分类号：_____ 单位代码：10300

密　级：_____ 学　　号：202212490130

南京信息工程大学

硕士专业学位论文

论文题目：基于卷积神经网络的
人脸检测和表情识别算法研究

申请人姓名：叶宇轩

指导教师：周先春

类别名称：电子信息

领域名称：通信工程（含带宽网络、移动通信等）

培养学院：电子与信息工程学院

提交时间：2025 年 6 月 18 日

二〇二五　年　六　月

目 录

摘要

在人脸的相关的应用领域中，人脸检测和表情识别是两个重要任务。随着人工智能技术的快速发展和广泛应用，基于深度学习的人脸检测和表情识别方法取得了显著的进展。而目前的人脸检测模型通常存在结构复杂、模型参数和计算量大等问题，表情识别模型存在不能充分提取各层次的细节特征，对部分区分度较小的表情易产生误判的问题，因此本文基于人脸检测和表情识别的问题展开研究。主要研究内容如下：

(1) 针对目前的人脸检测算法存在结构复杂、模型参数和计算量大等问题，提出一种基于卷积神经网络的轻量化人脸检测模型(MYDCFNet)。首先将 YOLOv4 的主干网络 CSPDarkNet53 替换为轻量型的 MobileNetV3 网络，在其颈部设计一种新的激活函数 NHS，通过调整α的大小在计算效率和非线性特性间取得平衡，其次提出新型深度可分离卷积 NESC，在特征融合过程中将 PANet 网络中的普通卷积替换为 NESC，结合自适应演化思想，动态调整卷积核结构来提升特征提取能力，然后提出一种采用多尺度空洞卷积并行结构的特征增强模块替代 SPP 模块，捕获更广泛的上下文信息，最后设计一个新的卷积注意力模块 DCAM，在上采样过程中，通过动态调整卷积核大小融合多尺度特征并加强特征提取。实验结果证明，本模型参数较原始模型显著减少了约 80 %，计算速度提升了近 67 %，检测精度达到了 97 %，提升了近 3 %，可实现实时的人脸检测。

(2) 针对目前的表情识别算法不能充分提取各层次的特征以及部分表情类间相似度高而导致误判的问题，提出一种基于多级特征提取和融合的表情识别网络模型(DFENet)。首先在数据集处理阶段，将人脸表情数据集划分为高相似集与低相似集，并构建由高低相似分支网络组成的多分支网络(RHLNet)。其中，高相似分支通过通道信息增强表情类间区分度，降低高相似类别的误分类率，低相似分支在维持低相似集区分度的同时，确保整体网络的均衡性。其次采用 DenseNet 密集连接网络为主干网络，提出一种多级特征提取模块 FEM，利用三个不同卷积核尺度的密集块分别提取低级、中级、高级特征，设计特征融合模块 FFM，先对局部特征进行独立提取，再对整体图像在输入前和提取全局特征后进行两个阶段的融合，进而将 FEM 与 FFM 结合为整体网络，通过 FFM 引导 FEM 提取并融合多层次面部表情特征。最终实验结果证明，提出的方法在 FER2013 和 CK+数据集上的识别精度达到了 94.81 %和 96.75 %，可以有效地提高人脸表情的识别率，并能有效地进行人脸表情的识别。

关键词：卷积神经网络，人脸检测，表情识别，注意力机制，多级特征提取

Abstract

In the face related application field, face detection and expression recognition are two important tasks. With the rapid development and wide application of artificial intelligence technology, deep learning-based face detection and expression recognition methods have made remarkable progress. However, current face detection models usually have problems such as complex structure, model parameters and large amount of calculation, and expression recognition models cannot fully extract detailed features at all levels, and some expressions with low differentiation are prone to misjudgment. Therefore, this paper conducts research based on face detection and expression recognition. The main research contents are as follows:

(1) A lightweight face detection model (MYDCFNet) based on convolutional neural network is proposed, aiming at the problems of complex structure, model parameters and large amount of computation in current face detection algorithms. Firstly, CSPDarkNet53, the backbone network of YOLOv4, was replaced by a lightweight MobileNetV3 network, and a new NHS activation function was designed on its neck to strike a balance between computational efficiency and nonlinear characteristics by adjusting the size of α. Secondly, a novel deeply separable convolutional NESC was proposed. In the process of feature fusion, the common convolution in PANet network is replaced by NESC. Combined with the idea of adaptive evolution, the convolution kernel structure is dynamically adjusted to improve the feature extraction capability. Then, a feature enhancement module with multi-scale cavity convolution parallel structure is proposed to replace the SPP module to capture a wider range of context information. Finally, a new convolutional attention module DCAM is designed to fuse multi-scale features and enhance feature extraction by dynamically adjusting the convolution kernel size during upsampling. The experimental results show that compared with the original model, the parameters of this model are significantly reduced by about 80 %, the calculation speed is improved by nearly 67 %, the detection accuracy is up to 97 %, and the detection accuracy is improved by nearly 3 %.

(2) Aiming at the problem that the current expression recognition algorithms cannot fully extract the features at all levels and the similarity between some expression classes is high, which leads to misjudgment, an expression recognition network model based on multi-level

feature extraction and fusion (DFENet) is proposed. Firstly, in the data set processing stage, the facial expression data set is divided into high similarity subset and low similarity subset, and a multi-branch network (RHLNet) composed of high similarity and low similarity branch network is constructed. Among them, the branches with high similarity enhance the differentiation between expression classes through channel information, and reduce the classification error rate of the categories with high similarity. The branches with low similarity ensure the balance of the whole network while maintaining the differentiation degree of low similarity sets. Secondly, DenseNet dense connected network is adopted as the main trunk network, and a multi-level feature extraction module FEM is proposed. The low level, medium level and high level features are extracted by three dense blocks with different convolution kernel scales, and the feature fusion module FFM is designed. First, local features are extracted independently. Then, the whole image is fused in two stages before input and after global features are extracted, and then FEM and FFM are combined into an overall network. FFM guides FEM to extract and fuse multi-level facial expression features. The final experimental results show that the recognition accuracy of the proposed method on FER2013 and CK+ data sets reaches 94.81 % and 96.75 %, which can effectively improve the recognition rate of facial expressions and effectively recognize facial expressions.

Key words: Convolutional neural network, Face detection, Facial expression recognition, Attention mechanism, Multi-level feature fusion

第一章 绪论

1.1 研究背景及意义

随着科技的快速发展，人们对于从图像中获取清晰、准确信息的需求日益增加，深度学习在人脸图像处理领域[1]得到广泛应用，计算机视觉[2]是一门以计算机为工具，通过对图像、视频进行分析、处理等一系列问题的一门学科。计算机视觉的研究背景可以追溯到二十世纪六十年代，当时科学家们开始尝试开发能够模拟人类视觉系统的计算机程序。

计算机视觉技术在众多应用场景中扮演着关键角色，对各行各业都产生了深刻的影响，具有很重要的研究价值。计算机视觉技术的主要应用领域包含以下几个方面：(1) 图像分类和识别[3-5]：帮助系统识别图像中的各种场景和它们的特征，让计算机能够理解并判断输入图像的内容，在自动驾驶、交通运输、人脸识别等方面具有非常重要的应用价值。(2) 目标检测和追踪[6-8]：通过分析视频或图片，让计算机自动从图像或视频流中时间序列中，对物体进行定位、分类和追踪，对视频监控系统、智能安防系统具有显著作用。(3) 三维建模和虚拟实境[9]：通过解析图像或视频让计算机在三维空间中重建物体或场景，并通过交互式技术让用户沉浸于虚拟环境中，共同推动了虚拟现实(Virtual Reality, VR)[10]、增强现实(Augmented Reality, AR)[11]技术的发展。(4) 医疗图像分析[12-15]：通过对各种医学影像(如 X 光、CT、MRI、超声等)进行自动或半自动处理与分析，协助医生进行病情诊断，治疗计划及手术后评价，推动了医疗服务向更加精准与智能的方向发展。(5) 视频分析和行为理解[16-19]：通过从动态影像中提取出丰富的语义信息与行为模式，从而完成事件检测、人物行为识别、动作分割等任务，为安防、交通、体育、娱乐等领域提供了强大的视频感知与分析能力。计算机视觉技术的这些应用，不但促进了相关领域的科技创新和发展，而且对人类社会的发展具有重要意义。

1.2 国内外研究现状

人脸检测和表情识别是计算机视觉和图像处理领域的重要研究方向，为了满足各类应用场景对人脸识别实时性和准确性的要求，国内外学者针对人脸检测和表情识别提出了多种理论和算法的研究，取得了显著的进展。当前的研究主要可以分为传统方法和深度学习方法两大类，本节将围绕这两类方法的研究现状进行梳理和介绍。

1.2.1 基于传统和深度学习的人脸检测方法

人脸检测[20]对于众多人脸相关应用具有关键性的作用。现有的人脸检测技术主要有采用传统的手动特征提取方式和利用深度学习两种方式进行识别。下面将围绕这两种方法的相关研究进行梳理并加以分析。

(1) 基于传统方法的人脸检测算法

对于传统的人脸检测方法，主要有 Haar(Haar Cascade Classifier)特征分类器检测[21]和基于 HOG(Histogram of Oriented Gradients)特征的人脸检测[22]方法。Viola 和 Jones[23]利用 Haar 类特征对图像进行多尺度滑窗检测，在每个窗口中快速计算 Haar 特征的响应值，然后结合 AdaBoost 分类器[24]逐级过滤，最终定位人脸区域，让计算更加简单高效，能够实时地进行人脸检测。VJ 人脸检测器[25]通过 AdaBoost 人脸检测框架，从大量 Haar 特征中选择一小部分最具判别力的特征，并将其组合成若干级(Stage)组成的级联结构，实现简单、检测速度快。Mathias 等人[26]提出方向梯度直方图(HOG)，它是一种利用图像中的梯度方向分布来提取目标的边界和纹理信息的方法，借助梯度信息来更好地识别面部的关键轮廓与特征点，从而提升对光照、背景变化的适应性。Sudhaker 等人[27]采用 Gabor 滤波器对人脸进行检测，该方法生成的人脸图像具有不同的角度、朝向，并通过匹配算法与已有的人脸特征进行比对。Banerjee A 等人[28]提出了一种新的基于局部二元模型(Local Binary Patterns, LBP)的肤色检测算法，使用 LBP 描述局部纹理特征，通过训练分类器对滑窗进行人脸判定。然而，上述传统的人脸检测方法，需要大量的专业知识和经验并且耗时耗力，而且由于人工提取特征表达能力有限，所以对多个图像检测效果并不理想。

(2) 基于深度学习的人脸检测算法

随着人工智能技术的迅猛发展和卷积神经网络(Convolutional Neural Networks, CNN)[29]的广泛应用，科研学者们提出了一系列的深度学习网络模型，特别是在人脸图像的特征提取与目标检测方面[30-31]，深度学习的优势更为明显，基于深度学习的人脸检测方法也取得了显著的进展。2016 年，Zhang 等人[32]提出一种基于 P-Net、R-Net 和 O-Net 三层级联的多任务卷积神经网络 MTCNN，通过多任务学习实现人脸检测和关键点定位，能够适应复杂的情况。Vimal C 等人[33]基于 VGG16 架构，结合迁移学习和轻量化设计提出一种实时人脸检测系统，在面部被口罩遮挡时也能检测多个人脸。Sunnetci K M 等人[34]通过使用 GoogLeNet 深度学习架构的 loss3-classifier 层提取深度图像特征，结合线性、二次和高斯核的支持向量机(SVM)分类器进行二分类任务，实现了高精度和低

复杂性方面的优势。Chen B 等人[35]提出一种改进的 Xception 模型，通过引入特征金字塔网络实现多级特征的提取，从而为最终决策提供丰富的信息支持。Yan H 等人[36]针对传统 Faster RCNN 人脸检测模型在复杂场景下出现的漏检和误检问题，提出一种基于线性加权 NMS 的人脸检测方法，通过动态调整重叠候选框的置信度从而保留更多真实目标框，显著增强了复杂环境下的检测鲁棒性，解决了多目标重叠时的漏检问题。Peng S 等人[37]通过改进 Inception-Resnet 模型，将残差缩放因子设置为可训练参数，并以参数化 PReLU 替代 ReLU 激活函数，能有效地利用负相关性，减少卷积核的冗余度，从而显著提高训练的稳定性和性能。

基于深度学习的人脸检测方法相较于传统方法具有更好的检测效果，尤其在复杂环境中能够大幅提升人脸检测的准确性和鲁棒性，同时能更好的适应大规模数据和多样化场景，为实时检测和多场景应用提供了有力的支持。

1.2.2 基于传统和深度学习的表情识别方法

人类的面部表情是情绪与心理状态最直接、最具普适性的外在表现之一，在人机交互、心理诊断、智能安防、娱乐等众多应用场景中扮演了重要角色。目前的面部表情[38]识别技术主要有两种，一种是基于传统的面部表情识别技术，另一种是基于深度学习的面部表情识别。下面将围绕这两种方法的相关研究进行梳理并加以分析。

(1) 基于传统方法的表情识别算法

1978 年，Ekman 和 Friesen[39]对人类的 6 种不同情绪(高兴、悲伤、惊讶、恐惧、愤怒、厌恶)的面部表情进行了开创性的研究，并构建了包含数千种不同表情的面部表情数据库，对每个表情对应的脸部变化进行了详细的描述。2001 年，Hafed 等人[40]提出一种利用离散余弦变换来进行脸部表情识别的方法，运用离散余弦变换技术对人脸图像进行频域分解，通过提取与表情相关的频域系数作为特征表示，并以此来实现面部表情的识别，最终得到了比较好的实验结果。2002 年，Liu 等人[41]提出将核主成分分析(Kernel Principal Component Analysis, KPCA)与 Gabor 小波相结合的表情识别算法，该方法通过 Gabor 小波提取人脸局部的纹理与边缘特征，然后利用 KPCA 在非线性特征空间中进行降维与特征选择，从而更好地保留表情差异信息。2003 年，Wen 等人[42]通过对表情图像进行预处理，提取出人脸的均值 Gabor 子波特征，并将其应用于面部表情识别中，该方法可有效避免光照等因素的影响，提高人脸识别准确率。上述传统的表情识别方法主要依赖于人工设计的特征，在捕捉面部表情的细节特征上能力有限，难以反映复杂情绪的

多样性，同时在光照、姿态等复杂环境下的表情识别效果不理想，鲁棒性不足。

(2) 基于深度学习的表情识别算法

近年来，伴随着深度学习技术的快速发展，人脸表情识别技术得到了突破。2009 年，Shan 等人[43]提出采用计算机视觉算法中的局部二值模式进行综合研究表情识别模型。Alex 等人[44]在 2012 年度图像识别大赛中提出了大型深度卷积神经网络 AlexNet，图像分类的准确率首次提升至 80 %以上，深度学习技术也得到了快速的发展。HE 等人[45]提出了一种基于深度残差学习的图像识别方法，使深层神经网络的训练变得更加简单。Cotter 等人[46]提出了一种新的轻量型模型，通过使用深度可分卷积、快速下采样等方法来实现模型的轻量化，但识别精度仍需进一步提升。Mollahosseini 等人[47]提出了一种基于深层神经网络结构的自动面部表情识别方法，其最大优点在于能够精确识别人脸不同部位的面部表情，提高人脸的精确识别率，并降低神经网络的训练参数，但受姿态、光照等因素的影响，其性能存在一定的差异。Vaswani 等人[48]提出了一个基于注意力机制的深度学习框架 Transformer，该框架最初在自然语言处理(Natural Language Processing, NLP)领域[49]中获得了突破，如今在计算机视觉等领域也成为了热门研究方向，它与 CNN、RNN[50]等传统网络不同，Transformer 摒弃了循环和卷积结构，依靠多头自注意力(Multi-Head Self-Attention)来建模序列中各位置间的关联。受到 Transformer 强大的表示能力的启发，Ma 等人[51]利用双分支 CNN 提出了一种具有特征融合的视觉转换器(Vision Transformer with Feature Fusion, VTFF)进行人脸表情识别，它使用 ResNet 对图像进行提取，并将其放入 Transformer 中对图像进行分类。Zhang 等人[52]创新性的设计了一种基于 Transformer 架构的网络模型，它包含了空间模板的空间编码器、时间聚合器，以及用于时间维度的分类头，其基本思想是先将光流场的特征信息输入到 Transformer 编码器，再利用 LSTM 对其进行时间和空间特征的融合处理。Asifullah K 等人[53]提出了一种将卷积运算和自注意力机制混合的视觉转换器模型(Vision Transformer, ViT)，将图像分割成若干补丁，将每个补丁视为一个序列元素，最后将这些补丁向量输入到 Transformer 编码器中进行特征抽取与分类，最终取得了与 CNN 相当或更优的性能，表明在视觉任务中也可以使用纯注意力结构替代卷积。

