024 维特征向量和浅层特征增强模块的输出 512 维特征向量融合得出。通过表可以看出对于不同的输出维度 F1 分数和识别率是不一样的，说明输出特征向量维度有一个适合的值，过大或者过小都会影响最后模型的效果。再次结合表可以看出输出特征向量维度为 2048 时识别率最高，但是 F1 分数是最低的，这是因为 AFEW 数据集中存在数据分布不均衡和小样本数据问题，造成部分表情误判现象，导致 F1 分数存在波动。但是从整体而言，对于表情识别更多的侧重于识别率，因此本模型倾向于选择 2048 维特征向量输出。而本模型最好的输出特征向量维度为 2048，因此接下来的实验比较都是采用 2048 维特征向量输出。

表 3.3 输出特征向量维度比较（%）

模型	F1 分数	准确率
特征增强卷积网络（1536）	40.31	44.65
特征增强卷积网络（2048）	39.44	45.19
特征增强卷积网络（3072）	40.91	44.92

表 3.4 展示了特征增强卷积网络在 AFEW 数据集上的仿真结果，同时将特征增强卷积网络进行了消融实验，分别展示了单帧特征增强特征卷积网络和帧间特征增强卷积网络的测试结果。为了证明特征增强卷积网络的有效性，将我们所提出的框架与 ResNet34-LSTM、DenseNet121-LSTM、端到端增强特征神经网络、ELRCN（VGG16-LSTM）进行了对比。第一行是 AFEW 数据集的官方基准线，第二行、第三行和第四行分别是 ResNet34-LSTM、DenseNet121-LSTM 和 ELRCN（VGG16-LSTM）在 AFEW 数据集上的 F1 分数和准确率，第五行是特征增强卷积网络的消融实验，第六行是特征增强卷积网络的 F1 分数和准确率。对于 ResNet34-LSTM 和 DenseNet121-LSTM，其中 ResNet34 和 DenseNet121 都没有在表情数据集上的预训练权重，因此采用了 FER2013 数据集对其进行预训练，但是效果相对而言不佳。同时 DenseNet121 有 121 层，层数较多，在训练到一定程度后出现过拟合现象，因此效果不如 ResNet34。事实证明，本章所提出的网络框架明显优于灵感来源的 ELRCN（VGG16-LSTM）网络以及其他优秀网络框架，F1 分数和准确率皆高于其他网络，同时对于所提出的单帧特征增强和帧间特征增强消融实验证明从单帧和多帧两个角度进行特征增强对于表情识别皆有促进作用。对于两个消融实验和特征增强卷积网络的混淆矩阵如表 3.5，3.6，3.7 所示。

表 3.4 AFEW 数据集测试结果比较（%）

模型	F1分数	准确率
AFEW baseline	-	38.81
ResNet34[89]-LSTM	31.59	36.10
DenseNet121[90]-LSTM	29.59	31.02
VGG16-LSTM[91]	31.00	40.00
VGG16-帧间特征增强卷积网络	39.37	44.39
特征增强卷积网络	39.44	45.19

表 3.5 AFEW 数据集（单帧特征增强卷积网络）测试结果

	生气	轻蔑	害怕	开心	中性	伤心	惊喜
生气	41	1	5	3	4	4	6
轻蔑	8	5	1	10	8	5	3
害怕	14	1	10	5	5	7	2
开心	3	1	0	52	1	4	1
中性	10	3	5	9	26	4	2
伤心	7	5	1	11	14	21	1
惊喜	16	1	6	9	5	1	7

表 3.6 AFEW 数据集（帧间特征增强卷积网络）测试结果

	生气	轻蔑	害怕	开心	中性	伤心	惊喜
生气	39	4	3	3	5	3	7
轻蔑	5	8	1	8	9	5	4
害怕	10	2	6	8	4	9	5
开心	4	0	0	53	2	2	1
中性	4	3	5	10	31	5	1
伤心	4	3	4	8	13	25	3
惊喜	17	2	3	5	9	5	4

表 3.7 AFEW 数据集（特征增强卷积网络）测试结果

	生气	轻蔑	害怕	开心	中性	伤心	惊喜
生气	32	3	7	3	4	7	8
轻蔑	4	8	3	5	7	10	3
害怕	6	5	9	6	2	13	3
开心	2	0	0	51	2	7	0
中性	2	4	7	8	30	8	0
伤心	3	3	4	7	10	31	2
惊喜	10	3	7	4	10	6	5

CK+数据集也是动态视频数据集常用的一种数据集，为了证明特征增强卷积网络的有效性，在该数据集上也进行了仿真对比。

CK+数据集是比较特殊的数据集，其没有具体区分训练集、验证集以及测试集，因此我们需要手动进行区分。本章采用了 5 折交叉验证方法，将整个 CK+数据集中的数据随机等概分为五份，随机将其中四份作为训练集，将剩余一份作为测试集，同时为了公平起见，实验五次。同时由于 CK+数据集数据量较少，因此为了增强样本数，降低过拟合现象，对数据进行了数据增强，扩充了数据。

在 CK+数据集上的测试结果如表 3.8 所示，对于所比较的网络框架与在 AFEW 数据集上的比较网络是一致的，同时为了验证帧间角度特征增强的有效性，在 CK+数据集上进行了消融实验，事实证明，针对视频数据，采用帧间特征增强，表情识别准确率有一定的提高，同时对于整个特征增强卷积网络的识别效果也做了实验，证实了网络的优越性。

表 3.8 CK+数据集测试结果比较（%）

模型	准确率
ResNet34[89]-LSTM	94.14
DenseNet121[90]-LSTM	91.09
VGG16-LSTM[91]	95.71
VGG16[77]-帧间特征增强卷积网络	97.95
特征增强卷积网络	98.38

3.3.2 SFEW 数据集和 FER2013 数据集

本章提出的特征增强卷积网络分为单帧特征增强和帧间特征增强，其中单帧特征增强作为静态图像数据集特征提取的主干网络，也可以作为动态视频数据集的单帧特征提取，帧间

特征增强需要输入之间存在相关性，因此只适用于动态视频数据集。单帧特征增强作为特征增强卷积网络的一部分，为了验证其有效性，因此在静态图像数据集 SFEW 数据集和 FER2013 数据集上进行实验仿真。同时为了与现有的一些优秀网络框架做对比，将 VGG16，ResNet34，DenseNet121，DenseNet169 在两个静态图像数据集上进行了测试，测试结果如表 3.9、3.10 所示。事实证明，我们所提出的特征增强卷积网络对于静态图像数据集上的表情识别具有良好的作用。

表 3.9 SFEW 数据集测试结果比较（%）

模型	准确率
VGG16[77]	53.40
ResNet34[89]	53.79
DenseNet121[90]	54.01
DenseNet169[90]	54.54
特征增强卷积网络	56.67

表 3.10 FER2013 数据集测试结果比较（%）

模型	准确率
VGG16[77]	72.80
ResNet34[89]	73.10
DenseNet121[90]	73.25
DenseNet169[90]	73.76
特征增强卷积网络	75.94

3.4 本章小结

在本章提出了一种特征增强卷积网络模型，从单帧和帧间两个角度进行特征增强，将 VGG-16、空洞卷积、SENet 和帧间注意力机制进行有效融合，在 AFEW 数据集、CK+数据集、SFEW 数据集和 FER2013 数据集进行了实验仿真，比较其在测试集上的 F1 分数和识别准确率，证明了该网络模型对于表情识别的优越性。

第四章 基于降参型特征增强卷积网络的视频表情识别研究

4.1 降参型网络

在深度学习刚兴起的时候，网络主要通过加深来尽可能的提取深层特征，但是在网络加深的同时，特征图也在变小，这一定程度上导致了梯度消失的问题。在本章主要讨论了如何解决网络加深带来梯度消失的问题，同时进一步提高网络的表情识别率。

对于深度学习，网络加深会提取到更多的深层特征，虽然特征图被缩小，但是每个特征点都具有极其丰富的语义信息。即使如此，这种网络加深并不是越深越好，其存在一个瓶颈，超过这个瓶颈会出现梯度消失问题，从而导致识别效果不升反降。ResNet 网络通过引入残差模块来解决因网络加深而带来的梯度消失问题，通过残差模块可以将网络的线层信息和深层信息连接起来，从而克服了网络的“退化”现象，这说明了残差模块的引入一定程度上会降低梯度消失的风险。深度可分离卷积（depthwise separable convolution）[92]主要用于一些轻量型网络中，如 mobilenet 网络，其将卷积操作分解为两部分，分别为深度卷积（Depthwise Convolution，DC）和逐点卷积（Pointwise Convolution，PC），其中深度卷积保证输出通道与原通道一致，逐点卷积用于对输入每个特征图进行卷积运算，最后将两者得到的输出融合，得到一个输出，相比于单个卷积层操作降低了网络参数量同时减少了网络训练时间。GoogleNet 中引入 Inception 模块，将网络深度用网络宽度替代，在保证模型特征提取不损失的情况下减少参数量，从而有效避免了梯度消失的风险。

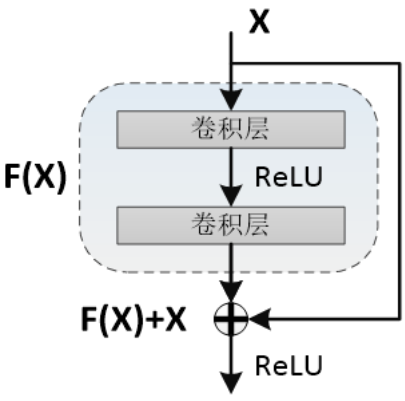


图 4.1 残差模块

残差结构如图 4.1 所示，其表达式为：

$$H(x)=F(x)+x \tag{4-1}$$

其中 H(x)是经过残差模块得到的输出，x 是该残差模块的输入，F(.)代表的是图中阴影部分的卷积映射操作，在经过该映射操作后得到 F(x)，因此可以看出 x 其实就是残差模块的恒等映射，是直接将输入跨层至输出，作为输出的一部分。残差模块就是通过该恒等映射解决网络的“退化”问题，在保证深层特征的同时将浅层特征不经过任何处理直接送到了深层，同时在网络反向传播更新参数的时候，由于浅层特征 x 直接传输到了深层，因此求导后保证了网络梯度大于等于 1，根据迭代原则，网络中的每个参数都能得到更新，避免了因梯度消失而产生的参数不更新问题。现阶段，残差模块为广大研究者使用，通过应用残差模块增加了网络的泛化能力。