目前，基于深度学习的表情识别算法相较于传统方法具有显著优势，能够自动学习和提取多层次的特征，减少传统方法中对于人工提取特征的依赖，能更好地适应不同场景。随着卷积神经网络架构的不断优化，深度学习方法在计算效率和泛化能力方面也得到了进一步提升，广泛应用于实时表情识别任务。

1.3 主要研究内容

本文采用四个公开的人脸数据集，使用卷积神经网络对人脸检测和表情识别等相关问题展开研究，本文具体研究内容如下：

(1) 针对目前的人脸检测算法往往存在结构复杂、模型参数和计算量大等问题，提出一种基于卷积神经网络的轻量化人脸检测模型(MYDCFNet)。首先，将 YOLOv4 的主干网络 CSPDarkNet53 替换为轻量型的 MobileNetV3 网络，并在其颈部设计一种新的激活函数 NHS，通过调整α的大小在计算效率和非线性特性间取得平衡。采用提出的新型深度可分离卷积 NESC，在特征融合过程中将 PANet 网络中的普通卷积替换为 NESC，引入进化计算思想，通过自适应演化调整深度卷积核的结构，提高特征提取效果。然后，针对传统空间金字塔池化结构(Spatial Pyramid Pooling, SPP)中的最大池化操作可能导致的特征信息损失问题，提出采用多尺度空洞卷积并行结构的特征增强模块替代 SPP 模块，捕获更广泛的上下文信息。最后，设计一个新的卷积注意力模块 DCAM，在上采样过程中，通过动态调整卷积核大小融合多尺度特征并加强特征提取，使其更加高效地处理输入数据。最终实验结果证明，本方法在网络模型大小和检测精度两方面都取得一定改善，能够实时的进行人脸检测。

(2) 针对目前的表情识别算法不能充分提取各层次的特征以及部分表情类间相似度高导致误判的问题，提出一种基于多级特征提取和融合的表情识别网络模型(DFENet)。首先，在数据集处理阶段，将人脸表情数据集划分为高相似集与低相似集，并构建由高低相似分支网络组成的多分支架构(RHLNet)。高相似分支通过通道信息增强表情类间区分度，从而降低高相似类别的误分类率，而低相似分支在维持低相似集区分度的同时，确保整体网络的均衡性。然后，使用 DenseNet 密集连接网络为主干网络，提出一种多级特征提取模块 FEM，利用三个不同卷积核尺度的密集块分别提取低级、中级、高级特征，设计特征融合模块 FFM，先对局部特征进行独立提取，再对整体图像在输入前和提取全局特征后进行两个阶段的融合，进而将 FEM 与 FFM 结合为整体网络，通过 FFM 引导 FEM 提取并融合多层次面部表情特征。最终实验结果证明，本方法在 FER2013 和 CK+数据集上的识别精度达到了 94.81 %和 96.75 %，在识别精度方面取得一定改善，能对人脸表情进行有效识别。

1.4 论文章节安排

本文主要围绕基于深度学习的人脸检测与表情识别算法展开研究。全文共有五个章节，每章的主要内容和具体安排如下：

第一章：绪论。阐述了人脸检测和表情识别的研究背景及意义，并对目前国内外的研究状况进行介绍，最后针对这些算法存在的不足对本文的研究内容以及论文的章节安排进行简要介绍。

第二章：基于卷积神经网络的基础理论。首先对卷积神经网络中的基本结构进行介绍，由此引入了不同经典模型的网络架构，最后对本文提出的网络所涉及到的基础知识进行介绍。

第三章：基于卷积神经网络的轻量化人脸检测算法。提出一种基于卷积神经网络的轻量化人脸检测模型(MYDCFNet)，本章首先并对提出的模型进行整体介绍，对主干网络、激活函数、深度可分离卷积和卷积注意力模块的改进与具体实现分别进行介绍和分析。然后在两个数据集上进行实验并分析结果，通过消融实验验证改进网络的各模块的有效性，通过与主流的卷积神经网络进行对比实验验证该方法在人脸检测中的优势。

第四章：基于特征提取和融合的人脸表情识别算法。提出一种基于多级特征提取和融合的表情识别网络模型(DFENet)。本章首先介绍了算法的整体设计流程与框架，其次对高低相似分支网络(RHLNet)和特征提取和融合网络(DFENet)两个网络及其中具体的模块进行详细的介绍与分析。接着为了验证模块的有效性，在两个数据集上进行实验并分析结果，通过消融实验、对比实验以及表情识别结果，对所提出的算法各模块的性能和改进效果进行验证。

第五章：总结与展望。对本文提出的人脸检测和表情识别两种算法进行总结归纳，指出存在的不足之处，并进一步介绍了未来的研究方向。

第二章 基于卷积神经网络的基础理论

图像的检测与识别是计算机视觉研究的一个核心问题，在许多与人脸有关的应用中起着至关重要的作用。由于模型参数量大、特征提取不明显等问题给后续人脸检测和表情识别的研究带来挑战。针对这些挑战，国内外的研究者提出了许多人脸检测和表情识别算法。本章重点介绍人脸检测和表情识别相关基础理论知识，主要包括卷积神经网络结构、不同模型的网络架构以及注意力机制的介绍，为本文所提算法奠定理论基础。

2.1 卷积神经网络基本结构

卷积神经网络(CNN)是近年来在图像处理等方面被广泛应用的一种神经网络。CNN能够自动从原始图像中提取特征，并利用这些特征进行任务分类或回归等操作。与传统的机器学习算法不同，CNN 不需要预先设定的特征提取器，而是通过层次化的卷积操作进行特征学习。卷积神经网络处理图像采用先局部后整体的方法，由输入层、卷积层、池化层和全连接层等基本模块构成，在此基础上，通过对卷积网络的学习，将卷积神经网络所得的分类得分与训练集中的每一个图像的标记一致。在该网络中，通过反复建立这样的结构，形成一个交互相连的叠合结构，直至将输入的图像变换成若干个小方块。在此模式中的最后一层中，被转化的小方块被转移到全连接层中。卷积神经网络结构如图 2-1 所示：

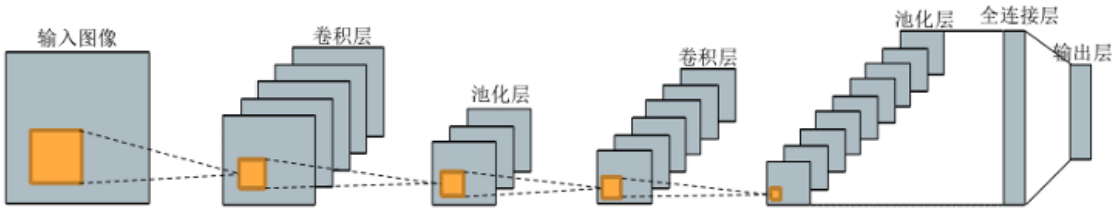


图 2-1 卷积神经网络结构图

(1) 输入层

输入层作为神经网络的初始组成部分，接收未经处理的原始输入数据并生成符合网络计算要求的张量格式。在图像处理任务中，输入层通常直接接收图像数据，例如彩色图像或灰度图像，将对应图像目标输入进入神经网络当中，并对输入图像尺寸、通道数等做出基本设置。

(2) 卷积层

卷积层是卷积神经网络的核心组件，用于在输入数据上提取局部空间特征。通过卷

积运算，卷积层可以在保持空间结构的同时，从不同层面学习到图像的边缘、纹理等。卷积原理如图 2-2 所示，这是一种对两个像素矩阵进行点乘求法和运算的操作，其结果代表了从原始图像中提取出来的部分特征。通过设定卷积核的大小、步长和填充等参数来确定卷积网络的大小。卷积神经网络通过对输入数据的卷积，对输入数据执行局部操作，从而获得一系列特征图谱。输出特性图包含了滤波后输入层的所有特性。其中数值越大则表示与特征越符合。

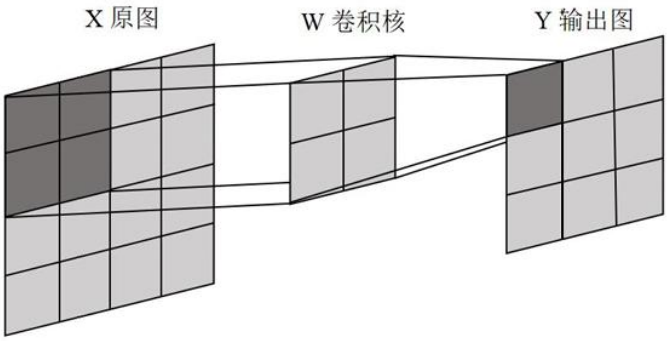


图 2-2 卷积原理示意图

(3) 池化层

池化层是卷积神经网络的重要组件，通过将卷积网络中提取出来的特性图进行降采样，从而减少网络的计算量和参数量，加速模型的训练过程。池化层通过某种方式对局部区域进行操作，将每个局部区域的多个值压缩成一个单一的输出值，这样做既能保留图像的主要特征，也能减少噪声。

在池化操作中，最具代表性的有最大池化(Max Pooling)和平均池化(Average Pooling)两种方法。其中，最大池化是在每一个局部区域中，以最大值为输出，这是最常见的池化方法，能够保留图像中最显著的特征；而平均池化是在每一个局部区域进行平均处理，并取其平均值作为输出，平均化可以使图像更光滑，并能降低局部特征的突出性。对整个特征图进行池化操作，通常采用最大池化或平均池化来输出单一的值。全局池化常用于分类任务的最后一层。如图 2-3 所示，通过使用 2×2 的窗口，步长为 2 来展示最大池化和平均池化操作，然后将尺寸 4×4 的特征图通过两个池化层后转化为尺寸为 2×2 的特征图。

池化层的目的在于将卷积后的特征作更深层次的压缩，降低了特征图的规模，从而减小计算量并防止过拟合。对于任意图像选择对其进行压缩，通过对特征图进行降维，可以大幅度的降低所要处理的数据量，使其在保持原有特征的同时，具有更加抽象的特征表示，并且可以清晰看出压缩后的结果依然保留了原始的特征。

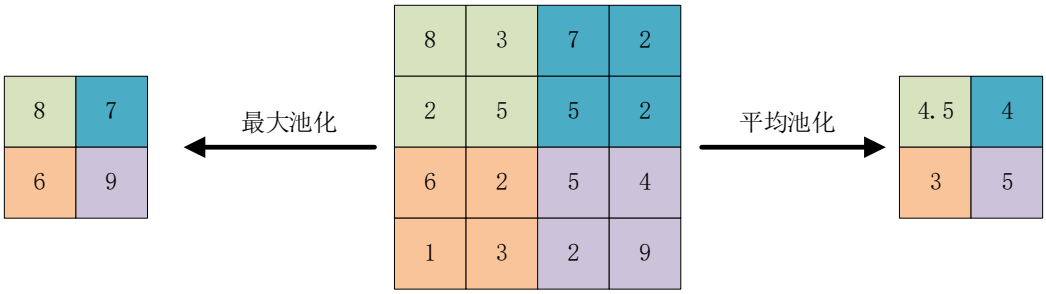


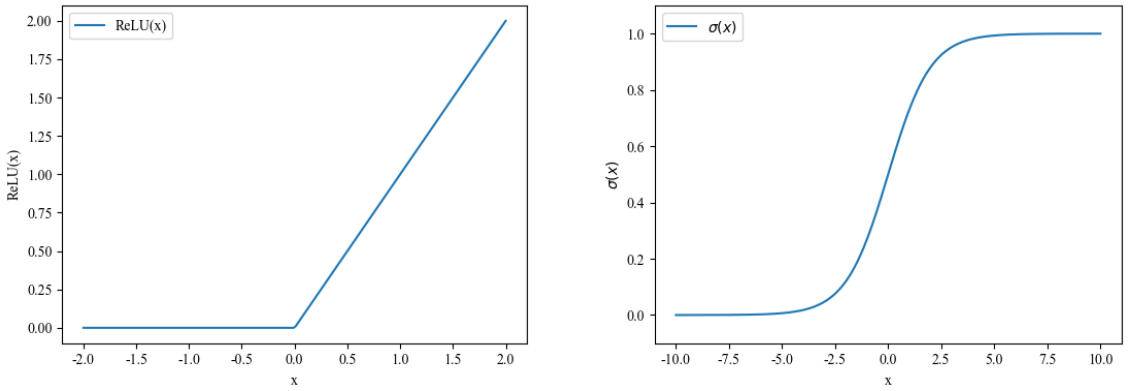
图 2-3 池化操作示意图

(4) 全连接层

全连接层是神经网络中的基本结构，它通常位于网络的最后部分。它的主要作用是对输入特征进行全局整合和非线性变换，将前面层提取的局部特征或全局特征进行组合，然后将整合的特征进行映射得到最终的预测结果或是分类标签。在全连接层中，每个节点与上一层的节点之间是相互连通的，在此基础上，对前一层的特征进行了融合，并将其映射到样本标签空间。全连接层将特征图展开并且拼接得到一维的特征数组，系统通过计算将目标图像的一维数组与训练的数组进行对比，从而得到相似度的匹配或判断。

(5) 激活函数

激活函数是神经网络中的重要组成部分，它通过对卷积神经网络的非线性分析，使其能够有效地解决线性系统无法解决的问题。AlexNet 成功地把 ReLU 用作 CNN 的激活函数，将 Sigmoid 激活函数转换为一个更为简单的 ReLU 激活函数。该方法的优点是可以简化 ReLU 激活函数的运算；此外，利用 ReLU 的激活函数，通过改变初始化参数的方式，ReLU 的活化功能可以很容易地对模型进行训练。尽管 ReLU 激活功能已经很早以前就被提出来了，但在 AlexNet 的问世之后，它才得到了进一步的发展。激活函数如图 2-4 所示。



(a) ReLU 激活函数　　(b) Sigmoid 激活函数

图 2-4 激活函数图像

(6) 损失函数

损失函数是机器学习和深度学习模型中一种非常重要的概念，它对神经网络的训练起着至关重要的作用。其本质是一个数学映射函数，通过计算预测值和实际标签之间的差异程度，明确模型参数的优化方向，从而使得模型的预测结果逐渐接近真实结果。

损失函数通常是在训练过程中作为优化目标函数进行最小化的。通过反向传播算法，损失函数的梯度被计算并传递回网络的每一层，进一步更新网络参数。在训练过程中，损失函数的值会反映出模型的预测效果，通过对模型参数的连续调节，使损失函数的值逐渐减少，以达到提高模型预测精度的目的。

2.2 理论基础知识

2.2.1 YOLO 系列算法

YOLO(You Only Look Once)[54]系列算法是最早是由 Redmon 等人在 2016 年提出的，其采用单阶段检测框架，将目标检测问题重新定义为一个端到端的回归问题，能够在单次前向传播的过程中输出图像所有预测的目标信息和类别概率，无需像传统方法那样先生成候选区域再逐一分类。这种端到端的检测方法使得 YOLO 能够实时检测多个目标，并在速度和准确性之间取得了良好的平衡。与传统的物体检测算法需要多次操作不同，YOLO 通过单次推断完成物体的检测，极大程度地提升了检测速度。YOLO 算法进行了多次版本的更新迭代，每一个版本在速度、精度和适用场景上都有所改善，被广泛应用于自动驾驶、视频监控等需要实时检测的场景，在精度和速度之间达到了良好平衡。

YOLOv4[55]对 YOLOv3[56]进行了进一步的优化，改进了数据增强、网络结构和训练技巧，提高了精度和速度的平衡。YOLOv4 的整体架构采用了模块化的设计思想，主要由三个部分构成，分别为用于初始特征提取的主干网络、实现多尺度特征融合的颈部网络以及实现最终预测的目标检测头，具体的网络模型结构如图 2-5 所示。

YOLOv4 采用 CSPDarknet53 作为主干网络(Backbone)进行初始特征提取[57]，在 CSPDarknet53 中嵌入了五个 CSP 模块，将基础层输出的特征图分割成两部分，并通过跨阶段特征融合策略将两个部分的输出进行整合，不仅显著降低了模型的参数量和计算量，同时通过保留丰富的梯度信息还能有效缓解网络训练过程中的梯度消失问题，从而实现了在减少参数量的同时依然能维持较好的检测精度。随后，在颈部网络部分采用空间金字塔池化(SPP)[58]和路径聚合网络(PANet)[59]。其中，在加强特征提取网络中引入的 SPP 通过使用多种尺寸的最大池化并将它们的输出特征图进行堆叠(Concat)操作，有效拓展了特征信息的感受野。PANet 引入了一种自下而上的路径增强机制，通过利用低层