深度可分离卷积分为深度卷积（Depthwise Convolution，DC）和逐点卷积（Pointwise Convolution，PC），其中深度卷积如图 4.2 所示，逐点卷积如图 4.3 所示。对于传统的卷积操作如图 4.4 所示，输入为 5×5×3，想要得到一个 3×3×4 的卷积，卷积核只需要设置为 3×3×4，同时步长为 1，填充为 0，该操作参数量为 108（3×3×3×4）。采用可分离卷积，其分为两步，首先深度卷积，该操作保证输出的通道数和输入的通道数保持一致，因此卷积核大小为 3×3×3，操作参数量为 27（3×3×3），接下来为逐点卷积，该操作改变输出通道数，使其变成网络所需通道数，因此卷积核大小为 1×1×4，操作参数量为 12（3×1×1×4），最终经过可分离卷积所需要的更新的参数量为 39（27+12），相比于传统卷积操作参数量降为原来的 36.1%，参数得到极大的降低，进而一定程度上避免了梯度消失的产生。

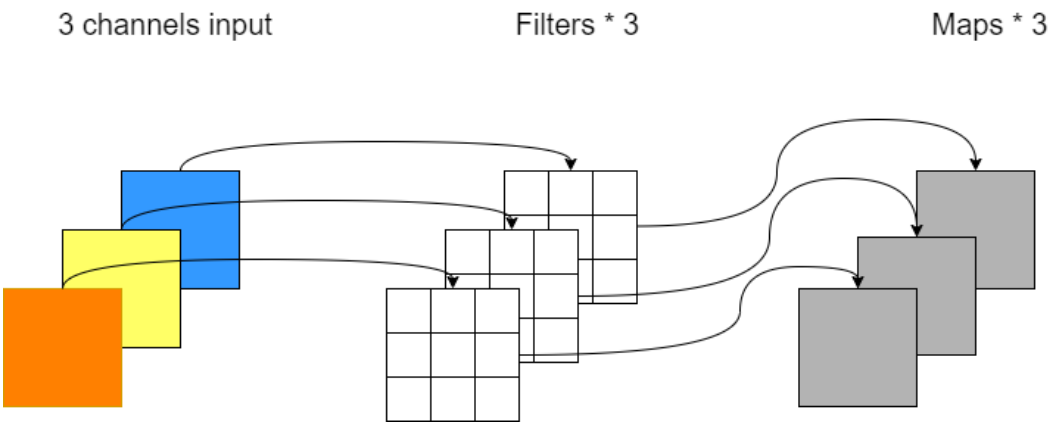


图 4.2 深度卷积

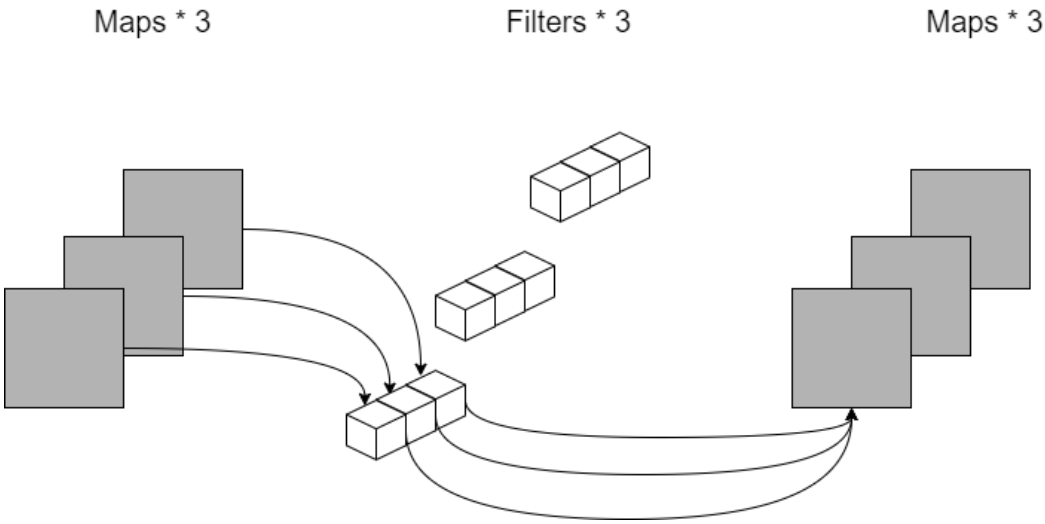


图 4.3 逐点卷积

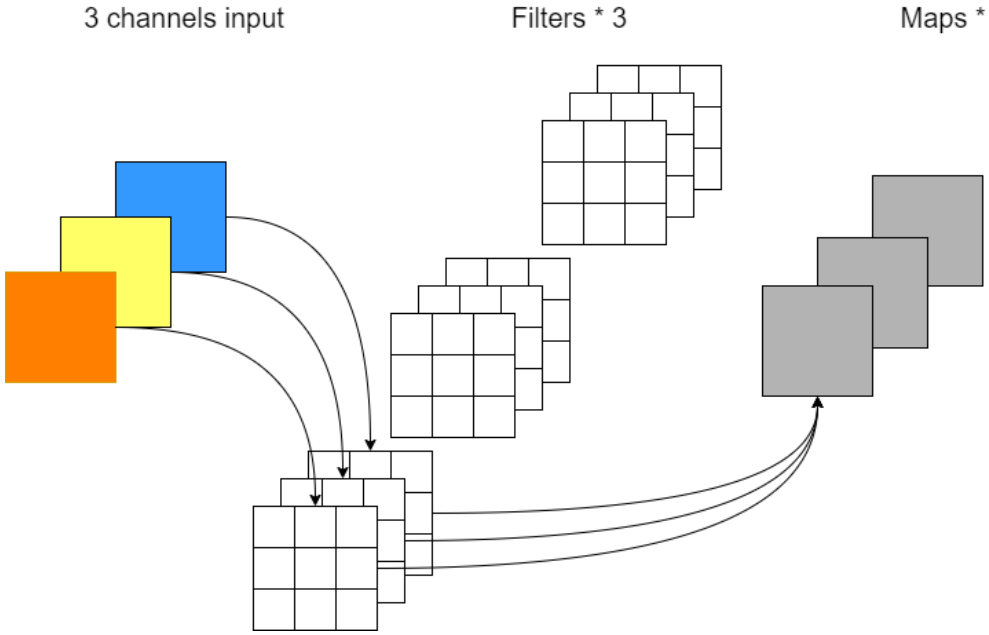


图 4.4 普通卷积

Google 首次提出并应用 Inception 思想来解决网络在增加到一定深度后达到性能饱和的问题，Inception 结构如图 4.5 所示。对于某一个卷积运算采用 Inception 结构来实现，通过不同的支路进行特征提取，最后将提取到的特征进行融合，得到一个输出，该输出的大小和单个卷积层所得到的输出是保持一致的。在减少参数量的同时极大的增加了网络的深度和宽度，同时参数的减少还降低了梯度消失的风险。在 Inception 结构中采用 1×1 卷积除了降维作用外还促进了特征的融合，增加了网络特征多样性，进一步提高了网络特征提取能力。

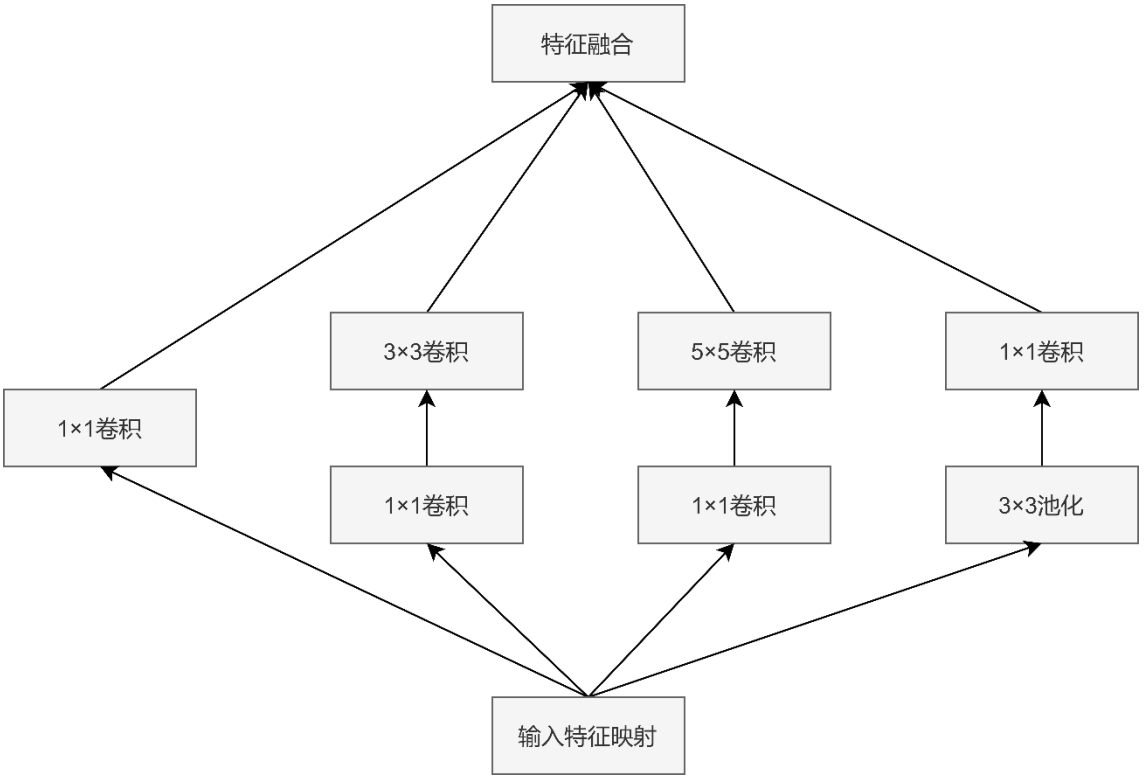


图 4.5 Inception 结构

4.2 基于降参型单帧特征增强卷积网络的视频表情识别网络框架

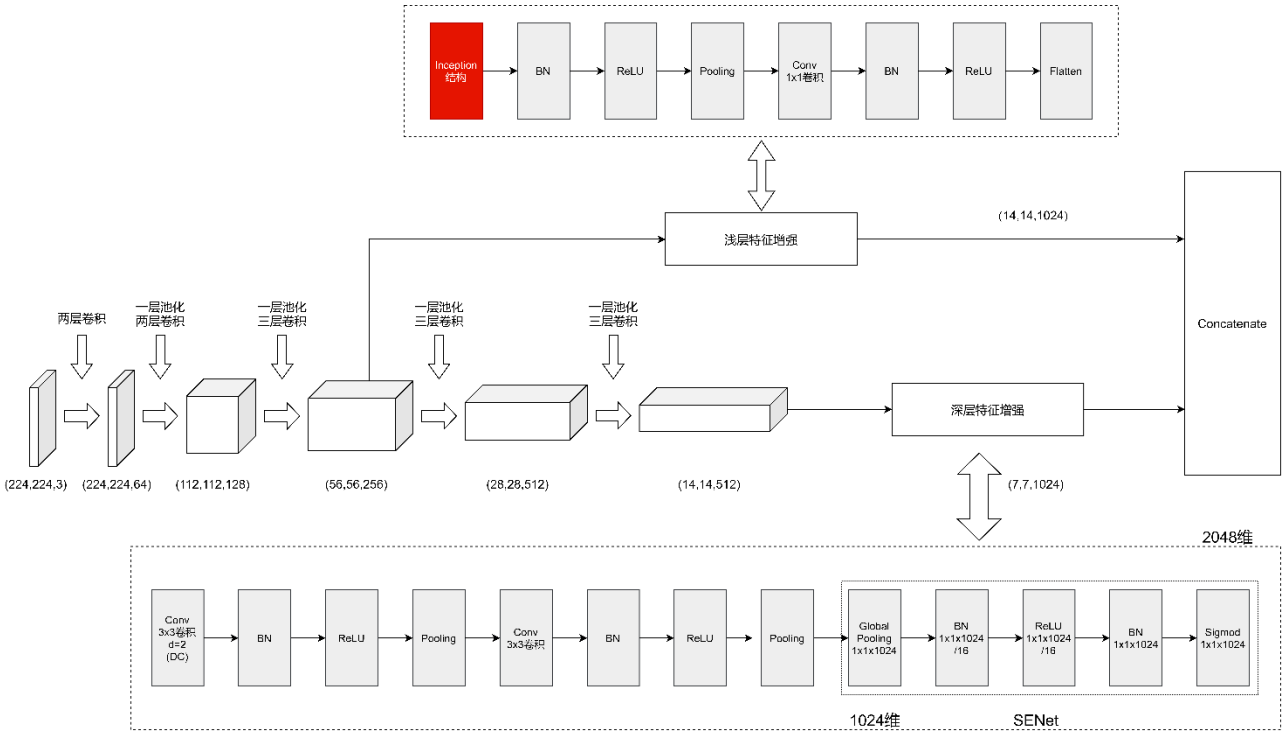


图 4.6 基于降参型单帧特征增强卷积网络框架

本章所提出的基于降参型单帧特征增强卷积网络框架如图 4.6 所示，多帧输入，首先通过 VGG 网络提取单帧图像表情信息，在提取单帧信息过程中，存在信息损失问题，因此采

用了第三章所提出的单帧特征增强模块进行特征补充，其中浅层特征增强模块采用降参型浅层特征增强模块替代，降低了网络的参数量；其次输入到帧间注意力机制中，将相邻多帧关联起来，提出多帧之间的渐变信息；最后分类输出，得到预测结果。

为了极大降低网络所需要更新的参数量，同时促进表情识别率的提高，本章提出了几种降参型浅层特征增强模块，以促进表情识别。本章所采用的基本网络框架为第三章所提出的特征增强卷积网络，在该基础网络上灵活应用 Inception 结构来进一步优化网络，在降低网络参数量的同时提高表情的识别准确率。同时在多个数据集上进行了验证，发现表情识别率确实存在提升。

在基础网络中，浅层特征增强采用了 7×7 大卷积块，而大卷积块的使用必然造成参数量的增加，如 7×7 卷积块的参数量 49，为了获得相同的感受野，可以采用两个 3×3 卷积块级联，此时参数量为 18，参数量得到了极大的减少。因此，现在使用小卷积块是深度学习应用的趋势，通过小卷积块的使用既加深网络的深度，又可以获得相同的感受野，保证模型的质量。因此本节采用 Inception 改造浅层特征增强中的 7×7 卷积块，同时对于 Inception 也做了几种变体，经过改造后的 Inception 如图 4.7、图 4.8、图 4.9 所示。图 4.7 为基础型 Inception V1，由 1×1，3×3，5×5 卷积组成，多种卷积组合获取到输入特征图的不同信息，其中 1×1 卷积用于获得“不稀疏”特征，3×3 卷积、5×5 卷积用于获得“稀疏”特征，从而丰富了特征。图 4.8 是在图 4.7 的基础上进行了改造，图 4.7 中 5×5 卷积相对而言还是一个大卷积块，为了进一步降低网络中的参数量，采用了两个 3×3 卷积级联替代 5×5 卷积，在相同感受野的前提条件下参数得到进一步降低。图 4.9 是采用两个 3×3 卷积并联的方式来替代 5×5 卷积，具有相同感受野，相对于图 4.8 参数量又得到了降低。

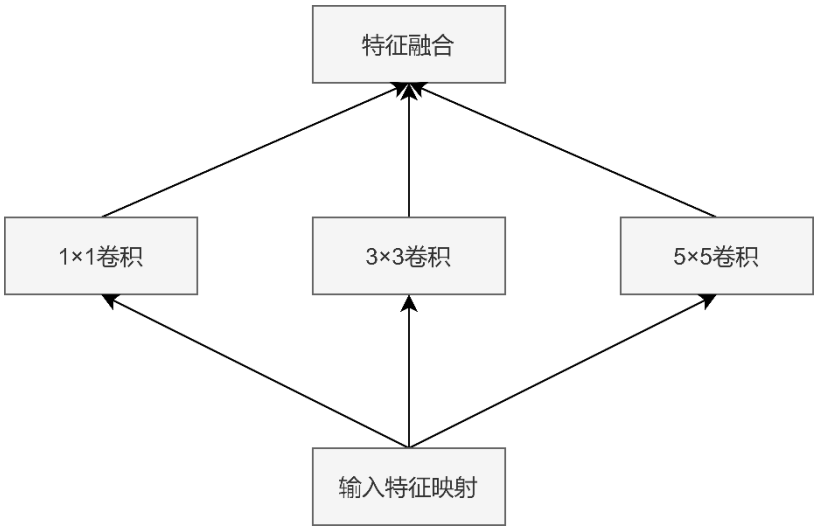


图 4.7 基础型 Inception V1

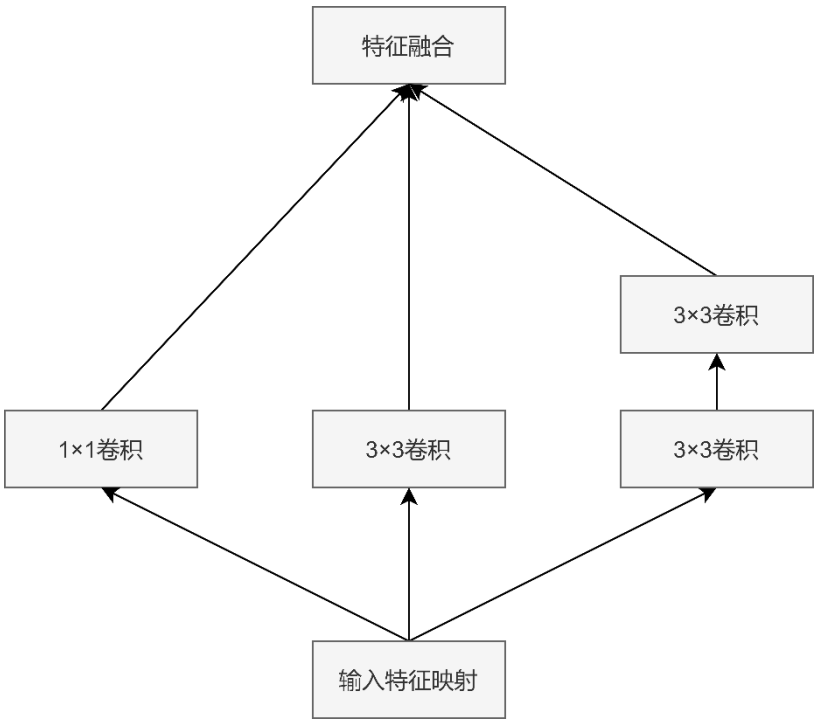


图 4.8 串联型 Inception V1

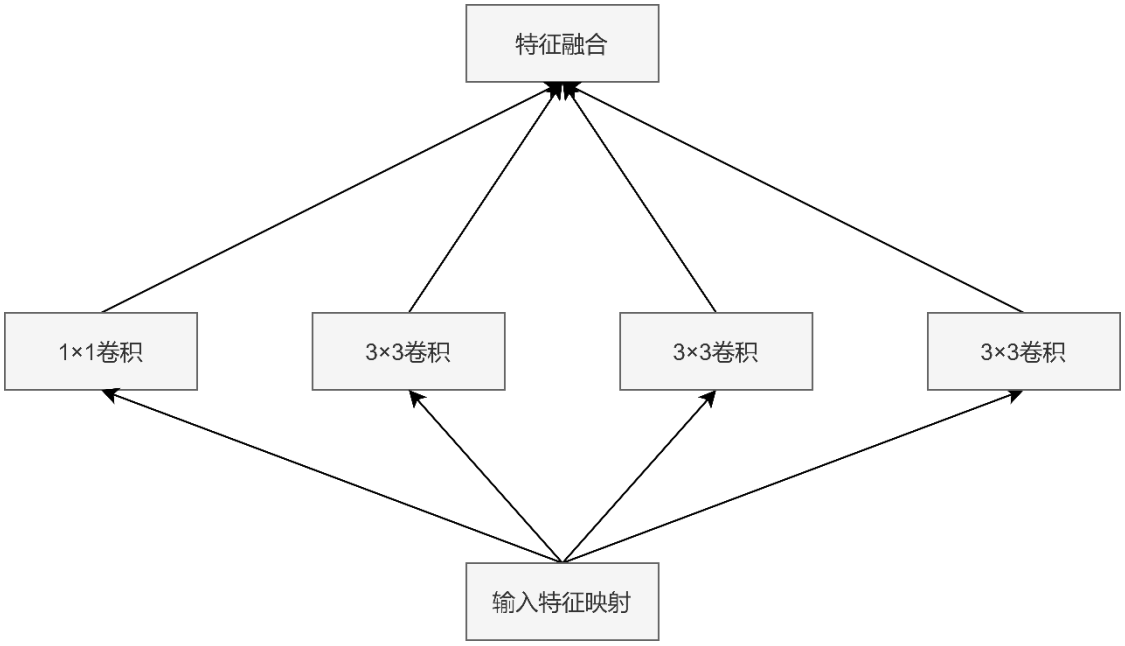


图 4.9 并联型 Inception V1

对于第三章所提出的特征增强卷积网络，其只有在浅层特征增强部分采用了大卷积块，因此对于 Inception 及其变体的使用主要是在浅层特征增强模块，主干网络和第三章一模一样，同时浅层特征增强模块的位置也和第三章所应用的位置保持一致。图 4.10 为降参型浅层特征增强模块（基础型 Inception V1），图 4.11 为降参型浅层特征增强模块（级联型 Inception V1），图 4.12 为降参型浅层特征增强模块（并联型 Inception V1）。