的位置信息来增强高层特征的表达能力，减少底层和上层之间的信息传递距离。同时，在该网络结构中使用 Mish 等激活函数，有助于进一步提高 CSPDarknet53 网络的分类准确率，提高了 YOLOv4 网络模型的泛化能力，相较于 YOLOv3 在速度与准确度上有显著的提升。

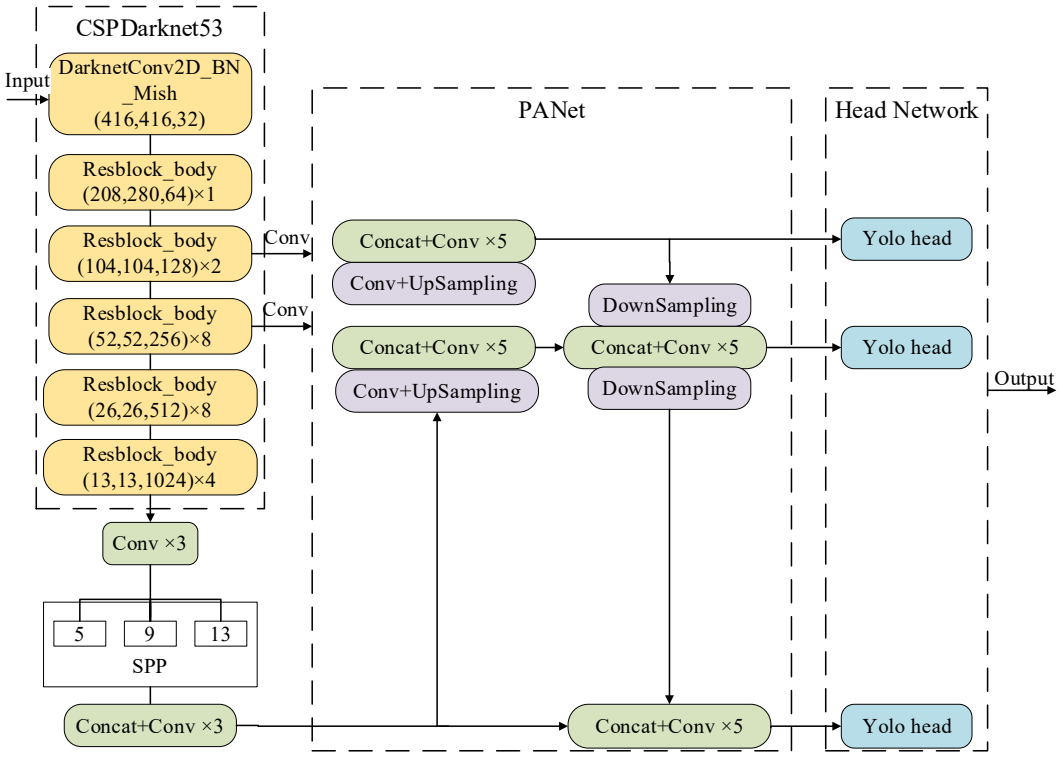


图 2-5 YOLOv4 网络结构

2.2.2 MobileNet 网络

MobileNet[60]是 2017 年为移动和嵌入式视觉应用提供的轻量级高效卷积神经网络，它通过轻量级的结构设计，显著降低了计算资源消耗，同时保持了相对较高的性能，能够在算力或硬件资源有限的设备上高效执行深度学习任务。MobileNet 采用了深度可分离卷积(Depthwise Separable Convolution)[61]，极大地降低了网络的计算量与参数量。MobileNet 网络系列已经推出多个版本，其中 MobilenetV3[62]是基于 MobilenetV1[63]的深度可分离卷积和 MobilenetV2[64]的反向残差块，再加入 SE 注意力模块(Squeeze-and-Excitation, SE)[65]得到的，SE 模块根据通道的重要程度计算特征权重，使得卷积神经网络能够突出关注关键的特征通道，并抑制不重要的特征通道。这样既提高了准确率，又不影响网络效率。

如图 2-6 所示，MobileNetV3 的主体由 bneck 结构构建而成，该结构包括主干和残差两部分。主干首先采用 1×1 卷积核扩展输入特征图的通道维度，以增强特征的表达能

力，通过 3×3 深度可分离卷积操作对高维特征进行空间滤波，接着对卷积输出实施全局平均池化操作，并在此基础上，通过两个连续的全连接层对特征进行非线性变换和降维，从而生成具有判别性的特征表示。最后，再以 1×1 卷积降维并输出结果。在激活函数方面，MobileNetV3 网络在前面部分采用了传统的 ReLU，而在更深层次的部分中采用 h-swish，这一做法既降低了整体运算量，又能获得更优的性能表现。

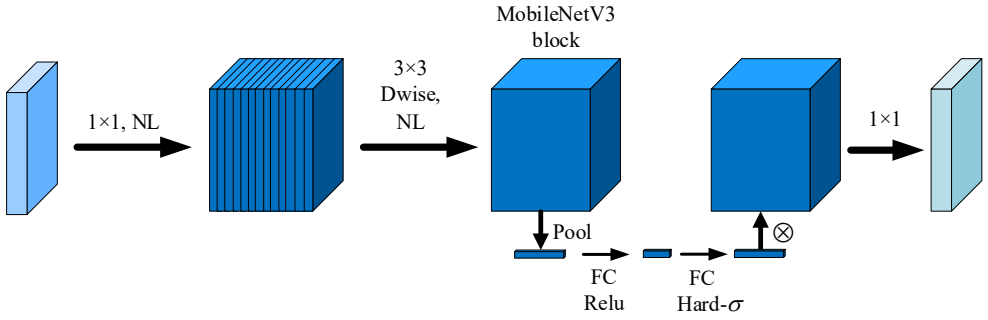


图 2-6 MoblieNetv3 bneck 结构图

深度可分离卷积是一种基于卷积运算的变种，它能有效地降低神经网络的计算复杂度，并能有效地降低网络中的参数。深度可分离卷积将卷积运算分解成深度卷积(Depthwise Convolution)[66]和点态卷积(Pointwise Convolution)[67]两个独立过程，如图 2-7 所示。正因如此，深度可分离卷积在追求模型轻量化的设计中获得了广泛使用，尤其适合部署在移动端等资源受限的场景。

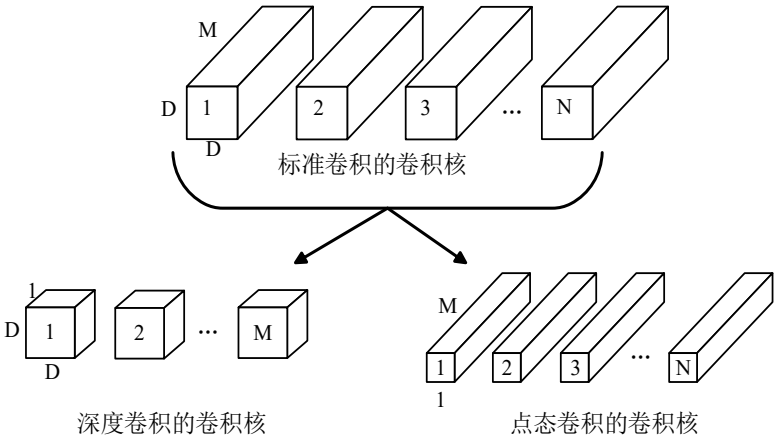


图 2-7 标准卷积核的分解过程

标准卷积过程是将 N 个卷积核与输入数据卷积，而深度可分离的 3×3 卷积首先使用 N 个卷积核进行分别卷积，然后将这个 3×3 的特征图与 N 个卷积核卷积，最终实现减少参数数量的效果，标准卷积核的分解过程如图 2-7 所示，通过判断输入通道数目的不同进而采用不同的卷积核进行卷积运算，最后调整通道通过单位为 1 的卷积核进行处理。

通过将卷积分解为两个独立阶段，深度可分离卷积可使参数与计算量仅相当于传统卷积的 1/N，大幅降低模型的复杂度并有效减少计算量。标准卷积的计算量为X_1，DW 卷积和 PW 卷积的计算量之和为X_2，减少的计算量为X_3，具体的计算公式如公式(2-1)至公式(2-3)所示：

$$X_1 = D_W \times D_H \times D \times D \times N \tag{2-1}$$

$$X_2 = D_W \times D_H \times D \times D + M \times N \times D_W \times D_H \tag{2-2}$$

$$X_3 = \frac{X_2}{X_1} = \frac{1}{N} + \frac{1}{D^2} \tag{2-3}$$

其中，标准卷积的输入映射尺寸为(D_W, D_H, M)，标准卷积的卷积核为(D, D, M, N)，输入映射的宽度和高度分别为D_W和D_H，输入通道数为M，卷积核的尺寸D，输出的通道数为N。由公式可知，与深度可分卷积相比，普通卷积无论在参数尺度还是计算量方面都要大很多，而且随着其卷积核尺寸的增大，这一差异将更为明显，表明深度可分卷积可以有效地提高计算效率。

2.2.3 ResNet 残差网络

2015 年，何恺明等人率先提出了一种基于深层卷积神经网络结构的残差网络(Residual Networks, ResNet)[68]，在 ImageNet 图像分类竞赛中获得了优异成绩，并已在各种机器视觉领域得到了广泛的应用。ResNet 的引入，成功缓解了传统神经网络在深度不断加深时的训练难题，并为深层神经网络的演进开辟了新的篇章。

ResNet 的主要思路在于采用残差学习(Residual Learning)，使网络专注于学习输入与输出之间的差值，而非直接拟合完整映射，从而降低深层网络的优化难度。并且，通过在卷积层块外设置跳跃连接(Skip Connections)或快捷连接(Shortcut Connections)，模型能在反向传播时更有效地更新可训练参数，避免梯度消失或爆炸，从而使信息在深层网络中得以直接传递，解决深度加深后出现的退化问题并加快训练。残差模块通常由卷积层、批归一化(Batch Normalization)、激活函数 ReLU 和跳跃连接组成，残差模块具体的两种连接形式如图 2-8 所示。

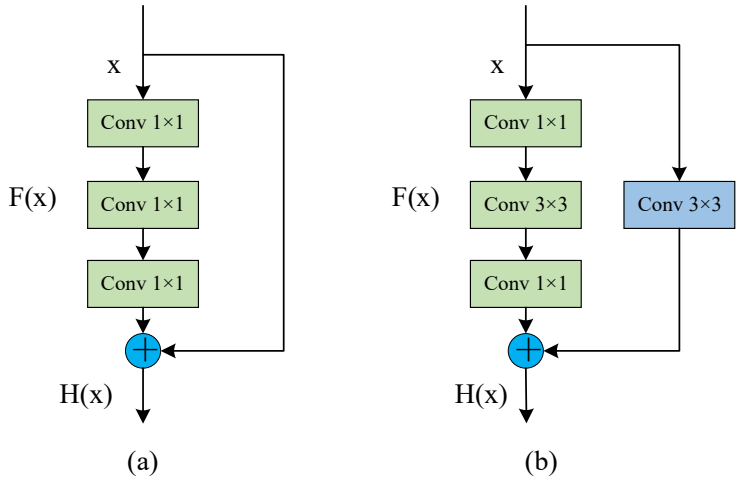


图 2-8 残差模块的两种连接方式

从上图可以看出，残差模块由卷积层和一个跳跃连接组成，假设某一层的输入是X，经过卷积、激活、批归一化等操作后得到输出$F(X)$，最后，利用跳跃连接将前一层残差模块的输出X与当前一层的$F(X)$相结合，并作为下一层残差模块的输入Y。如公式(2-4)所示：

$$Y = F(X, W_i) + X \tag{2-4}$$

其中，$F(X, W_i)$是对输入X进行卷积、激活等操作后得到的特征图，X是残差块的输入，W_i是网络中卷积操作的权重参数，Y是残差块的输出。通过叠加多个残差模块，残差神经网络在特征提取方面更具优势，而跳跃连接则能缓解深层网络中的梯度消失问题，还通过特征复用机制增强了网络的信息传递能力。

2.2.4 DenseNet 密集连接网络

密集连接网络(Densely Connected Convolutional Networks, DenseNet)[69]是一种深度神经网络，由 Huang 等人在 2017 年提出，目的在于通过密集跨层连接(Dense Connections)解决深度神经网络中的梯度消失问题，并提升特征复用效率。其核心设计原则为每一层的输入均来自前面所有层的输出，从而实现特征的多级复用与信息流的充分传递。模型构建思路与 ResNet 基本一致，但是它的前后层都使用加法变量连接的密集连接方法，可以在特征信道上实现特征的复用，从而达到对数据特征的极致利用。DenseNet 主要由特征重用(Dense Block)和连接层(Transition)组成，Dense Block 中采用批归一化、ReLU 激活函数和卷积层的结构，这样可以保证在进行每一个卷积层的操作时，数据均是标准化处理过后的数据。Transition 层作为连接相邻 Dense Block 的关键组件，有着压缩特征图尺寸压缩和控制通道数的重要作用，通过 1×1 卷积层对密集连接块输出的高维特征进行降维处理，随后通过 2×2 平均池化对特征图进行下采样，从而在保留重要特征信息的

同时显著降低计算复杂度，DenseNet 的结构如图 2-9 所示。

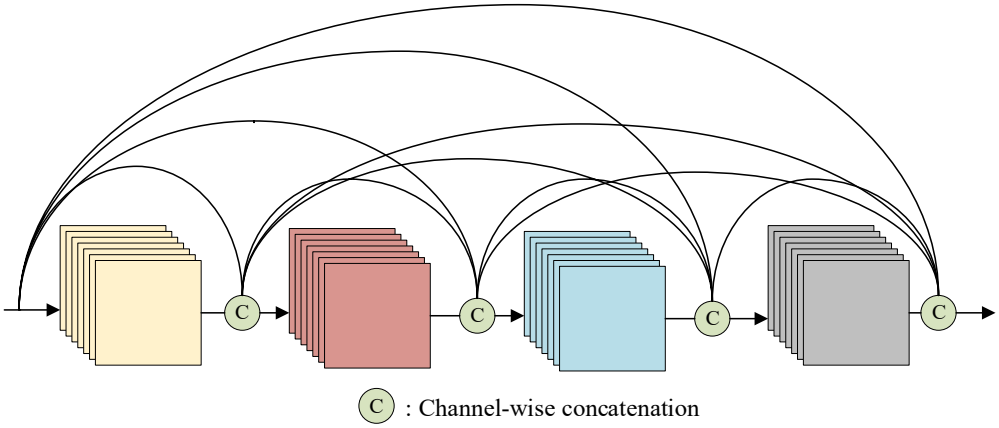


图 2-9 DenseNet 结构图

这种密集连接机制不同于传统的卷积神经网络，在传统的 CNN 中，每一层的输出只依赖于前一层，而在 DenseNet 中，每一层的输出依赖于所有前面的层的输出，如公式(2-5)所示：

$$x_n = F_n([x_0, x_1, ..., x_{n-1}]) \tag{2-5}$$

其中，$x_0, x_1, \dots, x_{n-1}$表示前 n 层的输出。$[x_0, x_1, \dots, x_{n-1}]$表示对这些输出进行拼接。$F_n()$是第 n 层的变换函数表示连接操作，x_n表示密集连接层的输出。

DenseNet 是一种创新的网络架构，它通过密集连接的设计增强了网络的特征提取能力，提高了梯度流动，减少了梯度消失问题，并且通过特征复用和较少的参数使得训练更加高效，提高模型的性能，使得网络具有更强的表达能力和更好的训练效果。DenseNet 在多个计算机视觉任务的领域中有着良好的性能。随着深度学习的进一步发展，DenseNet 将继续在更广泛的应用中发挥其优势。

2.3 注意力机制

在人的感知过程中，注意力扮演着关键角色，视觉系统运行时往往会自动聚焦于图像中更显著、引人注目的部分。注意力机制(Attention Mechanism)[70]最初源于自然语言处理领域，伴随着深度学习的迅速发展，现已被广泛引入到计算机视觉、语音处理等多种领域。注意力机制借鉴了人类处理信息时聚焦关键信息的过程，使模型能够集中于重要特征，从而提升性能与效率，注意力由此应运而生，用来提高 CNN 在分类任务中的性能。在卷积神经网络中，注意力机制通过对输入数据的不同部分赋予不同的权重，从而使网络能够关注到输入的关键部分，从而改善模型的学习效果。特别是在处理长序列、