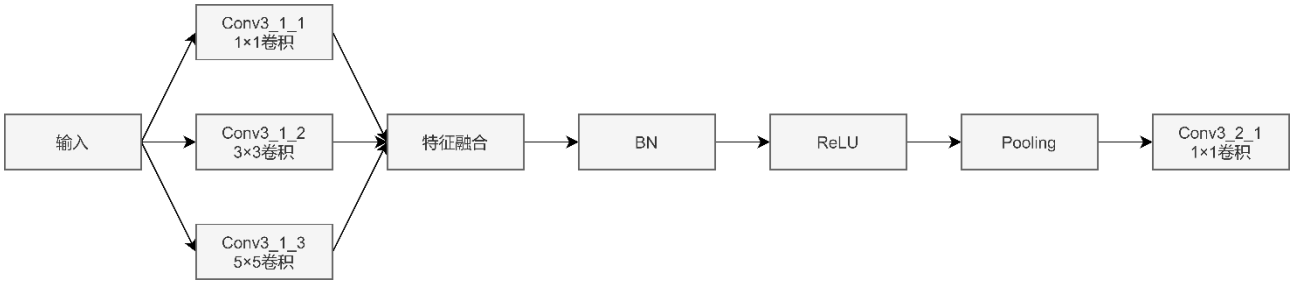


图 4.10 降参型浅层特征增强模块（基础型 Inception V1）

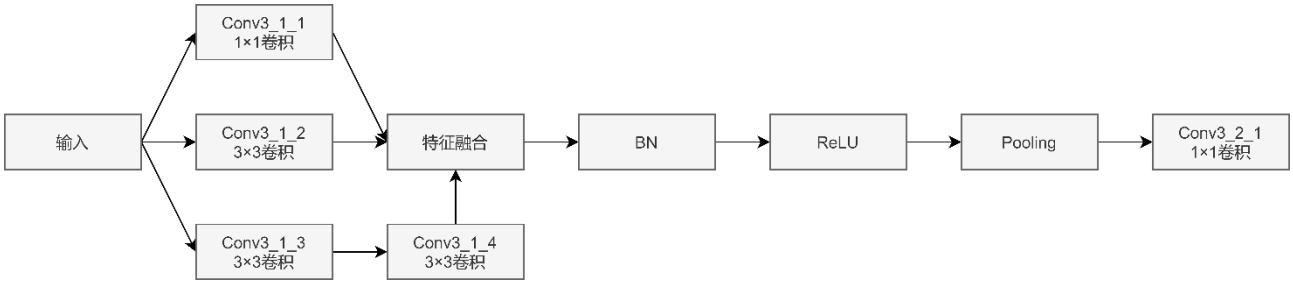


图 4.11 降参型浅层特征增强模块（级联型 Inception V1）

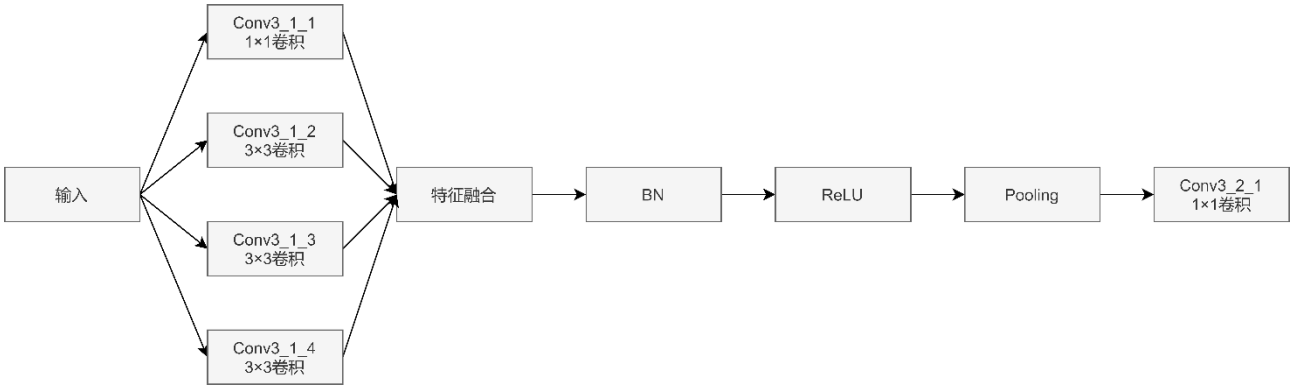


图 4.12 降参型浅层特征增强模块（并联型 Inception V1）

4.3 实验结果与分析

对于本章所提出的降参型浅层特征增强模块，为了验证其优越性，同时在动态视频数据集：AFEW 数据集、CK+数据集，静态图像数据集：SFEW 数据集、FER2013 数据集上进行仿真实验，结果表明在网络参数量减少的同时人脸表情识别率得到进一步的提高。实验代码使用 Keras 在 Ubuntu 16.4 系统下完成，主机配备 2 块 NVIDIA GTX 1080Ti。

表 4.1 降参型浅层特征增强模块参数设置（基础型 Inception V1）

网络层	输出大小	参数
Conv3_1_1	28×28×256	1×1 卷积
Conv3_1_2	28×28×512	3×3 卷积
Conv3_1_3	28×28×256	5×5 卷积
特征融合	28×28×1024	-
Pooling	14×14×1024	-
Conv3_2_1	14×14×1024	1×1 卷积

表 4.2 降参型浅层特征增强模块参数设置（级联型 Inception V1）

网络层	输出大小	参数
Conv3_1_1	28×28×256	1×1 卷积
Conv3_1_2	28×28×512	3×3 卷积
Conv3_1_3	28×28×256	3×3 卷积
Conv3_1_4	28×28×256	3×3 卷积
特征融合	28×28×1024	-
Pooling	14×14×1024	-
Conv3_2_1	14×14×1024	1×1 卷积

表 4.3 降参型浅层特征增强模块参数设置（并联型 Inception V1）

网络层	输出大小	参数
Conv3_1_1	28×28×256	1×1 卷积
Conv3_1_2	28×28×256	3×3 卷积
Conv3_1_3	28×28×256	3×3 卷积
Conv3_1_4	28×28×256	3×3 卷积
特征融合	28×28×1024	-
Pooling	14×14×1024	-
Conv3_2_1	14×14×1024	1×1 卷积

表 4.4 网络参数量比较

网络	参数量	参数计算	参数量比较
浅层特征增强模块	50176×256	7×7×1024×256	-
降参型浅层特征增强模块（基础型Inception V1）	11264×256	1×1×256×256+3×3×512×256+5×5×256×256	22.43%
降参型浅层特征增强模块（级联型Inception V1）	9472×256	1×1×256×256+3×3×512×256+3×3×256×256+3×3×256×256	18.87%
降参型浅层特征增强模块（并联型Inception V1）	7158×256	1×1×256×256+3×3×256×256+3×3×256×256+3×3×256×256	14.27%

表 4.1、表 4.2、表 4.3 为降参型浅层特征增强模块参数设置，表 4.4 为网络参数量比较，很明显可以看出相较于原浅层特征增强模块，参数量都得到了一定的减少，同时降参型浅层特征增强模块（并联型 Inception V1）参数量是减少最多的。

整个网络采用随机梯度下降算法进行参数的更新以及优化，动量常数设置为 0.9，为了尽可能快的达到网络最优，采用了预训练权重，网络初始学习率为 1e-5，该学习率会随着训练轮数增加不断衰减，以使得各参数能够达到全局最优。为了防止网络过拟合，同时配合采用早停法（Early Stopping），网络图像输入大小统一为 224×224。表 4.5 为本章网络模型的训练参数信息。

表 4.5 网络模型的训练参数设置

动量（Momentum）	0.9
初始学习率（Learning rate）	1e-5
优化算法	随机梯度下降法
epothes	50
Dropout	0.5
batchsize	10

4.3.1 AFEW 数据集和 CK+数据集

本章是基于第三章浅层特征增强模块大卷积块的使用造成参数量过多从而易导致梯度消失问题而做的改进，因此模型的输入仍然是连续十帧输入，同时前一次输入和后一次输入存

在五帧重合。

本章首先基于 AFEW 数据集进行验证对比。考虑到浅层特征增强模块大卷积块的使用使得网络需要训练学习的参数量过多，为了在相同感受野的情况下尽可能降低网络的参数量，我们提出了降参型浅层特征增强模块，通过 Inception 结构降低网络中的参数量，同时也能起到不改变感受野的效果。

表 4.6 AFEW 数据集测试结果比较（%）

模型	F1 分数	识别率
AFEW baseline	-	38.81
ResNet34[89]-LSTM	31.59	36.10
DenseNet121[90]-LSTM	29.59	31.02
VGG16-LSTM[91]	31.00	40.00
特征增强卷积网络	39.44	45.19
降参型特征增强卷积网络（基础型 Inception V1）	40.84	45.72
降参型特征增强卷积网络（级联型 Inception V1）	41.59	45.99
降参型特征增强卷积网络（并联型 Inception V1）	41.13	46.25

本章在 AFEW 数据集上的验证结果如表 4.6 所示，对于三种 Inception，我们分别实验仿真得到了其 F1 分数和识别率，降参型增强卷积网络（基础型 Inception V1）F1 分数为 0.4084，识别率为 0.4572，降参型增强卷积网络（级联型 Inception V1）F1 分数为 0.4159，识别率为 0.4599，降参型增强卷积网络（并联型 Inception V1）F1 分数为 0.4113，识别率为 0.4625，很明显可以看出识别率在提高，这说明网络在不改变感受野的同时需要训练学习的参数量减少会促进网络对于表情的识别效果。同时基于降参型浅层特征增强网络效果明显优于如 ELRCN（VGG16-LSTM）等优秀网络框架。对于 Inception 而言，其具有三种类型的卷积核（1×1，3×3，5×5），可以进行不同尺度的特征提取，从而获得多尺度信息，将多尺度信息融合可以获得更好的图像表征。同时结合整个表可以看出并联型 Inception 效果明显优于基础型和级联型，这是因为相较于基础型和级联型，在保持 Inception 结构的同时参数量得到了进一步的降低，进而促进了识别率的提高。基于该表所展示的数据，接下来在其他数据集上的测试都基于降参型增强卷积网络（并联型 Inception V1）。

在 CK+数据集上的验证结果如表 4.7 所示，相对于第三章提出的增强卷积网络，三种 Inception 变形结构识别率均在提高，但是由于增强卷积网络的识别率已经到 98.38%，因此提高的并不是很多。

表 4.7 CK+数据集测试结果比较（%）

模型	识别率
ResNet34[89]-LSTM	94.14
DenseNet121[90]-LSTM	91.09
VGG16-LSTM[91]	95.71
特征增强卷积网络	98.38
降参型特征增强卷积网络（并联型 Inception V1）	98.59

4.3.2 SFEW 数据集和 FER2013 数据集

本章主要改进点在单帧特征增强部分，该部分也是静态图像数据集特征提取的重要网络，因此该改进点对于静态图像数据集也具有一定的作用。为了验证对于静态图像数据集的作用，我们在 SFEW 数据集和 FER2013 数据集上进行了验证，具体如表 4.8、表 4.9 所示。通过 4.3.1 节可以看出降参型特征增强卷积网络（并联型 Inception V1）效果是最好的，因此在静态图像数据集上的测试以降参型特征增强卷积网络（并联型 Inception V1）为模型进行测试。通过表可以看出，基于 Inception 的网络识别率有一定的提高，这是因为 Inception 结构含有三种类型的卷积核，分别为 1×1、3×3、5×5，多种卷积核进行特征提取，从而获得多尺度信息，最后多尺度信息融合等价于 7×7 卷积操作，这种多尺度信息关注到了不同区域特征，使得特征更加丰富。