复杂图像或大规模数据时，注意力机制可以有效提高网络的表现。对于输入的所有信息，模型通过分配不同的权重，使得某些部分的贡献更大，其他部分的贡献则相对较小。这样模型就能够专注于最重要的信息，而忽略掉无关的信息，从而提高整体的学习效率和性能。

根据任务和具体的应用场景，常见的注意力机制可以分成下面两种不同类型：(1) 软注意力机制(Soft Attention Mechanism, SAM)[71]：软注意力通过动态计算输入特征不同区域的权重分布，实现对关键信息的自适应聚焦，由于整个过程是连续可微的，所以软注意力机制能够嵌入到深度学习框架中，并通过反向传播算法进行端到端的优化。常见的软注意力机制有通道注意力机制、空间注意力机制和卷积注意力机制等。(2) 硬注意力机制(Hard Attention Mechanism, HAM)[72]：一种离散的注意力机制，它通过选择输入中的某些部分来聚焦，而不像软注意力那样通过加权来融合所有输入，它选择了最重要的部分并且忽略了其他部分。由于硬注意力涉及离散选择，它通常是不可微的，因此在训练时需要使用强化学习或近似方法来优化。常见的硬注意力机制有循环注意力模型(Recurrent Attention Model, RAM)等。

本文主要应用了软注意力机制，因此本小节将对该类型的注意力机制进行阐述。

(1) SE 注意力机制

SE 注意力模块主要关注不同特征通道之间的关系，执行压缩和激发操作，其具体体现在 MobileNetV3 的颈部结构中，结构如图 2-10 所示。其中 X 表示输入数据，其维度为$[H' \times W' \times C']$，分别代表输入特征的高、宽和通道数，通过F_{tr}卷积操作得到处理后的向量 U，其维度为$[H \times W \times C]$，F_{sq}表示 SE 模块的压缩操作，F_{ex}表示 SE 模块的激发操作，F_{scale}表示的对特征进行缩放操作，将 SE 模块的输出作用于卷积层的输出，得到最终的输出特征$\tilde{X}$。

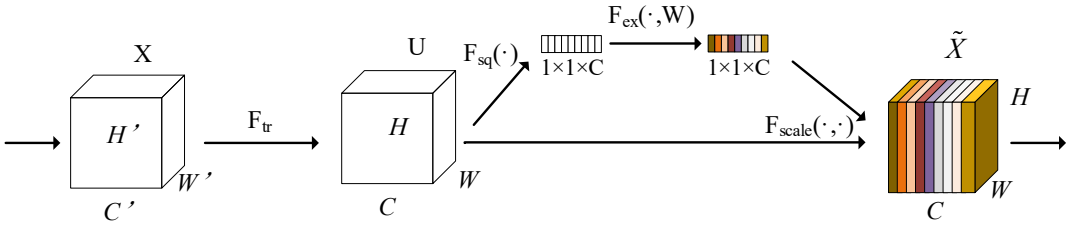


图 2-10 SE 模块结构图

将特征y作为输入，首先通过全局平均池化将特征$y \in R(W \times H \times C)$压缩成$y' \in R(1 \times 1 \times C)$。然后使用两个全连接层来拟合通道之间的相关性。采用 Sigmoid 进行归一化处理，通道的权重向量如式(2-6)所示：

$$f^c = \sigma(FC(\delta(FC(y'))))\tag{2-6}$$

其中，y'代表通过全局平均池化压缩后的特征，f^c代表通道的权重向量，FC 代表全连接层，δ代表 ReLU，σ代表 Sigmoid，SE 模块的输出如式(2-7)所示：

$$f' = f^c \cdot y\tag{2-7}$$

其中，y代表输入的特征，f'代表模块的输出。SE 模块执行了压缩和激发操作，通过自适应调整各个通道的权重来增强对特定特征通道的关注，同时减少对不相关特征通道的关注。这种机制有助于提高网络的性能和泛化能力，使网络更有效地学习和利用输入数据中的信息。

(2) 卷积注意力机制

卷积注意力模块(CBAM)[73]是一种高效且轻量化的前馈神经网络组件，通过协同整合通道和空间两个维度的注意力机制来增强特征表示能力，目的在于增强网络对重要特征的关注并提升模型性能。其结构如图 2-11 所示：

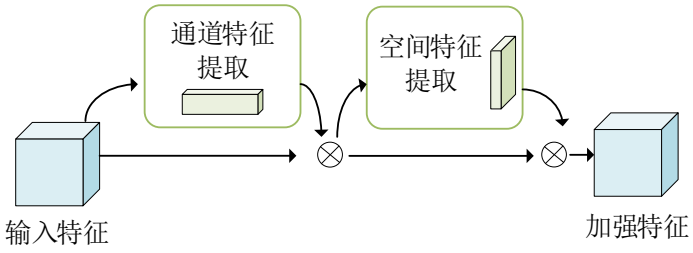


图 2-11 卷积注意力机制结构图

由上图可知，卷积注意力模块主要由通道注意力模块(Channel Attention Module, CAM)[74]和空间注意力模块(Spatial Attention Module, SAM)[75]两个部分构成，可以自适应地学习输入特征图中每个通道之间的相互依赖关系，并且在通道注意力模块和空间注意力模块的作用下，将这些特征图中的关键信息突出，进一步提高网络的性能。卷积注意力机制结合通道注意力与空间注意力，使模型能够自适应地聚焦于关键通道及空间区域，从而提升其特征提取能力，增强模型的表达与判别效果。由于该模块设计轻量，几乎不增加额外计算开销，因此可无缝集成至任意 CNN 架构，和基础网络一起进行端到端的训练，实现卷积神经网络的性能的优化。其中，通道注意力模块与空间注意力模块的工作机制如图 2-12 和图 2-13 所示。

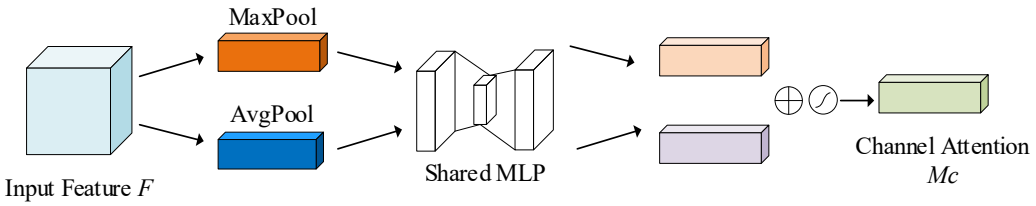


图 2-12 通道注意力模块

从图 2-12 中可以看出，通道注意力模块的核心原理主要是基于多尺度特征融合与动态权重分配，首先对输入特征图 F 分别执行全局最大池化和全局平均池化操作，分别生成两个具有全局上下文信息的一维特征向量；然后将两个特征向量输入共享全连接层(Shared MLP)中进行非线性变换，学习通道间的关系，最后，将 MLP 输出的两个特征向量通过逐元素相加的方式进行融合，生成通道注意力权重 Mc。该机制主要通过对输入特征图的通道进行加权，强调重要的通道信息，忽略不重要的通道信息。

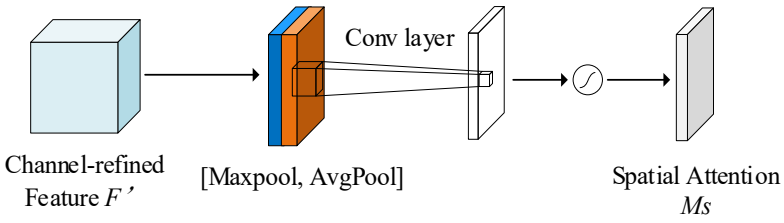


图 2-13 空间注意力模块

从图 2-13 中可以看出，空间注意力模块的核心原理主要是通过多步骤处理实现对特征图空间维度的动态加权，首先将经过通道注意力优化后的特征向量F'作为输入，依次通过全局最大池化和平均池化操作提取空间维度的全局上下文信息，通过单层卷积操作对池化结果进行特征融合与变换，最后，通过激活操作生成空间注意力权重 Ms。该机制能够捕捉特征图的空间结构信息，通过解析特征图的空间信息，对各位置进行加权，以增强重要区域的表达能力。

2.4 本章小结

本章主要介绍了人脸检测与表情识别的基础理论，首先介绍卷积神经网络的基本架构，并详细解析其各层结构、常用激活函数及损失函数。在此基础上，引出对神经网络模型的进一步探讨，详细介绍并分析了 YOLO、MobileNet、ResNet 和 DenseNet 模型的网络架构。然后介绍了涉及到的基础知识：深度可分离卷积、残差结构、密集连接以及注意力机制。通过本章的介绍可以对基于深度学习的人脸检测和表情识别的算法基础理论有一个较为全面的认识。

第三章 基于卷积神经网络的轻量化人脸检测算法

卷积神经网络在图像检测领域已经得到广泛应用，其中 YOLO、MobileNet 等为代表的经典网络模型极大的推动了该领域的研究与发展，在人脸检测任务中，深度学习方法已经取得了显著的效果。然而，随着卷积神经网络深度的增加，传统的深度学习模型如 YOLO 和 Faster R-CNN 存在计算量庞大、参数量多等问题，这使得这些网络难以在资源有限的设备上高效运行。因此，轻量化卷积神经网络如 MobileNetV3 成为了一个重要的研究方向，通过轻量化设计，能够显著降低计算复杂度和模型参数，同时保持较高的检测精度。

针对上述问题，本章提出一种基于卷积神经网络的轻量化人脸检测模型(MYDCFNet)。首先使用轻量型的 MobileNetV3 作为网络，同时在其颈部设计了一种新的激活函数 NHS，能在计算效率和非线性特性之间取得平衡。然后在特征融合过程中将使用提出的新型深度可分离卷积 NESC，通过动态调整深度卷积核的结构，更好地适应不同尺度的特征，采用多尺度空洞卷积并行结构的特征增强模块替代 SPP 模块，捕获更广泛的上下文信息。最后，为了加强特征提取，引入新的卷积注意力模块 DCAM，通过动态调整卷积核大小融合多尺度特征，使其更加高效地处理输入数据。

3.1 MYDCFNet 人脸检测网络架构设计

本节主要介绍构建的轻量化卷积神经网络人脸检测方法的总体框架。

图 3-1 是本章的人脸检测网络模型结构图，该网络主要由 MobileNetV3 主干网络、特征增强模块、PANet 网络以及最后的 Head Network 组成。(1) 图像首先输入主干网络采用二维卷积层进行初始特征提取，然后依次通过多个 bneck 模块，部分含有 SE 注意力机制增强关键通道特征响应，结合 NHS 激活函数以提取图像的多层次特征。(2) FEM 模块位于主干网络输出侧，并与 PANet 网络相衔接，用以增强特征表达能力。(3) PANet 实现多尺度特征融合，其中不同分辨率分支均采用了多次 NESC 堆叠，每个尺度分支中，在 NESC 堆叠后插入 DCAM 模块，在(52, 52)尺度分支上进行上采样操作，在(26, 26)尺度分支上依次进行上采样、下采样操作。(4) 最后在(52, 52)、(26, 26)和(13, 13)三个尺度上部署检测头，实现多尺度人脸检测。

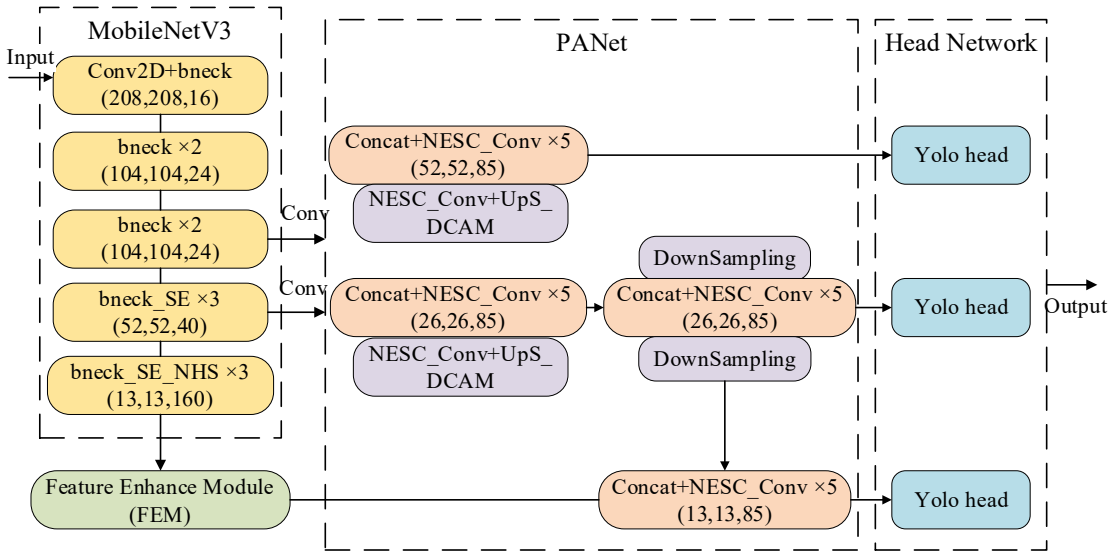


图 3-1 MYDCFNet 网络模型结构

本章在进行人脸检测研究的过程中首先将 YOLOv4 的主干提取网络 CSPDarkNet53 替换为轻量型的 MobileNetV3 网络，同时在特征融合过程中，将 PANet 网络中的普通卷积替换为深度可分离卷积(NESC)，提高特征提取效果，实现网络模型的轻量化，平衡速度和精度；然后，将原网络中的 SPP 模块删除，采用多尺度空洞卷积并行结构的特征增强模块，捕获更广泛的上下文信息；最后，设计了一个卷积注意力模块(DCAM)，通过动态调整卷积核大小，采用多个尺度的卷积核从而融合多尺度特征，并在每次上采样过程中，将特征输入 DCAM 模块，加强特征提取，构成了改进 MYDCFNet 网络模型结构。通过引入少量额外参数，提高网络的检测精度，同时在确保检测精度的前提下，实现网络结构的高效轻量化。

3.1.1 改进 MobileNetV3 主干网络

鉴于 MobileNetV3 网络模型轻量级的优势，本章改进 MYDCFNet 网络模型将 MobileNetV3 与 YOLOv4 进行融合，使用轻量型的 MobileNetV3 替代 YOLOv4 中的主干网络 CSPDarkNet53，减少模型参数量和计算量，达到网络模型的轻量化，提高系统的运行效率，在确保检测精度的前提下提高运行效率，同时使用了新的激活函数 NHS，在计算效率和非线性特性之间取得平衡，如图 3-2 所示。在主干网络 MobileNetV3 颈部结构的部分层级中设有 SE 注意力模块和 NHS 激活函数，SE 注意力模块可以执行压缩和激发操作，以增强对特定特征通道的关注，减少对不相关特征通道的关注，使网络更有效地学习和利用输入数据中的信息，NHS 激活函数可以在计算效率和非线性特性之间取得平衡，从而更好地适应不同深度学习任务和硬件环境。

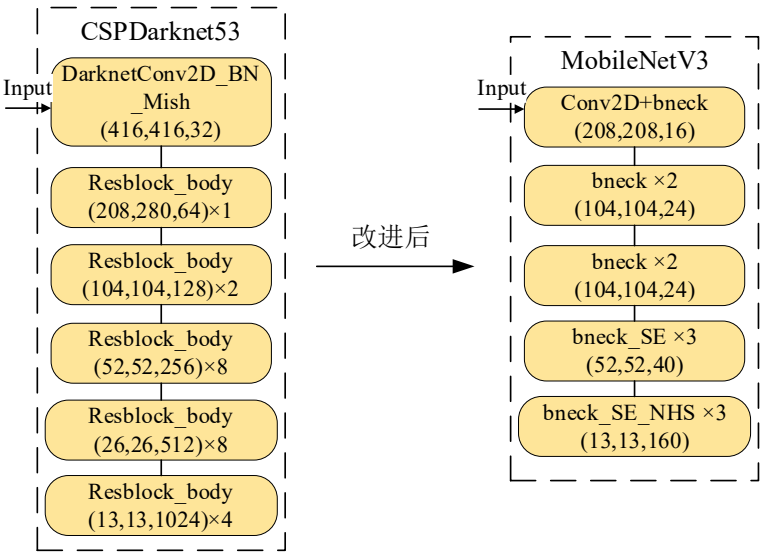


图 3-2 主干网络的改进

新模型的网络结构如表 3-1 所示。该网络输入图像的尺寸为 416×416×3，首先对输入图像进行普通卷积操作，将输入特征图转换为 16 个通道的特征图，中间层采用了一系列的 bottleneck(简写为 bneck)结构，每个 bottleneck 的输入特征图的通道数和尺寸在不同层级之间进行变化。由于 SE 模块会增加额外的计算量和参数量，设计时选择只在部分瓶颈层加入 SE 模块，主要在通道数较多、特征较复杂的层使用，用于增强特征图的重要信息。本发明进一步设计了一种新的激活函数 NHS，在所有 bottleneck 层级使用了 NHS 激活函数，增加了线性和非线性表达能力，表达式如式(3-1)、(3-2)所示：

$$NHS(x)=x\cdot\sigma(\alpha\cdot x) \tag{3-1}$$

$$\sigma(\alpha\cdot x)=\frac{1}{1+e^{-\alpha\cdot x}} \tag{3-2}$$

其中，x为输入，σ为 Sigmoid 函数，α为控制非线性程度的参数，e 为自然常数。NHS 激活函数借鉴了 Swish 的非线性特性，从而增强了模型对输入数据复杂关系的捕捉能力，进一步提升神经网络的表达效果。当α=0 时，NHS 激活函数变为 Hard Swish，此时的计算量较低，这是为了在需要计算效率的情况下使用 Hard Swish 的优势，随着α的增大，NHS 激活函数逐渐趋近于 Swish，保留了其非线性特性，增强模型的表达能力。NHS 激活函数可以通过调整α的大小，使 NHS 激活函数在计算效率和非线性特性之间取得平衡，从而更好地适应不同深度学习任务和硬件环境。

表 3-1 新模型基本网络结构

Input	Opetaror	Channels	SE	Nonlinearities	Stride
416×416×3	Conv2d	16	-	NHS	2
208×208×16	bneck,3×3	16	-	NHS	1
208×208×16	bneck,3×3	24	-	NHS	2
104×104×24	bneck,3×3	24	-	NHS	1
104×104×24	bneck,5×5	40	√	NHS	2
52×52×40	bneck,5×5	40	√	NHS	1
52×52×40	bneck,5×5	40	√	NHS	1
52×52×40	bneck,3×3	80	-	NHS	2
26×26×80	bneck,3×3	80	-	NHS	1
26×26×80	bneck,3×3	80	-	NHS	1
26×26×80	bneck,3×3	80	-	NHS	1
26×26×80	bneck,3×3	112	√	NHS	1
26×26×112	bneck,3×3	112	√	NHS	1
26×26×112	bneck,5×5	160	√	NHS	2
13×13×160	bneck,5×5	160	√	NHS	1
13×13×160	bneck,5×5	160	√	NHS	1
13×13×160	Conv2d,1×1	960	-	NHS	1
13×13×960	FEM	-	-	-	-

3.1.2 改进深度可分离卷积

基于 MobileNetV3 中的通用卷积模块替换为深度可分离卷积的想法，发现 PANet 模块中存在大量的普通卷积块。因此，本章提出了一种深度可分离卷积 NESC，并在 PANet 模块中用 NESC 替代传统的 3×3 卷积块。部分 NESC 卷积的替换示意如图 3-3 所示，能够有效降低模型的计算量和参数大小，以提高网络的检测速率和帧率。

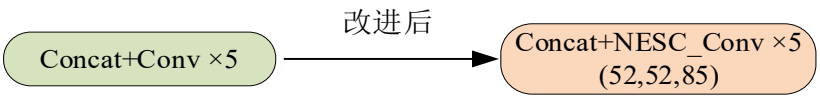


图 3-3 深度可分离卷积的替换

由于传统深度可分离卷积采用固定的深度卷积核结构，难以适应不同任务需求，因

此本章提出的 NESC 在传统深度可分离卷积的基础上引入了自适应演化模块(Evo)，使得网络能够根据任务需求动态优化调整卷积核的结构，增强特征提取能力，同时保持计算效率。具体的实现过程如式(3-3)至式(3-5)所示：

$$DW(x)_{i,j,k} = \sum_l x_{i,j,l} \cdot \omega_{i,j,l,k} \tag{3-3}$$

$$PW\left(Evo\left(DW(x)\right)\right)_{i,j,k} = \sum_l Evo\left(DW(x)\right)_{i,j,l} \cdot \omega'_{l,k} \tag{3-4}$$

$$NESC(x) = PW\left(Evo\left(DW(x)\right)\right) \tag{3-5}$$

其中x为输入，$NESC$表示深度可分离卷积操作，DW为深度卷积操作，PW为逐点卷积操作，Evo 为进化计算模块，i为行索引，j为列索引，k为通道索引，l为通道数，ω为深度卷积部分的权重参数，ω'为逐点卷积部分的权重参数。

由公式(3-3)至公式(3-5)可知，提出的 NESC 首先将输入特征图进行深度卷积，将每个通道进行卷积运算，得到具有深度卷积的特征图，然后引入自适应演化计算，通过 *Evo* 计算模块对深度卷积的输出进行调整，使得卷积操作更加适应任务需求，最后对进化计算调整后的深度卷积输出进行逐点卷积操作，逐点卷积在 Evo 计算模块的帮助下整合了深度卷积的输出，使其能够更加高效的整合通道信息，提高检测和识别的性能，生成最终的输出特征图特征。

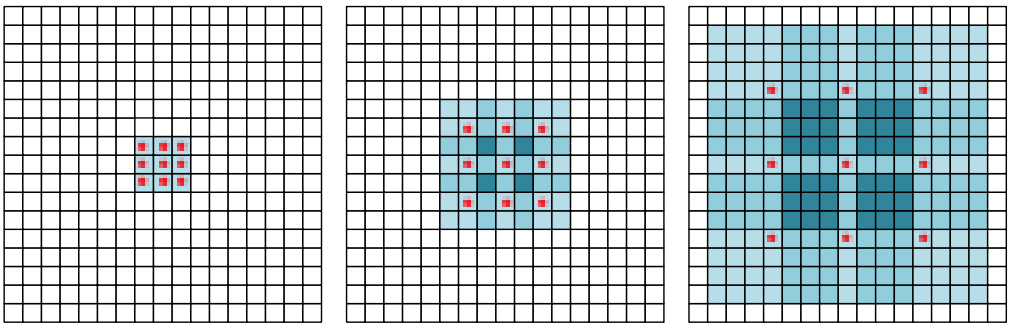
通过进化计算，NESC 能够自动调整深度卷积核的结构，以适应特定任务的需求。这种自适应性使得 NESC 更具灵活性和适应性，可以优化深度卷积核的设计，以更好地捕捉输入特征中的相关信息，同时 NESC 延续了深度可分离卷积的轻量化设计，通过将深度卷积与逐点卷积分解为两个独立步骤，从而有效降低计算复杂度和参数规模。

3.1.3 特征增强模块

(1) 空洞卷积

空洞卷积(Dilated Convolution)[76]是卷积神经网络中的一种特殊的卷积运算方式，相较于传统卷积，其特点是在卷积核内部引入间隔，从而在不增加参数量的情况下显著扩大感受野，同时有效保留特征图的分辨率。这使得空洞卷积在处理图像和信号时具有独特的优势，特别是在处理具有大尺度和多尺度特征的任务中。在不同膨胀率(Dilation Rate)设置下，空洞卷积的感受野变化如图 3-4 所示。其中，图(a)采用 3×3 的空洞卷积核，与标准卷积核的感受野相同；当图(b)的膨胀率为 2 时，感受野扩大至 7×7；当图(c)的膨胀率进一步增大至 4 后，感受野可达到 15×15。

因此，空洞卷积是一种有助于神经网络处理多尺度信息和大尺度感受野的重要工具。它在许多计算机视觉任务中被广泛使用，特别是在深度卷积神经网络中，以增强网络的感知范围和表征能力，对于改进图像处理和分析任务的性能非常重要。



(a) 3×3　　(b) 7×7　　(c) 15×15

图 3-4 空洞卷积示意图

(2) 网络结构

特征增强模块(Feature Enhance Module, FEM)是一种用于增强深度卷积神经网络特征表示能力的关键组件[77]，它在图像处理和计算机视觉任务中具有广泛的应用，旨在提高模型性能、准确性和鲁棒性。

在特征增强模块的设计中，首先通过 1×1 卷积核对上层特征图的通道维度进行调整，以确保与当前通道数相一致。随后，对特征图采用上采样操作，使其在尺寸上与当前层的特征图相匹配，并通过逐元素乘法进行特征融合。接着，将融合后的特征图划分为三部分，分别输入至不同膨胀率的空洞卷积层，以提取多尺度信息。最终，将各空洞卷积的输出结果整合，充分利用不同层的特征信息，从而提升特征的区分能力与模型的鲁棒性。

在 YOLOv4 网络中的空间金字塔池化(SPP)模块用于处理主干网络输出的特征图，其通过多尺度池化操作扩展特征图的感受野范围，实现局部细节特征与全局上下文信息的有效融合。然而针对人脸检测等一些目标尺度差异较小的情况，层间特征交互虽然有助于增强特征的区分能力，提高模型的稳健性，但是，最大池化操作可能导致部分关键信息的丢失，同时，单纯扩大感受野可能引入过多背景或冗余信息，影响特征的有效性。所以本章提出一种采用多尺度空洞卷积并行结构的特征增强模块来替代 SPP 模块，如图 3-5 所示。

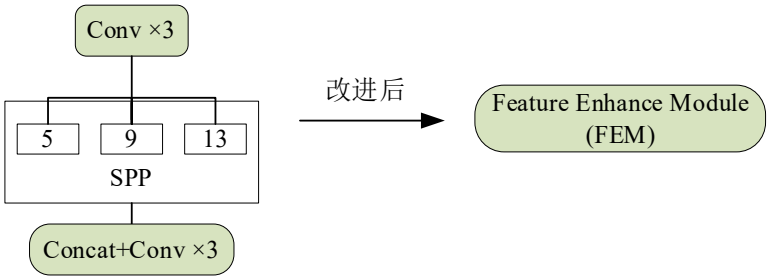


图 3-5 特征增强模块的改进

特征增强模块能够在不增加网络参数的情况下，扩大卷积层的感受野。另外，模块采用空洞卷积，能够在不降低特征图分辨率的情况下，增大特征图的感受野，通过将网络中不同层次的特征图进行多尺度融合，不仅能够整合浅层的细节信息和深层的语义信息，还能增强主干网络输出特征的表达能力。

特征增强模块的结构如图 3-6 所示，该模块通过多分支并行结构来丰富特征表达，输入特征首先经过 1×1 卷积压缩后分成多路分支，每路分支采用不同膨胀率(DR=1、2、3)的 3×3 卷积捕获不同尺度的上下文信息。随后将各分支的输出特征进行堆叠(Concat)，再用 1×1 卷积进行通道压缩并与输入特征形成残差连接，从而在保持计算效率的同时增强对多尺度特征的感知能力。

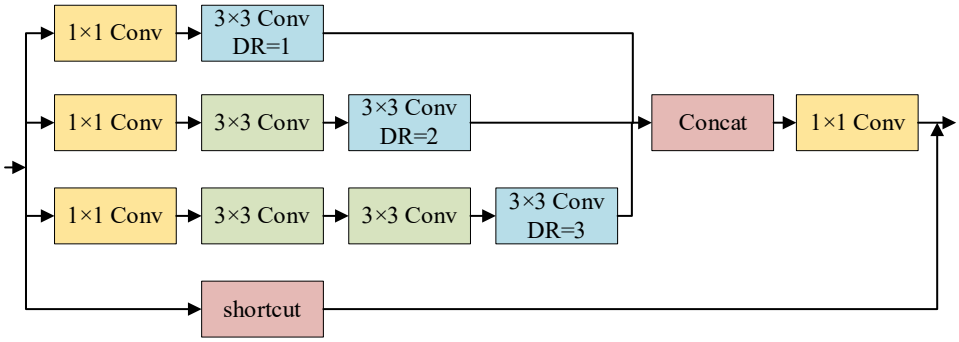


图 3-6 特征增强模块结构

在各路分支内，每路分支由不同数量的卷积层构成，特征图经过不同深度的卷积操作后，能够获得不同层次的语义信息。其中，旁路剪枝(shortcut)是一种残差连接，用于解决深层网络训练中的梯度消失和梯度爆炸，其可以通过移除一些特定的 shortcut 连接，进一步减少网络的计算负担。另外，该特征增强模块通过引入空洞卷积，在保持特征图空间分辨率不变的前提下，显著扩大了特征图的感受野范围。模块采用了膨胀率分别为 1、2、3 的多分支空洞卷积结构，每个分支以不同的采样间隔提取特征信息，从而捕获多尺度的上下文特征，最终通过通道维度上的特征拼接操作，将各分支的输出进行融合，生成具有丰富语义信息的增强特征图。这种设计不仅避免了池化操作导致的分辨率损失，

还通过多尺度特征的有效整合，显著提升了模块的特征表达能力。随着每个分支的空洞卷积的膨胀率逐步递增，各分支能获得不同的感受野范围，有利于捕获更广泛的上下文信息，从而显著增强模型的感知能力。在对各分支的输出特征图进行级联后，网络深层的空间信息与语义信息得以融合，使特征表达更加丰富。

3.1.4 注意力模块

本文网络模型采用两种注意力机制。第一种注意力机制是 MobileNetV3 主干网络颈部结构中的 SE 模块，第二种是注意力机制是在 PANet 网络上改进的 DCAM 模块，在每次上采样的过程中引入 DCAM 模块，加强特征提取。

(1) SE 模块

SE 模块执行了压缩和激发操作，通过这种方式能够增强对特定特征通道的关注，同时减少对不相关特征通道的关注。这种机制有助于提高网络的性能和泛化能力，使网络更有效地学习和利用输入数据中的信息。通道的权重向量和 SE 模块的输出表达式分别如公式(3-6)和公式(3-7)所示：

$$f^{c}=\sigma(FC(\delta(FC(y^{'})))) \tag{3-6}$$

$$f^{'}=f^{c}\cdot y \tag{3-7}$$

其中，y代表输入的特征，y'代表通过全局平均池化压缩后的特征，f'代表 SE 模块的输出，f^{c}代表通道的权重向量，FC 代表全连接层，δ代表 ReLU，σ代表 Sigmoid。

(2) DCAM 模块

设计了一个新的卷积注意力模块 DCAM，通过动态调整多个尺度的卷积核，融合多尺度特征，使模块能够更全面地捕获局部和全局特征，并在每次上采样过程中，将特征输入 DCAM 模块，加强特征提取，在轻量化后的网络中加入 DCAM 模块，如图 3-7 所示，在每次上采样过程中，将特征输入 DCAM 模块，加强特征提取。

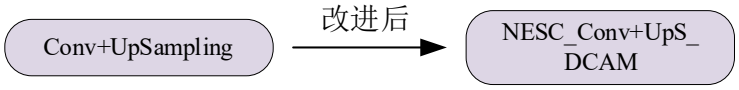


图 3-7 引进 DCAM 模块

DCAM 首先在不同卷积核上生成多尺度的特征，将多尺度特征在通道维度进行拼接，形成融合特征，与单一尺度的卷积核相比，多尺度卷积核可同时捕捉局部与全局信息，从而对不同层次的特征更加敏感。然后将特征在通道维度和空间维度进行两次加权操作，与多特征融合的特征逐元素相乘，得到最终的输出，这样可以使网络突出关键通

道与空间位置，增强特征表达。具体的实现过程如公式(3-8)至公式(3-12)所示：

$$X_i = Conv\left(X, size_i\right) \tag{3-8}$$

$$X_m = Concat\left(\left[X_1, X_2, ..., X_N\right]\right) \tag{3-9}$$

$$F^C = \sigma\left(FC\left(y\left(X_m\right)\right)\right) \tag{3-10}$$

$$F^{CS} = F^C \otimes \sigma\left(FC\left(y\left(X_m\right)\right)\right) \tag{3-11}$$

$$F = F^{CS} \otimes X_m \tag{3-12}$$

其中，X表示输入的特征，$size_i$表示第 i 个卷积操作所使用的卷积核大小，X_i表示自适应卷积核调整的第i个尺度的特征，i=1, 2, …, N，N 表示尺度的总数，X_m表示多个尺度特征拼接后的特征，σ代表 Sigmoid，F^C表示将通道特征图与 F 相乘所得到的结果，F^{CS}表示将空间特征图与F^C相乘所得到的结果，FC 代表全连接层，y代表通过全局平均池化压缩后的特征，$\otimes$表示逐元素相乘，F表示最终的特征表示。