表 4.8 SFEW 数据集测试结果比较（%）

模型	识别率
VGG16[77]	53.40
ResNet34[89]	53.79
DenseNet121[90]	54.01
DenseNet169[90]	54.54
特征增强卷积网络	56.67
降参型单帧特征增强卷积网络（并联型 Inception V1）	57.21

表 4.9 FER2013 数据集测试结果比较（%）

模型	识别率
VGG16[77]	72.80
ResNet34[89]	73.10
DenseNet121[90]	73.25
DenseNet169[90]	73.76
特征增强卷积网络	75.94
降参型单帧特征增强卷积网络（并联型 Inception V1）	76.51

4.4 本章小结

本章针对第三章所提出的增强卷积网络中浅层特征增强模块使用大卷积块从而造成网络参数量过多的现象进行了优化，提出了降参型浅层特征增强模块，采用 Inception 结构改造浅层特征增强模块中 7×7 卷积块。该优化考虑网络参数量增多会造成网络存在梯度消失的风险同时在反向传播更新参数时存在参数未更新现象，本章的设计希望在保持感受野的同时能够降低网络中的参数量。对于 Inception 结构，主要采用三种变形，分别为基础型 Inception V1、级联型 Inception V1、并联型 Inception V1，这三种结构网络参数量逐步降低，对于改进点分别在动态视频数据集和静态图像数据集中进行了验证对比，事实证明在不改变感受野的同时降低网络参数量会促进对于人脸表情的识别率。

第五章 基于多头先验注意力机制的视频表情识别研究

自从自注意力机制被提出来之后，采用自注意力机制的神经网络在各个任务中的性能都有了一定的提升，但是自注意力机制存在两个问题，第一处理长序列的效率较低，这是由于自注意力机制的计算和内存复杂性，第二不假设对输入有任何结构性偏见，注意力权重图全靠网络根据输入特征图训练学习，存在信息误差性分布。在本文每次输入都是十帧，因此输入序列较长，同时输入即为 CNN 产生的特征图，然后通过网络自行训练，存在训练效果不佳现象，为了解决这两个问题，本章采用了 transformer 中多头注意力机制[93]替代原帧间注意力机制，同时在多头注意力机制的基础上提出了多头先验注意力机制思想。

5.1 多头注意力机制

多头注意力机制源于自注意力机制，自注意力机制的基本理论已经在 2.3.2 节进行了详细阐述，而多头注意力机制其实就是将多个自注意力机制组合起来，其公式如式 5-1、5-2 所示。

$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{5-1}$$

$$MultiHead\left(Q,K,V\right) = Concat\left(head_1,\cdots,head_h\right)W^0 \tag{5-2}$$

不同于自注意力机制，多头注意力机制需要维护多组 W_K、W_V、W_Q 权重矩阵，每组权重矩阵相互独立，而多组权重矩阵也就对应了多头的概念，假设权重矩阵为 8 组，那么就是八头注意力机制，公式如 5-1 所示。然后将多头所产生的输出特征矩阵进行横向拼接，由于多头注意力机制要求输入特征矩阵和输出特征矩阵大小保持一致，而多头产生的拼接特征矩阵大小不满足一致的要求，因此会设置一个需要学习的权重矩阵 W^0，该权重矩阵大小提前已经设置好，通过拼接特征矩阵和 W^0 相乘，从而得到了一个与输入特征矩阵大小一致的输出特征矩阵，公式如 5-2 所示。相比于自注意力机制，多头注意力机制含有多组权重矩阵，通过每组独立权重矩阵可以将输入特征图投射到不同的表示子空间上，因而可以为注意力层提供不同的“表示子空间”信息，同时每组权重矩阵，可以使得输入特征矩阵专注到不同位置的特征，对于不同位置的特征进行特征学习以及注意力权重赋予，最后将不同位置的特征矩阵图拼接从而获得整个输入特征矩阵的全部注意力权重图，在经过 W^0 的维度控制，从而得到与输入大小一致同时已经被赋予了注意力权重的特征矩阵。

5.2 基于多头先验注意力机制的视频表情识别网络框架

自注意力机制简单来说就是获得基于输入图像的注意力权重图，然后通过注意力权重图乘以输入图像特征图，最后获得被赋予注意力权重的特征图，对于多头注意力机制，其每个头的权重矩阵都是单独训练以及赋值，存在部分权重矩阵训练效果不佳，导致注意力权重无法达到较好的拟合效果，因此导致部分信息丢失的问题。为了解决该问题，我们提出采用 CNN 提取到的特征图为多头注意力机制做特征补充，通过 DenseNet 的跨层连接思想，将 CNN 出来的特征图与多头注意力生成的具有不同权重的注意力特征图级联或者融合，从而通过 CNN 出来的特征图为多头注意力机制做特征补充以及特征引导，有效弥补了信息损失的问题，对于该结构取名为多头先验注意力机制。

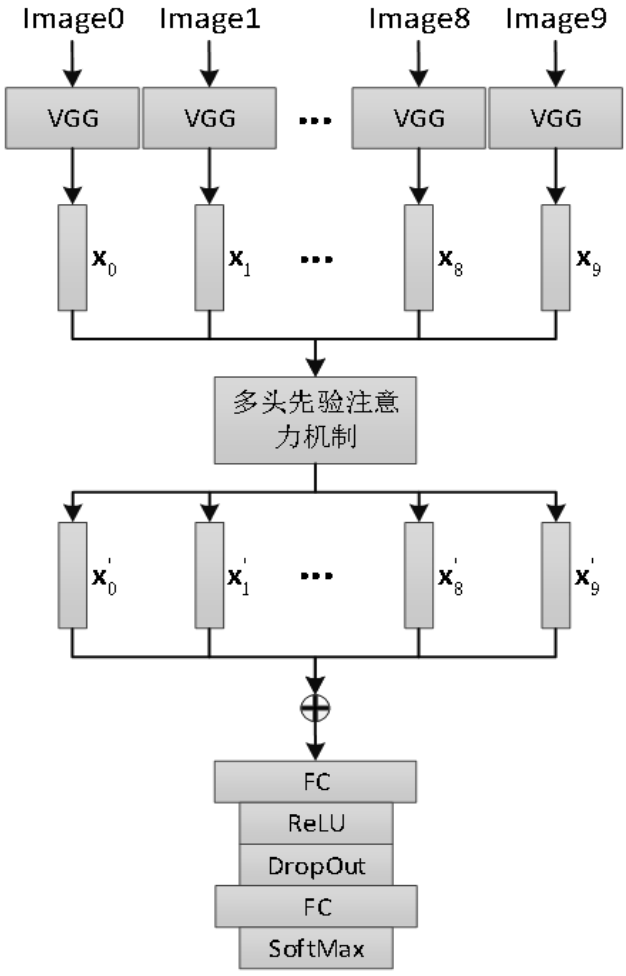


图 5.1 CNN-多头先验注意力机制框架

本章提出的 CNN-多头先验注意力机制框架如图 5.1 所示，连续视频帧输入到 CNN-多头先验注意力机制框架中，首先通过 VGG 网络提取单帧图像表情信息；其次输入到多头先验注意力机制中，将相邻多帧关联起来，提取多帧之间的表情渐变信息，多头注意力机制通过输入视频帧的相关性获得注意力权重图，然后将注意力权重图与输入特征图相乘得到具有注

意力权重的特征图，为了防止部分头的权重矩阵训练未达到较好的拟合效果，也就是说部分有用信息被抑制了，部分无用信息被凸显了，因此采用 CNN 出来的特征图与多头注意力机制产生的特征图融合或级联，用 CNN 出来的特征图补充特征以及引导注意力权重；最后得到一个与输入序列相同大小的输出特征序列。接下来设计了一个非线性回归模块，输出特征序列会经过全连接层完成特征映射，然后经过 ReLU 激活函数增加其非线性能力，同步使用 Dropout 技术，通过 Dropout 有效防止了网络过拟合现象的产生，在本框架中，Dropout 参数设置为 0.5，最后经过 Softmax 函数得到视频帧的预测结果。

5.3 多头先验注意力机制设计策略

对于多头先验注意力机制，主要依靠 CNN 出来的特征图为多头注意力机制做特征补充，对于多头注意力机制模块数目以及特征补充网络模型做了几种变形，从而验证哪种多头先验注意力机制的效果更好，这几种网络如图 5.2，5.3，5.4 所示。

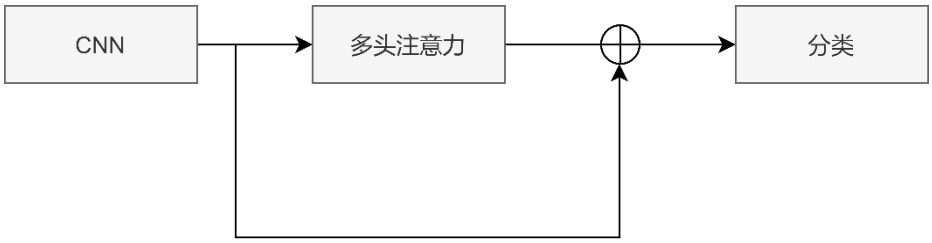


图 5.2 多头先验注意力机制（特征融合）

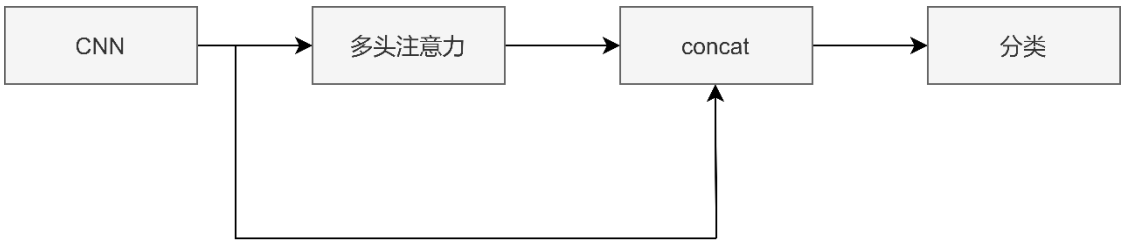


图 5.3 多头先验注意力机制（特征级联）

图 5.2，5.3 展示的是一个多头注意力机制与 CNN 出来的特征图级联或者融合的网络，这两种网络模型主要是为了验证是直接特征级联然后通过 1×1 卷积降维效果更好还是直接特征融合效果更好。在 5.4 节实验得出特征融合效果明显优于特征级联。