DCAM 模块如图 3-8 所示。通道注意力模块与空间注意力模块采取串行方式对输入特征进行处理，首先对输入特征进行合并得到特征X_m，输入到通道注意力模块和空间注意力模块中分别计算通道维度和空间维度上的权重系数，最后将这两个维度的注意力权重与原特征逐元素相乘，实现对特征图的通道和空间双重校正，从而获得优化后的特征。这种串行处理能够同时突出对重要通道的关注与对关键空间位置的强调，实现对多维度特征的综合增强。

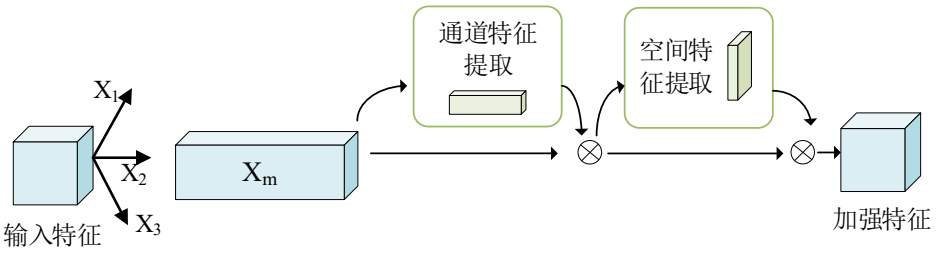


图 3-8 DCAM 模块

DCAM 模块通过动态调整多个尺度的卷积核，融合多尺度特征，使 DCAM 模块能够更全面地捕获局部和全局特征，并在每次上采样过程中，将特征输入 DCAM 模块，加强特征提取，DCAM 模块通过自适应感受野调整和多尺度特征整合，能够全面感知输入特征的不同尺度和层次的信息，提高了改进 MYDCFNet 网络模型的泛化能力。

3.2 实验数据集及相关配置

3.2.1 实验环境配置及训练细节

本研究所有实验均配置在 GPU 显卡型号为 NVIDIA GeForce RTX 4060 和 CPU 处理器型号为 Intel Core i7-14650HX@2.20GHz 的计算机上完成。其详细的系统配置环境如表 3-1 所示，完全满足深度学习实验的计算要求。实验环境基于 Anaconda3 构建，采用 Python 3.8.3 为编程语言，并使用 PyCharm 作为集成开发环境，在 Keras 深度学习框架下实现了卷积神经网络的搭建与训练。Keras 提供了便捷的神经网络构建方式，并支持在 CPU 与 GPU 之间切换运行，以提高计算效率并实现并行处理。

表 3-1 系统环境配置

环境配置	型号版本
CPU 处理器	Intel Core i7-14650HX@2.20GHz
内存	32GB
GPU 显卡	NVIDIA GeForce RTX 4060
CUDA 版本	11.1
CUDNN 版本	11.1
Python 版本	3.8.3
Keras 版本	2.2.5

本章基于 Python 的 Keras 库搭建了所提出的网络模型，在训练过程中设定迭代次数(epoch)为 100，设定训练批量大小(batch_size)为 32，为了优化模型的收敛性能，采用 Adam 优化器进行参数更新，其中设定动量系数(momentum)为 0.9，权值衰减率(weight decay)调整为 0.0005，设定初始学习率为 0.001，以优化模型收敛效果。为优化模型训练过程并防止过拟合，采用了一种动态学习率调整策略，当评估指标未表现出上升趋势时，将自动将学习率降低至初始值的 10％。在训练过程中，每个 epoch 都会遍历全部训练样本，当训练损失值低于验证损失时，立即停止训练，否则，模型将继续训练直至达到预设的最大迭代。

3.2.2 实验数据集

(1) 数据集介绍

本章选用的人脸数据集是网上公开的 Labeled Faces in the Wild(LFW)人脸数据集和

WIDER Face 人脸数据集。LFW 数据集包含约 13000 张人脸图像，每幅图像的分辨率为 250×250，并附带相应的身份标注。WIDER Face 作为常用的人脸检测数据集，涵盖了超过 32000 张图像，广泛用于相关研究，其中包括人脸数量从一到数百不等的不同规模的人脸，每个图像都有对应的注释文件，其中包含每个人脸的边界框坐标和人脸所属类别。其中 LFW 和 WIDER Face 数据集部分人脸样本分别如图 3-9 和 3-10 所示。

图 3-9 LFW 数据集部分人脸图像样本

图 3-10 WIDER Face 数据集部分人脸图像样本

(2) 数据集标注

获取的人脸图像部分数据集没有人脸的标注信息，所以需要对人脸信息进行标注。数据集需通过采用开源的标注工具 LabelImg 对人脸信息进行标注。LabelImg 标注数据的界面如图 3-11 所示。标注包括“人脸矩形区域位置”和“人脸语义信息”，即脸部及其周围肩膀区域的矩形区域位置，使用 PASCAL VOC 格式，标注后生成 XML 文件；使用 Python 编写程序将 XML 文件中保存的标注位置坐标转换为对应的文本文件，标注后的部分标注数据如图 3-12 所示。

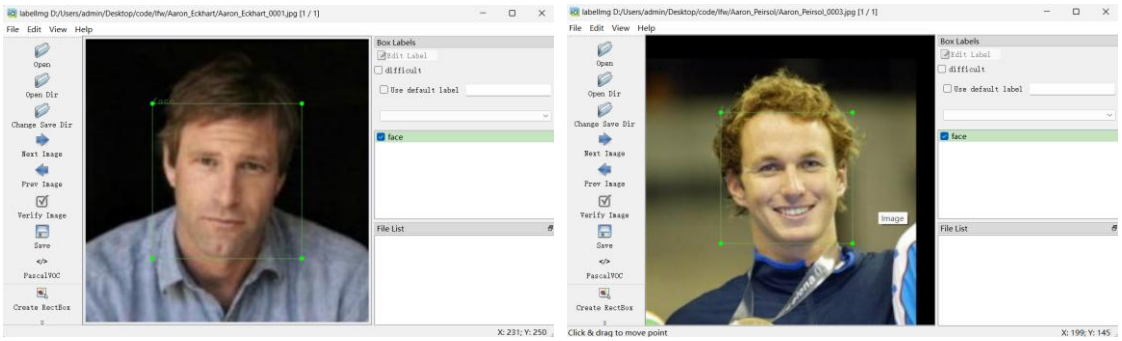


图 3-11 LabelImg 的标注界面

```
<annotation>
  <folder>Aaron_Eckhart</folder>
  <filename>Aaron_Eckhart_0001.jpg</filename>
  <path>D:\Users\admin\Desktop\code\lfw\Aaron_Eckhart\Aaron_Eckhart_0001.jpg</path>
  <source>
    <database>Unknown</database>
  </source>
  <size>
    <width>250</width>
    <height>250</height>
    <depth>3</depth>
  </size>
  <segmented>0</segmented>
  <object>
    <name>face</name>
    <pose>Unspecified</pose>
    <truncated>0</truncated>
    <difficult>0</difficult>
    <bndbox>
      <xmin>53</xmin>
      <ymin>57</ymin>
      <xmax>185</xmax>
      <ymax>194</ymax>
    </bndbox>
  </object>
</annotation>
```

```
<annotation>
  <folder>Aaron_Peirsol</folder>
  <filename>Aaron_Peirsol_0003.jpg</filename>
  <path>D:\Users\admin\Desktop\code\lfw\Aaron_Peirsol\Aaron_Peirsol_0003.jpg</path>
  <source>
    <database>Unknown</database>
  </source>
  <size>
    <width>250</width>
    <height>250</height>
    <depth>3</depth>
  </size>
  <segmented>0</segmented>
  <object>
    <name>face</name>
    <pose>Unspecified</pose>
    <truncated>0</truncated>
    <difficult>0</difficult>
    <bndbox>
      <xmin>64</xmin>
      <ymin>66</ymin>
      <xmax>178</xmax>
      <ymax>179</ymax>
    </bndbox>
  </object>
</annotation>
```


图 3-12 数据标注文件

3.2.3 多尺度训练策略

多尺度训练策略是一种用于目标检测领域的训练方法，通过在训练过程中使用不同的输入图像尺寸来提高检测器的性能和稳定性。在该策略中，训练图像被随机缩放到不同的尺寸，以增加训练样本的丰富性。同时，多尺度训练可以使检测器更好地适应不同尺寸的目标，并能够在不同的场景中进行更好的检测。

将标注后的人脸图像构建数据集并进行预处理，引入多尺度训练策略对原始数据集进行了缩放和裁剪并进行归一化处理，将不同尺度的图像作为训练样本，增加数据集的多样性。本文的数据集都引入了多尺度训练策略来增加数据集的多样性，使得模型可以更好地适应各种尺度下的人脸检测任务，以提高模型对不同尺度人脸的检测能力，从而提高模型的检测性能和鲁棒性，多尺度训练流程图如图 3-13 所示。

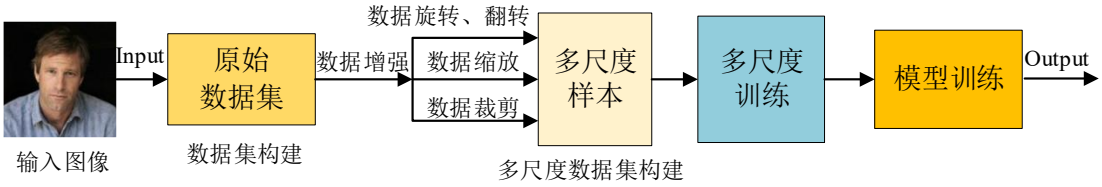


图 3-13 多尺度训练流程图

3.2.4 评价指标

为了能够进一步全方面的评价人脸检测网络模型的性能，本章采用准确率(Accuracy)、精确率(Precision)、灵敏度(Sensitivity)、F1_Score、参数量(Parameter)、计算量(GFLOPs)和帧率(FPS)作为实验结果的评价指标。其中参数量表示网络计算参数量，FPS 表示每秒识别图片的帧数。各评价指标的计算公式定义如式(3-13)至式(3-17)所示：

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3-13}$$

$$Precision = \frac{TP}{TP + FP} \tag{3-14}$$

$$Sensitivity(Recall) = \frac{TP}{TP + FN} \tag{3-15}$$

$$F1\_Score = 2 \times \frac{Precision \times Sensiticity}{Precision + Sensiticity} \tag{3-16}$$

$$FPS = \frac{M}{S} \tag{3-17}$$

公式(3-13)至(3-17)中，*FP* 表示实际为负但被预测为正的样本数量，*TN* 表示实际为负被预测为负的样本的数量，*TP* 表示实际为正被预测为正的样本数量，*FN* 表示实际为正但被预测为负的样本的数量，*M* 表示图像总数，*S* 表示识别过程中处理图像的时间。

3.3 消融实验

为了验证本章提出的 MYDCFNet 网络中各个模块对网络参数和检测性能产生的影响，本节设计了消融实验，在 YOLOv4 网络的基础上依次加入各个模块，在保证实验条件相同的情况下进行实验验证。

3.3.1 改进前后网络模型参数变化

M3-YOLOv4 网络将 YOLOv4 网络中的主干提取网络 CSPDarkNet53 替换成了 MobileNetV3，并引用了新的 NHS 激活函数，通过动态调整参数，在计算效率和非线性特性之间取得平衡。而改进的 M3-YOLOv4-D 网络结构基于 M3-YOLOv4 网络，将传统卷积替换成了新型深度可分离卷积 NESC。改进 MYDCFNet 网络模型进一步将 M3-YOLOv4-D 网络中的 SPP 模块删除，加入特征增强模块 FEM，另外在 PANet 网络中引入 DCAM 注意力模块，可以动态调整卷积核大小，适应不同尺度的输入特征，融合多个尺度的特征。改进网络的模型参数如表 3-2 所示。

表 3-2 网络模型参数对比

模型	参数量	Model Size(MB)	GFLOPs	FPS
YOLOv4	64.43M	246.68	8.81	61.12
M3-YOLOv4	40.04M	154.26	6.52	122.26
M3-YOLOv4-D	11.80M	46.24	7.59	131.87
MYDCFNet	11.81M	47.51	7.91	102.68

由表 3-2 可知，由于 YOLOv4 原主干网络的 CSPDarknet53 的参数量较大，所以在将主干网络替换为 MobileNetV3 后，M3-YOLOv4 模型的参数量显著降低，模型大小减少到 154.26MB，相较于原始模型减少了约 37 %。且检测速率明显提高近一倍。当将特征提取网络(PANet)中的普通卷积替换为深度可分离卷积 NESC 时，改进 MYDCFNet 模型参数量大大减少，模型大小相较于原始模型(YOLOv4)显著减少了约 81 %，检测速率进一步提高。在 MoblieNetV3 网络中进一步引入 FEM 特征增强模块和在 PANet 网络中引入 DCAM 注意力机制后，相比于 M3-YOLOv4-D，帧率有所下降，计算量有一定增加，而参数量和模型大小只有细微变化，说明引入 FEM 模块和 DCAM 模块对轻量化后的网络整体参数和模型影响较小，可以在保证一定检测精度的情况下，减少网络参数量，压缩了网络模型大小，实现了对网络结构的轻量化。

3.3.2 改进前后网络检测性能的对比

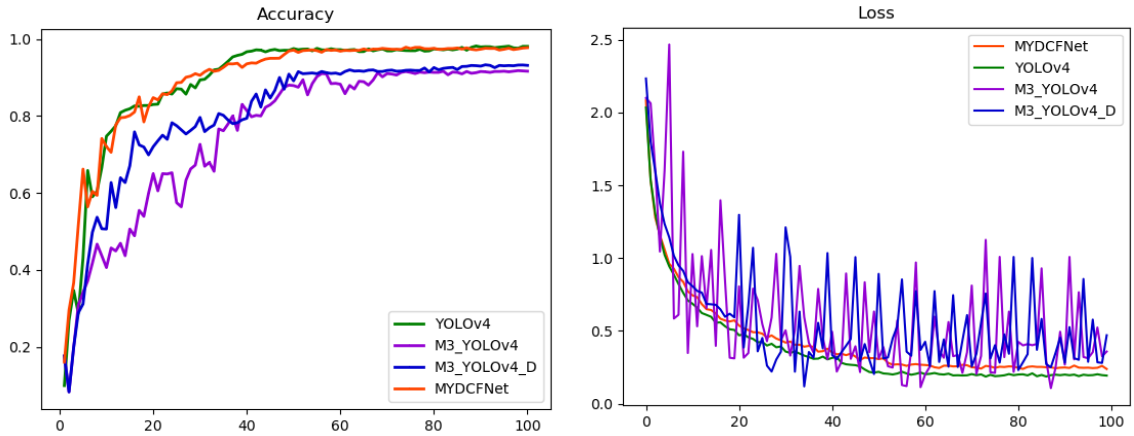
由于将主干提取网络替换为轻量型网络 MobileNetV3 后，主干提取网络的深度会随之减小，从而导致对输入特征提取部分的计算能力变弱。虽然模型的整体参数量大大减少，检测速率大大提升，但精度可能会随之下降。而且当将特征提取网络 PANet 中的普通卷积替换为深度可分离卷积时，虽然检测速率进一步提升，但由于网络特征提取的信息不丰富，总而导致精度进一步下降。所以为了保证特征提取的准确性，在 M3-YOLOv4-D 网络模型进一步引入 FEM 特征增强模块和 DCAM 注意力机制，构成了 MYDCFNet 网络模型，增强了网络的特征感知能力和信息表达能力，加强了网络在特征提取过程中的信息的整合，改进的各网络模型性能如表 3-3 所示。

表 3-3 不同改进网络模型的性能对比

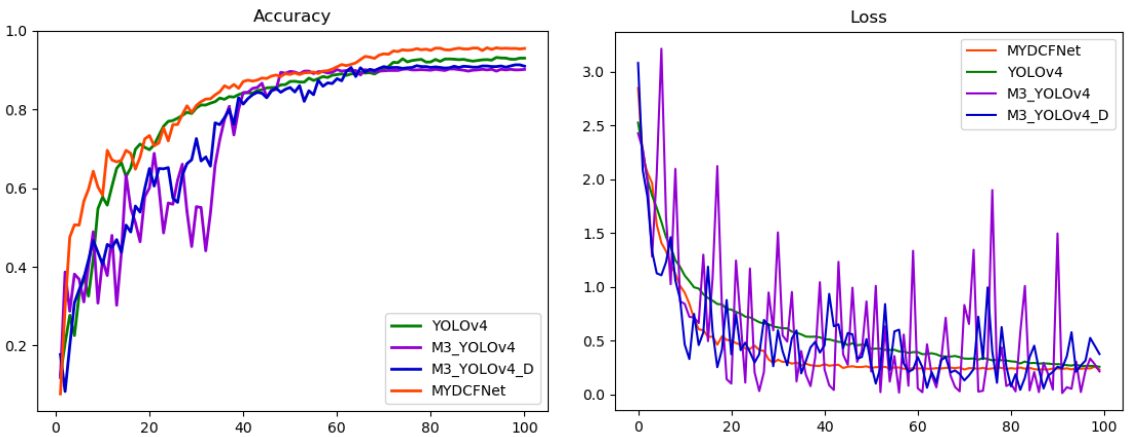
数据集	模型	主干网络	注意力机制	NESC	FEM	Acc(%)	FPS
LFW	YOLOv4	CSPDarkNet53	×	×	×	97.21	61.12
	M3-YOLOv4	MobilenetV3	×	×	×	93.68	122.26
	M3-YOLOv4-D	MobilenetV3	×	√	×	92.42	131.87
	MYDCFNet	MobilenetV3	DCAM	√	√	97.96	102.68
WIDER Face	YOLOv4	CSPDarkNet53	×	×	×	96.65	60.31
	M3-YOLOv4	MobilenetV3	×	×	×	91.25	120.12
	M3-YOLOv4-D	MobilenetV3	×	√	×	90.34	128.58
	MYDCFNet	MobilenetV3	DCAM	√	√	97.11	101.66

从以上结果来看，将主干网络轻量化以及将传统卷积替换为深度可分离卷积可以有效的减少网络模型的参数量，提高检测的帧率和检测速率，但检测精度会随之下降。而引入注意力机制后，能够在保证一定检测精度的情况下，大大减少网络参数量，显著提高帧率和检测速率，具有较好的实时性。

根据改进的 MYDCFNet 模型的实验结果，绘制了改进的几种方法在 LFW 和 WIDER Face 人脸数据集上的准确率曲线和损失曲线，如图 3-14 所示。改进的 MYDCFNet 网络模型的准确率验证曲线随着迭代次数的增加稳定上升，在经过 100 次迭代后趋于稳定，在模型参数量和计算量显著减少的情况下，也具有较高的检测精度，并且相较于其他几种改进的模型具有最高的检测精度。在损失曲线图像上，M3-YOLOv4 和 M3-YOLOv4-D 的损失曲线都存在明显的震荡现象，这表明在训练过程中这两个模型的优化难度较大。而改进的 MYDCFNet 模型震荡频率更小，并且在震荡过程中呈现稳定下降趋势。且模型收敛速度较快，当迭代次数在 60 次左右时，损失值基本稳定在 0.3 左右，之后网络结构趋于收敛，与改进之前的 M3-YOLOv4 和 M3-YOLOv4-D 模型相比，损失曲线震荡频率明显减小，在训练过程中模型可以快速收敛。



(a) LFW Dataset



(b) WIDER Face Dataset

图 3-14 改进网络的准确率曲线和损失曲线对比

3.4 实验结果与分析

3.4.1 不同网络检测性能的对比

为了进一步评估改进 MYDCFNet 网络模型的检测性能，将该改进 MYDCFNet 网络模型在 LFW 和 WIDER Face 人脸数据集上分别与主流识别深度学习模型(GoogleNet、Xception、Inception-Resnet-V1、Inception-Resnet-V2、Faster R-CNN、VGG19、AlexNet)完成对比实验。表 3-4 展示了本模型及对比模型在数据集 LFW 上的检测精度及各项性能参数，表 3-5 展示了在数据集 WIDER Face 的各项性能指标。如表 3-4 和表 3-5 所示，改进的 MYDCFNet 模型相对于其他模型在检测帧率上有明显提升。改进的 MYDCFNet 模型在 LFW 和 WIDER Face 人脸数据集上的检测精度最高，达到了 97.96 %和 97.11 %，检测帧率大大提升。相较于检测精度较差的 AlexNet 模型，改进的 MYDCFNet 模型检测精度提升了近 5 %。综上所述，MYDCFNet 模型在准确率和检测速度上都有了很大的提高，FPS 的值处于领先位置。所以本章改进的 MYDCFNet 模型的整体性能优势更加突出，能够同时满足识别精度和速度的需求。

基于 MYDCFNet 模型与七种对比算法的实验结果，分别在 LFW 和 WIDER Face 数据集上绘制了所有方法的准确率曲线。为了更清晰地展示模型的性能特点，将准确率验证曲线分为两个部分进行绘制。图 3-15 展示了 MYDCFNet 与非轻量级模型(Inception-ResNet-V1、Faster R-CNN、VGG19 和 AlexNet)的准确率验证曲线对比图，图 3-16 展示了 MYDCFNet 与轻量级模型(Inception-ResNet-V2、GoogleNet 和 Xception)的准确率验证曲线对比结果图。

表 3-4 LFW 数据集不同网络模型检测结果

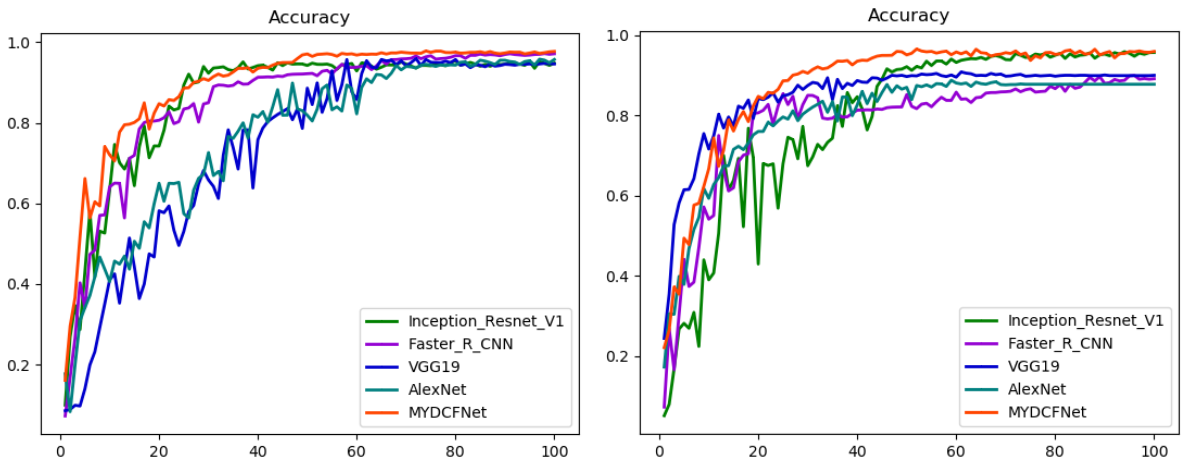
数据集	模型	Acc(%)	Sens(%)	Prec(%)	F1(%)	FPS
LFW	GoogleNet	97.11	95.45	97.23	96.51	64.22
	Xception	96.63	95.11	96.77	95.64	54.17
	Inception-Resnet-V1	96.59	95.21	96.64	95.45	58.13
	Inception-ResNet-V2	97.21	96.47	97.22	95.36	61.12
	Faster R-CNN	96.85	95.38	96.89	91.56	45.26
	VGG19	95.48	94.36	95.53	91.76	43.85

	AlexNet	93.61	93.19	93.65	89.73	83.45
	MYDCFNet	97.96	97.14	97.87	96.19	102.68

表 3-5 WIDER Face 数据集不同网络模型检测结果

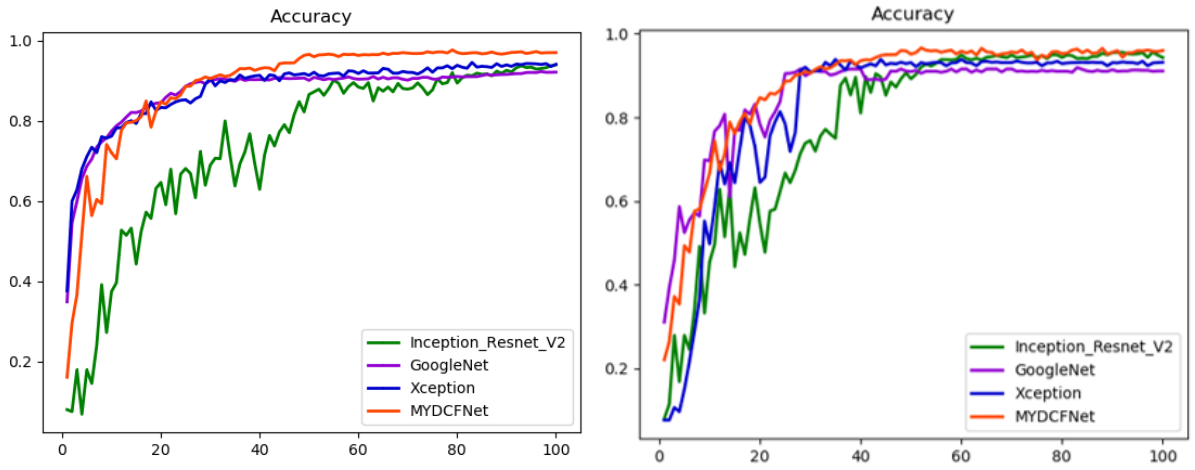
数据集	模型	Acc(%)	Sens(%)	Prec(%)	F1(%)	FPS
	GoogleNet	96.64	95.39	96.67	95.84	61.74
	Xception	95.89	95.02	96.03	94.48	52.45
	Inception-Resnet-V1	96.12	95.54	96.15	94.96	56.96
WIDER	Inception-ResNet-V2	96.65	96.01	96.74	93.22	60.31
Face	Faster R-CNN	95.43	94.15	95.51	89.57	44.81
	VGG19	94.87	93.56	94.95	88.25	43.40
	AlexNet	92.36	92.08	92.41	86.84	82.12
	MYDCFNet	97.11	97.56	98.14	96.14	101.66

实验结果表明，相较于传统轻量级网络，改进后的 MYDCFNet 模型在图像深度特征提取方面表现出更高的效率和稳定性，其验证曲线随着训练迭代的进行呈现出平稳上升的趋势；与计算复杂度较高的非轻量级网络相比，改进后的 MYDCFNet 模型在显著减少参数量的同时仍保持了优异的性能，经过 100 个 epoch 的训练，MYDCFNet 模型准确率曲线随着迭代次数的增加稳定上升，准确率曲线也更加平稳。MYDCFNet 模型的检测精度高于对比的网络结构，对人脸图像的检测更加高效。



(a) LFW Dataset (b) WIDER Face Dataset

图 3-15 MYDCFNet 和非轻量级网络的准确率验证曲线对比



(a) LFW Dataset　　(b) WIDER Face Dataset

图 3-16 MYDCFNet 和轻量级网络的准确率验证曲线对比

3.4.2 人脸检测结果

将上述训练的模型用于人脸检测，检测结果如图 3-17 所示。可以看到本章设计的改进 MYDCFNet 模型在人脸检测方面取得了较好的效果。检测框准确的框出了图像中的人脸，并且能够适应不同尺度、姿态、场景和光照条件下的人脸，并且检测的准确率普遍较高。这表明本章设计的人脸检测模型具有很好的泛化能力和鲁棒性，为人脸识别和其他相关任务的研究和应用提供了良好的基础。

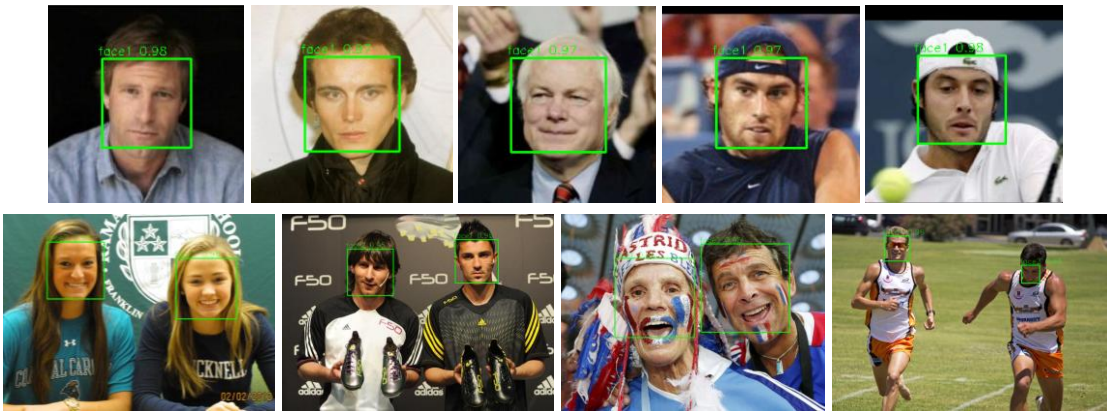


图 3-17 人脸检测结果图

3.5 本章小结

本章首先提出了一种基于改进的 MYDCFNet 网络的人脸检测方法，可以在确保检测精度的前提下，实现网络结构的高效轻量化。首先将 YOLOv4 的主干提取网络 CSPDarkNet53 替换为轻量型的 MobileNetV3 网络，同时在主干网络 MobileNetV3 的颈部设计了一种新的激活函数 NHS，通过调整α的大小，使新激活函数在计算效率和非线

性特性之间取得平衡，然后设计一种新型深度可分离卷积 NESC，在特征融合过程中将 PANet 网络中的普通卷积替换为 NESC，通过动态调整深度卷积核的结构，更好地适应不同尺度的特征，提高泛化能力；然后提出一种采用多尺度空洞卷积并行结构的特征增强模块替代 SPP 模块，捕获更广泛的上下文信息；最后，设计一个新的卷积注意力模块 DCAM，通过动态调整卷积核大小融合多尺度特征，在每次上采样过程中，将特征输入卷积注意力模块加强特征提取，通过仅引入少量的参数来提高轻量化网络的损伤检测精度，在保证一定检测精度的情况下，减少网络参数量，提高帧率和检测速率。在数据集的构建时引入多尺度训练策略，通过在训练过程中随机缩放输入图像的尺寸，使得模型可以更好地适应各种尺度下的人脸检测任务，最终的人脸检测准确率达到了 97 %，模型参数较原始模型显著减少了约 80 %，计算速度提升了近 67 %，检测精度提升了近 3 %，与其他人脸检测方法比较，本章提出人脸检测方法具有很好的泛化能力和鲁棒性，可以达到实时检测的要求，验证了提出的改进 MYDCFNet 网络在人脸检测任务上的有效性和可行性，可以改善在人脸检测过程中存在的模型参数大，检测速率慢等问题。

人脸检测是人脸图像研究的基础环节，为后续表情识别等进一步研究提供了良好的基础，其精度和效率对后续任务如表情识别等也有重要影响。表情是人类传递情绪、表达内心状态的重要方式，而表情识别是判断人类情真实绪的重要手段，在很多方面都具有重要的应用价值和社会意义。因此，基于本章人脸检测的研究之后，下一章将进一步进行人脸表情识别任务的研究。

第四章 基于多级特征提取和融合的人脸表情识别算法

人脸检测作为人脸表情识别的前置步骤，其精确性直接影响后续表情识别的可靠性和有效性，因此在人脸检测模型 MYDCFNet 的基础上，本章进一步开展对于人脸表情识别任务的研究。近年来，为提升人脸表情识别的准确性，众多研究者通过构建新颖的网络架构、改进损失函数等手段在该领域取得了显著进展，并有效提升了识别精度，但目前的表情识别算法仍存在不能很好提取各层次的特征和细节以及部分表情类间相似度高导致误判的问题。

针对上述问题，本章提出一种基于多级特征提取和融合的表情识别网络模型(DFENet)。首先在数据集处理阶段，将人脸表情数据集划分为高低两个相似集，构建由高低相似分支网络组成的多分支架构(RHLNet)。高相似分支通过通道信息增强表情类间区分度，从而降低高相似类别的误分类率，低相似分支在维持低相似集区分度的同时，确保整体网络的均衡性。其次构建了基于多级特征提取和融合的表情识别网络模型(DFENet)，使用 DenseNet 密集连接网络为主干网络，提出一种多级特征提取模块 FEM，采用三个不同卷积核尺度的密集块分别提取低级、中级和高级面部表情特征，设计特征融合模块 FFM，先对局部特征进行独立提取，再对整体图像在输入前和提取全局特征后进行两个阶段的融合，进而将 FEM 与 FFM 结合为整体网络，通过 FFM 引导 FEM 提取并融合多层次面部表情特征。