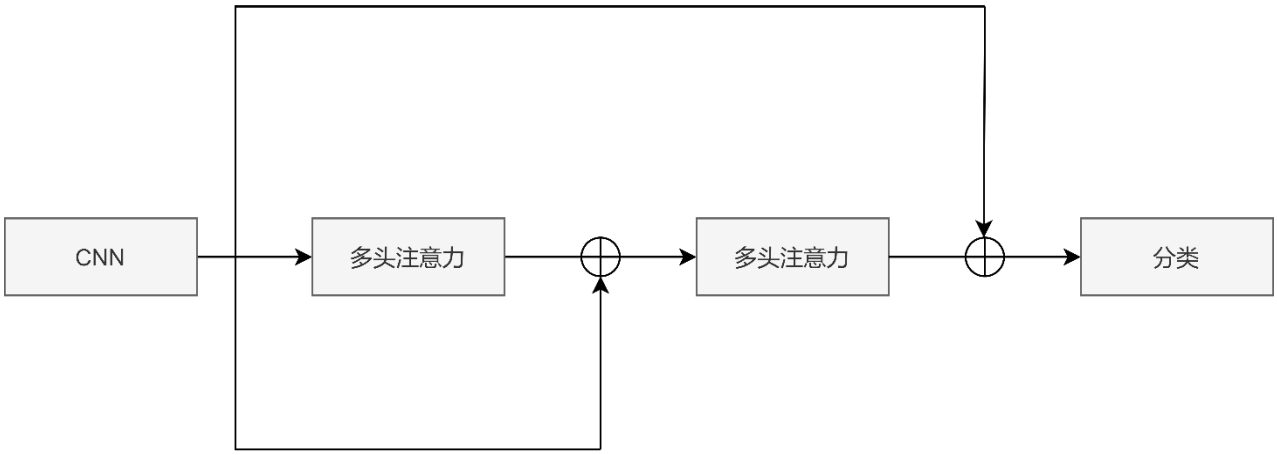


图 5.4 多头先验注意力机制（双特征融合）

通过图 5.2，5.3 可以得出到底是特征融合还是特征级联效果更好，5.4 节实验会充分证明特征融合效果更好。基于实验，接下来的多头先验注意力机制网络都以特征融合做出设计，从而设计了图 5.4 网络框架，该网络框架与图 5.2 比较，从而确定多头注意力机制数目以及跨层特征补充该如何设计才能使得特征得到较好的补充同时表情识别准确率更高。

5.4 实验结果与分析

为了验证本模型的有效性，本章主要在动态视频数据集：AFEW 数据集、CK+数据集上进行实验仿真，由于本章主要是在帧间特征提取模块做的优化，并未对单帧特征提取模块做出改变，因此不在静态数据集：SFEW 数据集、FER2013 数据集上进行仿真。实验代码使用 Keras 在 Ubuntu 16.4 系统下完成，主机配备 2 块 NVIDIA GTX 1080Ti。

表 5.1 为多头先验注意力机制模块和非线性回归模块的重要参数信息。多头先验注意力机制的输入为 n×1024 维人脸表情特征，对于 n 取值和前几章保持一致，为 10 帧输入，同时前后存在 5 帧重合。K、V、Q 大小皆为 1024×1024，相比于全连接操作，参数量进行了降低。经过多头先验注意力机制赋予注意力权重后会进入非线性回归模块，通过归一化、ReLU 激活函数增加非线性能力、Dropout 降低过拟合现象之后，将其送至全连接层，进而将特征映射为 7 维，最后通过 Softmax 进行表情分类。

该模型所使用的梯度下降算法为随机梯度下降算法，动量大小设置为 0.9。对于模型 VGG-16 使用了预训练权重模型，模型学习率初始值设置为 1e-4，随着网络训练次数的增加而减小。对于输入视频帧的大小设置为 224×224，代码编写基于 pytorch 框架。表 5.2 为本章网络模型的训练参数信息。

表 5.1 多头先验注意力机制参数设置

网络层	输出大小	参数
多头先验注意力机制	n×1024	K(1024×1024) V(1024×1024) Q(1024×1024)
Layer_norm	1024	-
ReLU	1024	-
DropOut	1024	0.5
FC	7	-
Softmax	7	-

表 5.2 网络模型的训练参数设置

动量（Momentum）	0.9
初始学习率（Learning rate）	1e-4
优化算法	随机梯度下降法
epothes	50
Dropout	0.5
batchsize	10

5.4.1 AFEW 数据集

表 5.3 展示的是用多头注意力机制直接替代帧间注意力机制，对于头的个数进行仿真实验，从而确定多头注意力机制中的多头设置为多少效果最佳。根据表 5.3 可以看出，头设置为 2，dropout 为 0.5 时识别率和 F1 分数是最高的，这是因为头数目较少情况下不仅能抑制网络过拟合现象的产生，而且又能增加多“表示子空间”信息。由于多头先验注意力机制中存在多头注意力机制，因此基于表 5.3 其头设置为 2，同时 dropout 设置为 0.5。

表 5.3 多头设置选择（%）

模型（CNN 为 VGG16[77]）	识别率	F1 分数
两头注意力机制[93]（dropout=0）	43.85	39.06
两头注意力机制[93]（dropout=0.5）	44.65	39.67
四头注意力机制[93]（dropout=0）	43.20	38.44
四头注意力机制[93]（dropout=0.5）	43.85	39.19
八头注意力机制[93]（dropout=0）	42.52	37.42
八头注意力机制[93]（dropout=0.5）	44.12	39.55

表 5.4 CNN-多头先验注意力机制模型测试结果（%）

模型	准确率	F1 分数
CNN-帧间注意力机制	44.39	38.12
CNN-多头注意力机制[93]	44.65	39.42
CNN-多头先验注意力机制（特征融合）	45.45	40.29
CNN-多头先验注意力机制（特征级联）	45.19	39.44

首先根据表 5.4 第一行和第二行数据可以看出用多头注意力机制替代帧间注意力机制表情识别准确率和 F1 分数都有一定的提高，这说明多头注意力机制在提取特征方面效果确实优于帧间注意力机制。其次根据表 5.4 第三行和第四行数据可以看出特征融合型比特征级联型效果更好，这是因为采用特征级联型会造成特征的空间信息损失，同时对于两种特征（一种为被赋予注意力权重的特征，一种为未赋予注意力权重的特征），两者级联只是单纯信息拼接，并没有达到特征补充的作用，因此效果不如特征融合型，基于此实验结果，接下来皆是采用特征融合型网络模型。

表 5.5 多头注意力机制个数设计测试结果（%）

模型	准确率	F1 分数
CNN-多头先验注意力机制（特征融合）	45.45	40.29
CNN-多头先验注意力机制（双特征融合）	45.81	40.92

最后结合表 5.4 和表 5.5 数据可以明显看出我们所提出的多头先验注意力机制效果优于帧间注意力机制以及多头注意力机制，同时对于多头先验注意力机制中多头注意力机制数目进行实验仿真得出多头注意力机制为两个时效果最好，CNN 出来的特征图与第一个多头注意力机制出来的被赋予注意力权重的特征图融合从而作为第二个多头注意力机制的输入，通过第二个多头注意力机制对其进行注意力权重赋予，再将 CNN 出来的特征图作为第二个多

头注意力机制的特征补充，从而进一步弥补了特征损失。

表 5.6 AFEW 数据集测试结果比较（%）

模型	准确率
AFEW baseline	38.81
ResNet34[89]-LSTM	36.10
DenseNet121[90]-LSTM	31.02
VGG16-LSTM[91]	40.00
CNN-多头先验注意力机制（双特征融合）	45.81

通过表 5.6 数据可以看出我们所提出的多头先验注意力机制明显优于现在的一些优秀网络框架，如端到端增强特征神经网络、ELRCN（VGG16-LSTM），从而论证了我们模型的有效性。

为了进一步体现网络设计的先进性，将传统的 VGG16 以及第三章提出的单帧特征增强网络、第四章提出的降参型单帧特征增强网络与第五章所提出的的多头先验注意力机制网络进行融合并进行实验比对，事实证明，多头先验注意力机制网络的应用能够在第三章以及第四章的基础上提高表情识别的准确率。

表 5.7 AFEW 数据集测试结果比较（%）

模型	准确率
CNN-多头先验注意力机制（双特征融合）	45.81
单帧特征增强-多头先验注意力机制（双特征融合）	45.98
降参型单帧特征增强-多头先验注意力机制（双特征融合）	46.52

5.4.2 CK+数据集

本章所做的改变在于用多头先验注意力机制替代帧间注意力机制，对于单帧特征提取模块未做改变，因此对于静态图像数据集的识别效果提高是没有作用的。本节在动态视频数据集 CK+数据集上进行实验仿真，从而进一步验证多头先验注意力机制对于表情识别准确率提升所起的积极作用。

结合表 5.8 可以看出我们所提出的多头先验注意力机制明显优于一些较为优秀网络模型框架，同时结合在 AFEW 数据集上的比较结果更加验证了我们所提出模型的优越性。

表 5.8 CK+数据集测试结果比较（%）

模型	准确率
ResNet34[89]-LSTM	94.14
DenseNet121[90]-LSTM	91.09
VGG16-LSTM[91]	95.71
CNN-多头先验注意力机制（双特征融合）	97.40

5.5 本章小结

本章基于帧间注意力机制处理长序列效果不佳以及无法提供“多空间信息”问题，提出采用 transformer 中多头注意力机制替代帧间注意力机制，同时实验仿真证明了多头注意力机制对网络性能提升所起的积极作用。同时基于多头注意力机制，提出了多头先验注意力机制，通过 CNN 出来的特征图为多头注意力机制做特征补充以及特征引导，从而避免了多头产生的训练效果不佳问题。对于提出的多头先验注意力机制在动态视频数据集 AFEW 数据集和 CK+数据集上进行了实验仿真，从而证明了模型的优越性。

第六章 总结与展望

6.1 本文工作总结

视频人脸表情识别作为现在比较热门的研究领域之一，对于促进非语言交流、人机交互具有划时代的意义，现阶段已经被广泛应用于课堂教学、自动驾驶等现实场景。由于大数据的发展，视频表情识别技术也由传统机器学习算法转向了深度学习算法，为此后期主要采用深度学习算法进行视频人脸表情识别研究，深度学习主要采用神经网络自适应提取特征，不需要手动提取，同时深度学习的训练需要大量数据，因此其适合处理大量数据，通过采用深度学习算法，视频人脸表情识别研究步入了一个新的里程。本文所采用的数据集是真实场景下产生的数据，存在如光照、背景、像素低等干扰因素，因此特征提取相对具有一定的难度。本文主要设计了特征增强卷积网络进行特征提取，同时基于特征增强卷积网络大卷积块参数量过大问题采用 Inception 结构进行优化，基于帧间注意力机制不适合处理长序列问题采用 transformer 中多头注意力机制进行优化，同时提出多头先验注意力机制的思想，在动态视频数据集和静态图像数据集验证了所提出以及所优化模型的可行性，具体工作如下：

（1）提出特征增强卷积网络，其中分为单帧特征增强和帧间特征增强，设计浅层特征增强模块，并引入空洞卷积和 SENet，达到增强单帧人脸表情特征的目的；采用帧间注意力机制实现多帧人脸表情特征增强，两者融合达到提高视频表情识别准确率的目的。对于该模型在四个数据集上进行了验证比对，证实了其模型的优越性。

（2）针对特征增强卷积网络中 7×7 大卷积块网络参数量过多的问题，采用 Inception 结构进行优化，提出降参型特征增强卷积网络。同时对于 Inception 在基础结构上提出了两种变体，分别为级联型 Inception 和并联型 Inception，加上原有的基础结构，共三种，对于三种结构的优越性在四个数据集上进行了验证比对，证实了在降低网络参数量的同时能够提高表情识别的准确率。

（3）针对特征增强卷积网络中帧间注意力机制不适合处理长序列问题，提出采用 transformer 中多头注意力机制代替帧间注意力机制，并针对因注意力权重赋予偏差造成信息丢失问题提出了多头先验注意力机制。该模型在 AFEW 数据集、CK+数据集上进行了仿真验证，证实了多头先验注意力机制能够提高人脸表情识别的准确率。

6.2 未来工作与展望

本文研究课题为基于视频的人脸表情识别研究，针对动态视频数据集和静态图像数据集搭建基于端到端的神经网络框架，以表情识别准确率和 F1 分数作为评判标准对网络进行优化，同时对于网络中的参数量进行降低。对于所采用的数据集，存在如光照、像素低、背景等干扰因素，这些干扰因素都影响着网络的表情识别准确率。在搭建网络以及优化网络过程中，发现还存在以下点可以进行研究：

（1）本文对于单帧特征提取主要采用 VGG16，帧间特征增强的设计、多头先验注意力机制的设计、空洞卷积的应用以及 Inception 结构的应用都是基于 VGG16 模型通过训练测试所得到的最优模型。但是现在也存在如 ResNet、DenseNet、GoogleNet 等基础 CNN 网络，这些网络相对于 VGG16 层数更多，所提取到的特征更丰富，而当前模型的设计不适合这些基础 CNN 模型，因此可以将其他的基础 CNN 模型作为新的研究方向。

（2）本文网络输入是视频帧，通过网络对视频帧进行特征提取以及分类，这属于单模态的领域，这种单模态特征相对而言信息具有局限性，因此可以考虑多模态，对视频、文本等其他输入设计特征提取网络，然后将多模态所获得的特征进行融合再进行分类，通过多模态的方式促进特征提取的丰富性，进一步增加表情的识别准确性。

（3）本文主要考虑深度学习对输入进行特征提取，完全依靠网络的特征提取能力，但是其实可以考虑增加一些传统特征，如 LBP 特征、Gabor 特征，通过传统特征对于网络提取到的特征进行补充，从而丰富了特征，对最后的分类能够起到一定的辅助作用。

参考文献

[1] D, C, P, et al. The Expression of Emotions in Man and Animals[J]. The American Journal of Psychology, 1981.

[2] Tian Y I , Kanade T , Cohn J . Recognizing action units for facial expression analysis[J]. IEEE Trans Pattern Anal Mach Intell, 2001.

[3] Lisetti C , Nasoz F , Lerouge C , et al. Developing multimodal intelligent affective interfaces for tele-home health care[J]. International Journal of Human - Computer Studies, 2003, 59(1-2):245-255.

[4] D'Mello S K , Graesser A C , Picard R W . Toward an Affect-Sensitive AutoTutor[J]. Intelligent Systems, IEEE, 2007, 22(4):53-61.

[5] Yannakakis G N , Togelius J . Experience-Driven Procedural Content Generation[J]. IEEE Transactions on Affective Computing, 2011, 2(3):147-161.

[6] Tian Y , Kanade T , Cohn J F . Facial Expression Recognition[M]. Engg Journals Publications, 2011.

[7] Wm A , Wh A . Facial emotion recognition using deep learning: review and insights[J]. Procedia Computer Science, 2020, 175:689-694.

[8] Wen G , Chang T , Li H , et al. Dynamic Objectives Learning for Facial Expression Recognition[J]. IEEE Transactions on Multimedia, 2020, PP(99):1-1.

[9] Ekman P , Friesen W V . constants across cultures in the face and emotion[J]. 2017.

[10] Matsumoto D . More evidence for the universality of a contempt expression[J]. Motivation & Emotion, 1992, 16(4):363-368.

[11] R. E. Jack, O. G. Garrod, H. Yu, R. Caldara, and P. G. Schyns, “Facial expressions of emotion are not culturally universal,” Proceedings of the National Academy of Sciences, vol. 109, no. 19, pp. 7241–7244, 2012.

[12] Ekman P E , Friesen W V . Facial action coding system (FACS)[J]. a human face, 2002.

[13] Gunes H, Schuller B . Categorical and dimensional affect analysis in continuous input: Current trends and future directions[J]. Image & Vision Computing, 2013, 31(2):120-136.

[14] Zeng Z , Pantic M , Roisman G I , et al. A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions[J]. IEEE Trans Pattern Anal Mach Intell, 2009, 31(1):39-58.

[15] Sariyanidi E, Gunes H , Cavallaro A.Automatic Analysis of Facial Affect: A Survey of Registration, Representation, and Recognition[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2015, 37(6):1113-1133.

[16] Martinez B, Valstar M F. Advances, challenges, and opportunities in automatic facial expression recognition[M]//Advances in face detection and facial image analysis. Springer, Cham, 2016: 63-100.

[17] Suwa M , Sugie N , Fujimora K. A preliminary note on pattern recognition of human emotional expression. 1978.

[18] Shan C, Gong S, Mcowan P W. Facial expression recognition based on Local Binary Patterns: A comprehensive study[J]. Image and Vision Computing, 2009, 27(6):803-816.

[19] Tai S L. Image Representation Using 2D Gabor Wavelets[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1996.

[20] Lowe D G . Distinctive Image Features from Scale-Invariant Keypoints[J]. International Journal of Computer Vision, 2004, 60(2):91-110.

[21] YAO A, SHAO J, MA N, et al. Capturing au-aware facial features and their latent relations for emotion recognition in the wild[C]//Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. ACM, 2015: 451-458.

[22] 张轩阁等. "基于光流与LBP-TOP特征结合的微表情识别." 吉林大学学报(信息科学版) 33.5(2015):516-523.

[23] G. Zhao and M. Pietikainen, “Dynamic texture recognition using local binary patterns with an application to facial expressions,” IEEE transactions on pattern analysis and machine intelligence, vol. 29, no. 6, pp. 915–928, 2007.

[24] Huang X, Zhao G, Hong X, et al. Spontaneous facial micro-expression analysis using spatiotemporal completed local quantized patterns[J]. Neurocomputing, 2016, 175: 564-578.

[25] 卢官明等. "基于LBP-TOP特征的微表情识别." 南京邮电大学学报(自然科学版) 6(2017):1-7.

[26] 周华平,张道义,孙克雷,秦黄利,桂海霞.基于ENM-Gabor差分权重的人脸表情特征提取方法[J].计算机应用与软件,2020,37(03):184-189+212.

[27] 谢惠华,黎明,王艳,陈昊.基于DE-Gabor特征的人脸表情识别[J].南昌航空大学学报(自然科学版),2021,35(02):82-91+124.

[28] Yin L , Wei X , Yi S , et al. A 3D facial expression database for facial behavior research[C]// International Conference on Automatic Face & Gesture Recognition. IEEE, 2006.

[29] C. Shan, S. Gong, and P. W. McOwan, “Facial expression recognition based on local binary patterns: A comprehensive study,” Image and Vision Computing, vol. 27, no. 6, pp. 803–816, 2009.

[30] P. Liu, S. Han, Z. Meng, and Y. Tong, “Facial expression recognition via a boosted deep belief network,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1805–1812.

[31] A. Mollahosseini, D. Chan, and M. H. Mahoor, “Going deeper in facial expression recognition using deep neural networks,” in Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on. IEEE, 2016, pp. 1–10.

[32] Zhao G , Pietikainen M . Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2007, 29:915-928.

[33] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, “Joint fine-tuning in deep neural networks for facial expression recognition,” in Computer Vision (ICCV), 2015 IEEE International Conference on. IEEE, 2015, pp. 2983–2991.

[34] X. Zhao, X. Liang, L. Liu, T. Li, Y. Han, N. Vasconcelos, and S. Yan, “Peak-piloted deep network for facial expression recognition,” in European conference on computer vision. Springer, 2016, pp. 425– 442.

[35] Pantic, M.; Valstar, M.; Rademaker, R.; Maat, L. Web-based database for facial expression analysis. In Proceedings of the 2005 IEEE International Conference on Multimedia and Expo, Amsterdam, Netherlands, 6 July 2005; pp. 5–pp, doi:10.1109/ICME.2005.1521424.

[36] DHALL A, GOECKE R, LUCEY S, et al. "Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark," 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), Barcelona, 2011, pp. 2106-2112, doi: 10.1109/ICCVW.2011.6130508.

[37] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee et al., “Challenges in representation learning: A report on three machine learning contests,” in International Conference on Neural Information Processing. Springer, 2013, pp. 117–124.

[38] A. Dhall, O. Ramana Murthy, R. Goecke, J. Joshi, and T. Gedeon, “Video and image based emotion recognition challenges in the wild: Emotiw 2015,” in Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. ACM, 2015, pp. 423–426.

[39] A. Dhall, R. Goecke, J. Joshi, J. Hoey, and T. Gedeon, “Emotiw 2016: Video and group-level emotion recognition challenges,” in Proceedings of the 18th ACM International Conference on Multimodal Interaction. ACM, 2016, pp. 427–432.

[40] A. Dhall, R. Goecke, S. Ghosh, J. Joshi, J. Hoey, and T. Gedeon, “From individual to group-level emotion recognition: Emotiw 5.0,” in Proceedings of the 19th ACM International Conference on Multimodal Interaction. ACM, 2017, pp. 524–528.

[41] DHALL A, GOECKE R, LUCEY S, et al. Collecting large, richly annotated facial-expression databases from movies[J]. IEEE multimedia, 2012, 19(3): 34-41.

[42] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Advances in neural information processing systems, 2012, pp. 1097–1105.
[43] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556, 2014.
[44] C. Szegedy,W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.
[45] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[46] Didan Deng et al. MIMAMO Net: Integrating Micro- and Macro-Motion for Video Emotion Recognition[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(03) : 2621-2628.
[47] Khor H Q, See J, Phan R C W, et al. Enriched long-term recurrent convolutional network for facial micro-expression recognition[C]//2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). IEEE, 2018: 667-674.
[48] Li, Jiaxing, et al. "Facial Expression Recognition with Faster R-CNN." Procedia Computer Science 107 (2017): 135-140.
[49] Hu M, Wang H, Wang X, et al. Video facial emotion recognition based on local enhanced motion history image and CNN-CTSLSTM networks[J]. Journal of Visual Communication and Image Representation, 2019, 59: 176-185.
[50] Ruan D , Yan Y , Lai S , et al. Feature Decomposition and Reconstruction Learning for Effective Facial Expression Recognition[J]. 2021.
[51] Davis L S . Covariance discriminative learning: A natural and efficient approach to image set classification[C]// Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2012.
[52] HU J, SHEN L, ALBANIE S, et al. Squeeze-and-Excitation Networks. 2020, 42(8):2011-2023.
[53] Woo S , Park J , Lee J Y , et al. CBAM: Convolutional Block Attention Module[J]. Springer, Cham, 2018.
[54] 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]// Computer Vision & Pattern Recognition Workshops. IEEE, 2010.
[55] T. F. Cootes, G. J. Edwards, and C. J. Taylor, “Active appearance models,” IEEE Transactions on Pattern Analysis & Machine Intelligence, no. 6, pp. 681–685, 2001.
[56] Zhu X , Ramanan D . [IEEE 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) - Providence, RI (2012.06.16-2012.06.21)] 2012 IEEE Conference on Computer Vision and Pattern Recognition - Face detection, pose estimation, and landmark localization in the wild[J]. 2012:2879-2886.
[57] Asthana A, Zafeiriou S, Cheng S, et al. Robust discriminative response map fitting with constrained local models[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2013: 3444-3451.
[58] X. Xiong and F. De la Torre, “Supervised descent method and its applications to face alignment,” in Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on. IEEE, 2013, pp. 532–539.
[59] S. Ren, X. Cao, Y. Wei, and J. Sun, “Face alignment at 3000 fps via regressing local binary features,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1685–1692.
[60] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, “Incremental face alignment in the wild,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 1859–1866.
[61] Y. Sun, X. Wang, and X. Tang, “Deep convolutional network cascade for facial point detection,” in Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on. IEEE, 2013, pp. 3476–3483.
[62] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, “Joint face detection and alignment using multitask cascaded convolutional networks,” IEEE Signal Processing Letters, vol. 23, no. 10, pp. 1499–1503, 2016.

[63] A. Mollahosseini, D. Chan, and M. H. Mahoor, “Going deeper in facial expression recognition using deep neural networks,” in Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on. IEEE, 2016, pp. 1–10.

[64] F. De la Torre,W.-S. Chu, X. Xiong, F. Vicente, X. Ding, and J. F. Cohn, “Intraface,” in IEEE International Conference on Automatic Face and Gesture Recognition (FG), 2015.

[65] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, “Facial landmark detection by deep multi-task learning,” in European Conference on Computer Vision. Springer, 2014, pp. 94–108.

[66] Gharbi M, Chen J, Barron J T, et al. Deep bilateral learning for real-time image enhancement[J]. ACM Transactions on Graphics (TOG), 2017, 36(4): 1-12.

[67] Sethuraman R, Kerin R A, Cron W L. A field study comparing online and offline data collection methods for identifying product attribute preferences using conjoint analysis[J]. Journal of Business Research, 2005, 58(5): 602-610.

[68] Z. Meng, P. Liu, J. Cai, S. Han, and Y. Tong, “Identity-aware convolutional neural network for facial expression recognition,” in Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on. IEEE, 2017, pp. 558–565.

[69] X. Liu, B. Kumar, J. You, and P. Jia, “Adaptive deep metric learning for identity-aware facial expression recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), 2017, pp. 522– 531.

[70] G. Levi and T. Hassner, “Emotion recognition in the wild via convolutional neural networks and mapped binary patterns,” in Proceedings of the 2015 ACM on international conference on multimodal interaction. ACM, 2015, pp. 503–510.

[71] D. A. Pitaloka, A. Wulandari, T. Basaruddin, and D. Y. Liliana, “Enhancing cnn with preprocessing stage in automatic emotion recognition,” Procedia Computer Science, vol. 116, pp. 523–529, 2017.

[72] A. T. Lopes, E. de Aguiar, A. F. De Souza, and T. Oliveira-Santos, “Facial expression recognition with convolutional neural networks: coping with few data and the training sample order,” Pattern Recognition, vol. 61, pp. 610–628, 2017.

[73] Z. Yu, Q. Liu, and G. Liu, “Deeper cascaded peak-piloted network for weak expression recognition,” The Visual Computer, pp. 1–9, 2017.

[74] Z. Yu and C. Zhang, “Image based static facial expression recognition with multiple deep network learning,” in Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. ACM, 2015, pp. 435–442.

[75] I. Abbasnejad, S. Sridharan, D. Nguyen, S. Denman, C. Fookes, and S. Lucey, “Using synthetic data to improve facial expression analysis with 3d convolutional networks,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1609– 1618.

[76] J. Cai, Z. Meng, A. S. Khan, Z. Li, J. OReilly, and Y. Tong, “Island loss for learning discriminative features in facial expression recognition,” in Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on. IEEE, 2018, pp. 302-309.

[77] X. Liu, B. Kumar, J. You, and P. Jia, “Adaptive deep metric learning for identity-aware facial expression recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), 2017, pp. 522-531.

[78] M. Liu, S. Li, S. Shan, and X. Chen, “Au-aware deep networks for facial expression recognition,” in Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on. IEEE, 2013, pp. 1-6.

[79] BACCOUCHE M, MAMALET F, WOIF C, et al. Spatio-Temporal Convolutional Sparse Auto- Encoder for Sequence Classification[C]//BMVC. 2012: 1-12.

[80] YAO A, SHAO J, MA N, et al. Capturing au-aware facial features and their latent relations for emotion recognition in the wild[C]//Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. ACM, 2015: 451-458.

[81] KHORRAMI P, LE P T, BRADY K, et al. How deep neural networks can improve emotion recognition on video data[C]//Image Processing (ICIP), 2016 IEEE International Conference on. IEEE, 2016: 619-623.
[82] ZHANG F, ZHANG T, MAO Q, et al. Geometry Guided Pose-Invariant Facial Expression Recognition[J]. IEEE Transactions on Image Processing, 2020, 29: 4445-4460.
[83] CHEN S, WANG J, CHEN Y, et al. Label Distribution Learning on Auxiliary Label Space Graphs for Facial Expression Recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 13984-13993.
[84] Wang K , Peng X , Yang J , et al. Suppressing Uncertainties for Large-Scale Facial Expression Recognition[J]. IEEE, 2020.
[85] 陈乐, 童莹, 陈瑞,等. 端到端增强特征神经网络的视频表情识别[J]. 重庆理工大学学报：自然科学, 2019, 33(9):7.
[86] ZHANG K, HUANG Y, DU Y, et al. Facial expression recognition based on deep evolutional spatial-temporal networks[J]. IEEE Transactions on Image Processing, 2017, 26(9): 4193-4203.
[87] P. W et al., "Understanding Convolution for Semantic Segmentation[C]," 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), 2018, pp. 1451-1460.
[88] FAJTL J, SOKEH H S, ARGYRIOU V, et al. Summarizing videos with attention[C]/Asian Conference on Computer Vision. Springer, Cham, 2018: 39-54.
[89] S. E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, C, . Gulc.ehre, " R. Memisevic, P. Vincent, A. Courville, Y. Bengio, R. C. Ferrari et al., “Combining modality specifific deep neural networks for emotion recognition in video,” in Proceedings of the 15th ACM on International conference on multimodal interaction. ACM, 2013, pp. 543-550.
[90] Huang G , Liu Z , Laurens V , et al. Densely Connected Convolutional Networks[J]. IEEE Computer Society, 2016.
[91] KHOR H Q, SEE J, PHAN R C W, et al. Enriched Long-term Recurrent Convolutional Network for Facial Micro-Expression Recognition[C]//Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on. IEEE, 2018: 667-674.
[92] Howard A G, Zhu M, Chen B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications[J]. arXiv preprint arXiv:1704.04861, 2017.
[93] Vaswani A , Shazeer N , Parmar N , et al. Attention Is All You Need[J]. arXiv, 2017.

附录 1 攻读硕士期间撰写的论文

[1] 唐武宾、童莹、曹雪虹，端到端增强卷积网络的视频人脸表情识别研究，软件导刊，已录用。

附录 2 攻读硕士学位期间参加的科研项目

（1）国家自然科学基金青年项目(61703201)，江苏省自然科学基金青年项目(BK20170765)

致谢

时光荏苒，不知不觉到了快要毕业的日子，回首过往，有过因科研而带来的焦虑，有过因找工作而带来的压力，有过因发论文而带来的喜悦，等等，多种情感交会，组成了这丰富多彩的研究生生涯。在这即将临别之际，感谢我身边给予帮助的任何一个人，其中包括了我的父母、我的老师、我的朋友、我的同学，是他们的存在给了我勇气，让我越过了很多人生的高峰，也培养了我独立处理问题的能力，这份能力使我能够在未来的工作中更有自信。因此，在此再次感谢对我有过帮助的你们。

首先感谢我的导师曹雪虹教授，对于我学习过程中的困难给予了悉心指导，对于我的生活给予了足够多的关心，在当初选择研究生研究课题时，能够让我选择自己感兴趣的方向以及领域。同时，在科研过程中，曹老师对于其中一些研究点以及思路给予了建设性的意见，对于小论文、开题报告、中期报告的撰写进行了督促把关，使我能够高标准的完成研究生研究课题。在科研过程中，本人曾因创新点效果不佳而产生了不自信的情绪，曹老师能够及时发现并给予开导、安慰，让我感觉到了老师对我的关怀。在此，衷心表示对曹老师的感谢与感激！

其次感谢童莹老师在科研过程中对我无微不至的帮助以及指导，在实验室时，童老师会定期组织汇报，让我们讲自己所做的工作在会上作阐述，同时基于疑问点进行提问，一次一次的例会，使我能够更加认真对待自己所做的课题，对于课题其中的难点进行挖掘，在碰到自己所不理解的地方，童老师能够进行指导，在碰到理解存在偏差的地方，童老师能够进行纠正，因此，我的成长离不开童老师的悉心帮助。童老师对于科研有着严谨的态度，对待工作有着如拼命三郎的精神，在跟童老师三年的相处过程中，她的精神以及做事风格深深的影响了我，未来，我也会向童老师学习，成为一个优秀的人。

接下来感谢与我朝夕相处的师门同学以及舍友，感谢师门葛垚、陈乐、王志强、任丽、赵曼雪、马杲东、陈雅玲以及苏擎凯对我平时科研的帮助，与你们相处的时光，我很开心。感谢舍友曾骏、王紫腾、何志敏、生柳振、赵亚楠，因你们的陪伴，我的研究生生活不孤单。希望未来无论身在何方，我们依然是最好的朋友，友谊长长久久。最后，衷心祝愿我研究生期间遇到的每一个朋友未来能够事业有成，做一个对社会、对国家有用的人！

最后，感谢评委老师们能够抽出时间阅读我的论文，您们辛苦了！希望您们能够根据我的论文提出宝贵的意见，谢谢！