4.1 DFENet 人脸表情识别网络架构设计

本节主要介绍构建的多级特征提取和融合的人脸表情识别方法的总体框架。

本网络模型的具体识别步骤如图 4-1 所示，识别过程分为数据增强、高低相似集的分类、特征提取和特征融合的识别阶段。首先将人脸表情数据集分为高相似集和低相似集，构建一个由高相似分支和低相似分支组成的多分支网络，高相似分支通过通道信息增强表情类间区分度，从而降低高相似类别的误分类率，低相似分支网络在适度保持低相似集区分度的同时，保证网络的平衡性，解决部分表情类间相似度高导致误判的问题。然后提出了一种具有不同卷积核尺度的特征提取模块，该模块提取多级特征作为整个特征提取网络的输出，提出一种特征融合模块 FFM，先对局部特征进行独立提取，再对整体图像在输入前和提取全局特征后进行两个阶段的融合，以自上而下的方式自适应的融合这些不同级别的特征，并构建新的面部表情特征。

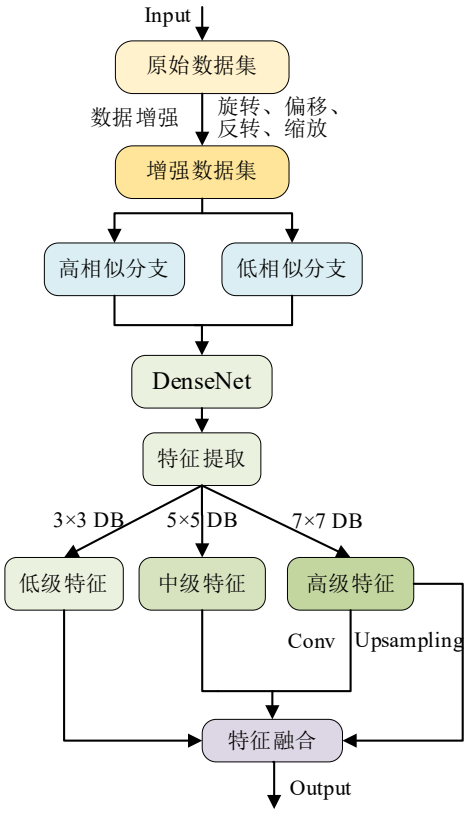


图 4-1 表情识别的流程图

4.1.1 DFENet 整体网络结构

本章提出的多级特征提取和融合网络(DFENet)的整体结构如图 4-2 所示，由图可见所提出的网络结构主要包括特征提取模块(FEM)和特征融合模块(FFM)两个模块。整个人脸表情识别过程可以分为两个部分。首先在特征提取阶段，为了实现特征复用并提高特征提取效率，采用密集连接网络(DenseNet)作为特征提取模块的主干网络，特征提取模块通过三个不同尺度的密集块分别生成面部表情的低级、中级和高级三种的不同层次特征，并各自对应不同层次的密集块。其次在特征融合阶段，将这三个特征映射作为特征融合模块(FFM)的输入，模块使用全局和局部注意力机制对多尺度的特征融合形成特征映射，从而形成用于最终表情分类的特征表示。同时，FFM 通过动态学习不同层次特征来调整融合权重，引导 FEM 更加关注重要特征。FFM 主要实现对不同层次面部表情特征的自适应融合，特别是将高级特征与低级特征之间的全局和局部关系融合，从而显著提高了识别性能，并最终将融合后的特征用于完成面部表情分类任务。

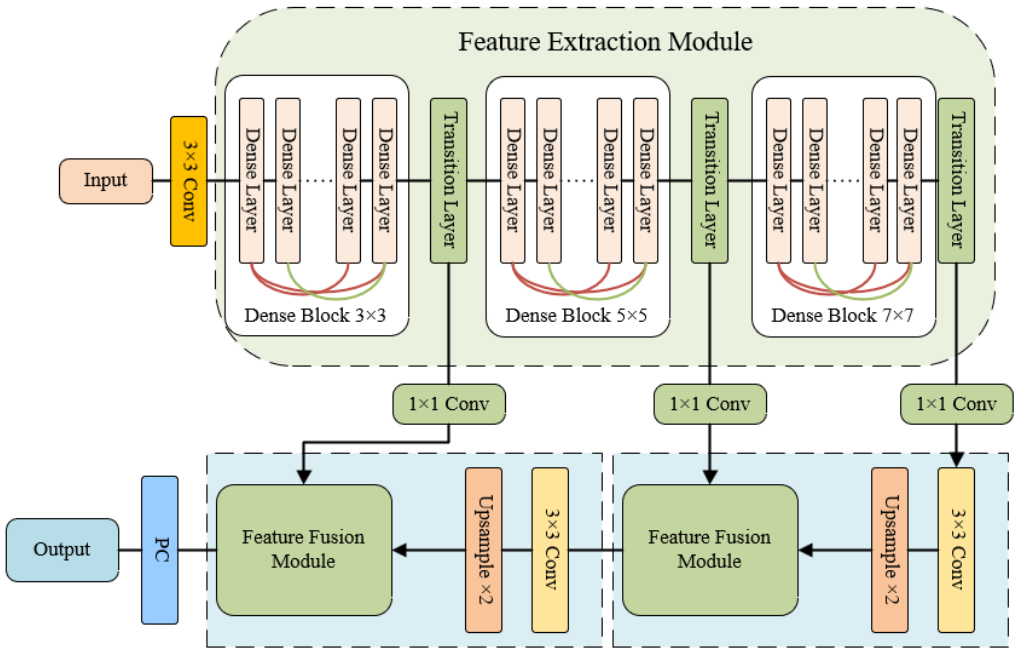


图 4-2 多特征提取与融合模块

(1) 多特征提取模块

FEM 由三个不同尺度的密集块组成，其主要目的在于尽可能多地提取多层次的面部表情特征(低级特征到高级特征)。如图 4-3 所示，特征提取模块包含三个密集块，每个密集块的输出依次作为下一个密集块的输入，这三个密集块分别称为 3×3 密集块、5×5 密集块和 7×7 密集块，不同尺寸的密集块在提取面部表情特征的能力上有所不同，3×3 密集块主要提取低级面部表情特征，5×5 密集块主要提取中级面部表情特征，7×7 密集块主要提取高级面部表情特征。每一个密集块包含不同数量的密集层，每层特征图的尺度均保持一致，并且层与层之间采用密集连接的方式，确保每一层的输出都能被后续层利用，每个密集块的输出是由之前所有层输出的特征通过拼接而成，不仅充分利用了不同核尺度卷积在表情特征提取上的优势，而且能有效整合多层次的特征信息，而不是简单地将最终特征图直接传递到后续网络。

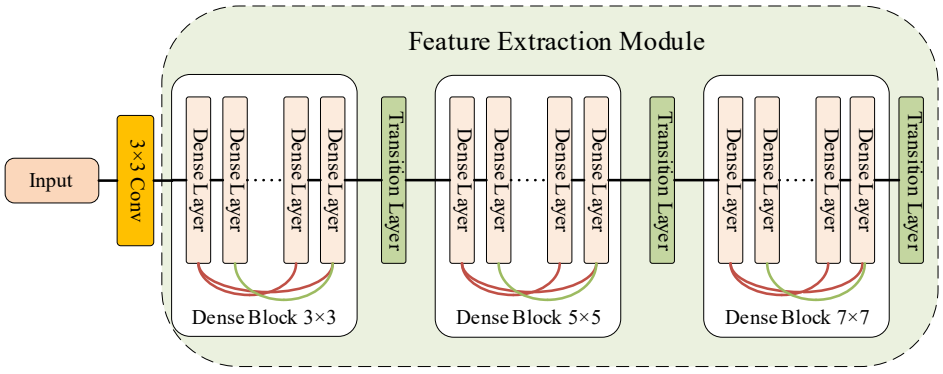


图 4-3 特征提取模块

为了提取面部表情的深层特征，本章引入了密集连接网络(DenseNet)结构，实现特征的充分复用并尽可能提高特征提取效率，如图 4-4 所示，DenseNet 中的密集层由两层卷积组成，其中卷积核参数 K 是每个密集层中卷积核的大小，K 的值分别为 3、5 和 7，取决于不同的密集块。同时每个卷积操作均配合批量归一化(BN)和激活函数(ReLU)操作，以稳定训练过程并增强非线性表达能力。

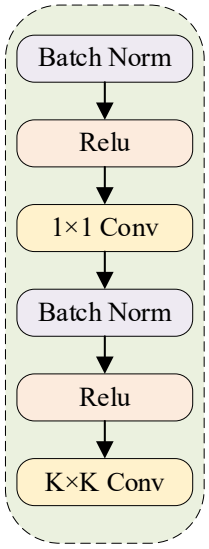


图 4-4 密集层

为了进一步减少整个模型的计算量，过渡层由 1×1 卷积和 2×2 平均池化操作组成，如图 4-5 所示，过渡层用于衔接不同的密集块，并通过卷积与池化运算对特征图的尺寸及通道数进行压缩，以提升计算效率。随着网络深度的增加，密集块输出的特征图通道数会显著增长，可能会导致计算效率下降和训练过程变慢。而过渡层采用了 1×1 卷积操作来降低特征图的通道维度，从而维持模型的复杂度，然后通过采用 2×2 的平均池化层对特征图的空间分辨率进行下采样，在压缩感受野的同时保留了重要的特征信息。该方法有助于降低模型复杂度，同时提升网络的泛化性能。

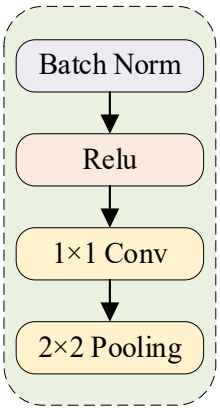


图 4-5 过渡层

(2) 多特征融合模块

本章提出的特征融合模块(FFM)采用独立网络提取人脸局部特征，并在整体图像输入网络之前及全局特征提取完成后分别执行两次特征融合，以提升特征表达能力，所提出的特征融合模块如图 4-6 所示，输入特征首先经过两个 3×3 卷积层，为后续分支的处理提供充分特征基础，然后模块采用双注意力分支分别对全局注意力和局部注意力的特征进行融合，全局注意力分支采用多个 3×3 卷积层堆叠，提取图像的全局特征，并通过平均池化操作聚合全局信息，经激活函数调整特征权重，突出全局关键信息。局部注意力分支也采用多个 3×3 卷积层提取人脸局部特征，接着通过 1×1 卷积调整通道维度，经激活函数强化局部特征的表达。最后将全局和局部注意力分支处理后的特征进行融合，输出全局与局部信息的特征，从而实现两次特征融合，有效的提升了特征表达能力。

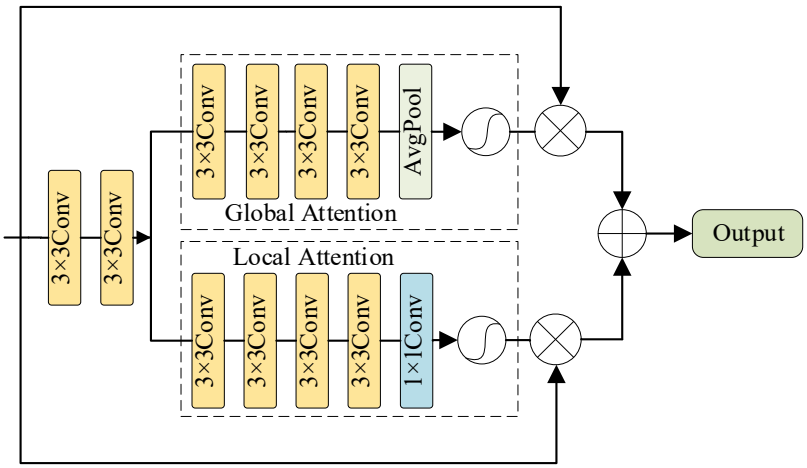


图 4-6 特征融合模块

特征融合模块(FFM)主要由全局注意力和局部注意力组成，这种设计可用于融合不同层次的面部表情特征。首先对输出的较高层次的特征图(H_{Map})进行上采样和卷积操作，然后将处理后的高层次特征与低层次的特征图(L_{Map})融合，生成融合后的特征图，如公式(4-1)所示：

$$F = Conv(Up(H_{Map})) + L_{Map} \tag{4-1}$$

其中，H_{Map}是从模型中较高层提取的特征图，L_{Map}从模型中较低层提取的特征图，$Conv$表示 3×3 卷积核的卷积操作，Up 表示上采样操作，F 表示融合了高层次特征和低层次特征的特征图。

将输出的特征图分别使用全局平均池化和 1×1 卷积生成全局融合和局部融合的注

意力，将全局权重与原始特征逐通道相乘，突出重要通道，局部权重与原始特征逐空间位置相乘，增强关键区域，然后将全局和局部注意力进行融合，输出最终的融合特征。具体实现过程如公式(4-2)至公式(4-4)所示：

$$F_{Global} = \sigma(A_p(Conv(F))) \tag{4-2}$$

$$F_{Local} = \sigma(Conv(F)) \tag{4-3}$$

$$F_{Output} = F_{Global} \times F + F_{Local} \times F \tag{4-4}$$

其中，σ表示 Sigmoid 函数，F_{Global}表示全局融合权重，F_{Local}表示局部融合权重，F_{Output}表示 FFM 的输出。

进一步将 FEM 与 FFM 相结合，形成一个整体网络，如图 4-2 所示。通过 FFM 引导 FEM 提取所需的多层次面部表情特征，有效融合多层次面部表情特征。对包含高级面部表情特征的特征映射进行上采样。采用 FFM 方法对高、低层次的面部表情特征进行融合。将前面三个包含不同层次面部特征信息的特征图通过 FFM 转换成一个包含全局和局部面部表情特征的特征图，为了进一步减少特征图的通道数量，在进入 FFM 之前，每个密集块的输出将通过一个 1×1 的卷积来实现降维的目的。在特征融合过程中，包含更高级面部表情特征的特征图通过双线性操作进行上采样。然后，将更高级和较低级的面部表情特征进行融合。

4.1.2 高低相似分支网络结构

在人脸表情数据集中，由于某些表情类别之间的区分度较低，所以容易引发误判的问题。例如，愤怒与惊讶表情的特征相似性较高，导致愤怒表情常被错误识别为惊讶。所以本章在处理人脸表情数据集时构建了由高低相似两个分支组成的高低相似多分支网络(RHLNet)，整体网络结构如图 4-7 所示。RHLNet 由主干网络 ResNet50 和两个分支网络组成，即低相似分支与高相似分支。在双分支网络架构中，ResNet50 主干网络的第三个卷积块输出特征被用于低相似分支的输入，而第四个卷积块输出特征则作为高相似分支的输入。

ResNet50 是一个深度残差网络，它由多个卷积层和残差块组成，用于高效提取图像特征，本网络结构通过 ResNet50 网络提出每张图像的特征向量，网络通过卷积操作提取图像的局部和全局特征，在 Conv2_x 至 Conv5_x 中，每个阶段都包含多个残差块进行特征提取，具体的网络结构如表 4-1 所示。该网络输入图像的尺寸为 224×224×3，对

输入图像进行普通卷积操作，将输入特征图转换为 64 个通道的特征图，中间层采用了一系列的 bneck 结构，每个 bneck 结构均采用降维再升维的操作，首先通过 1×1 卷积降维，通过减少输入特征的通道数降低后续计算量，然后通过 3×3 卷积提取局部特征，保留空间信息，最后再通过 1×1 卷积升维，将通道数恢复到原始维度，增强特征表达能力。在 ResNet50 的输出层之前，网络会通过平均池化层，将特征图转换成一个 2048 维的特征向量，这个向量表示了图像的高层次信息。

表 4-1 ResNet50 主干网络结构

Input	Opetaror	Channels	Nonlinearities	Stride
224×224×3	Conv2d	64	ReLU	2
112×112×64	MaxPool	64	-	2
56×56×64	bneck,3×3	256	ReLU	1
56×56×256	bneck,3×3	256	ReLU	1
56×56×256	bneck,3×3	256	ReLU	1
56×56×256	bneck,3×3	512	ReLU	2
28×28×512	bneck,3×3	512	ReLU	1
28×28×512	bneck,3×3	512	ReLU	1
28×28×512	bneck,3×3	512	ReLU	1
28×28×512	bneck,3×3	1024	ReLU	2
14×14×1024	bneck,3×3	1024	ReLU	1
14×14×1024	bneck,3×3	1024	ReLU	1
14×14×1024	bneck,3×3	1024	ReLU	1
14×14×1024	bneck,3×3	2048	ReLU	1
7×7×2048	bneck,3×3	2048	ReLU	2
7×7×2048	bneck,3×3	2048	ReLU	1
7×7×2048	GAP	2048	-	-
1×1×2048	FC	-	Softmax	-

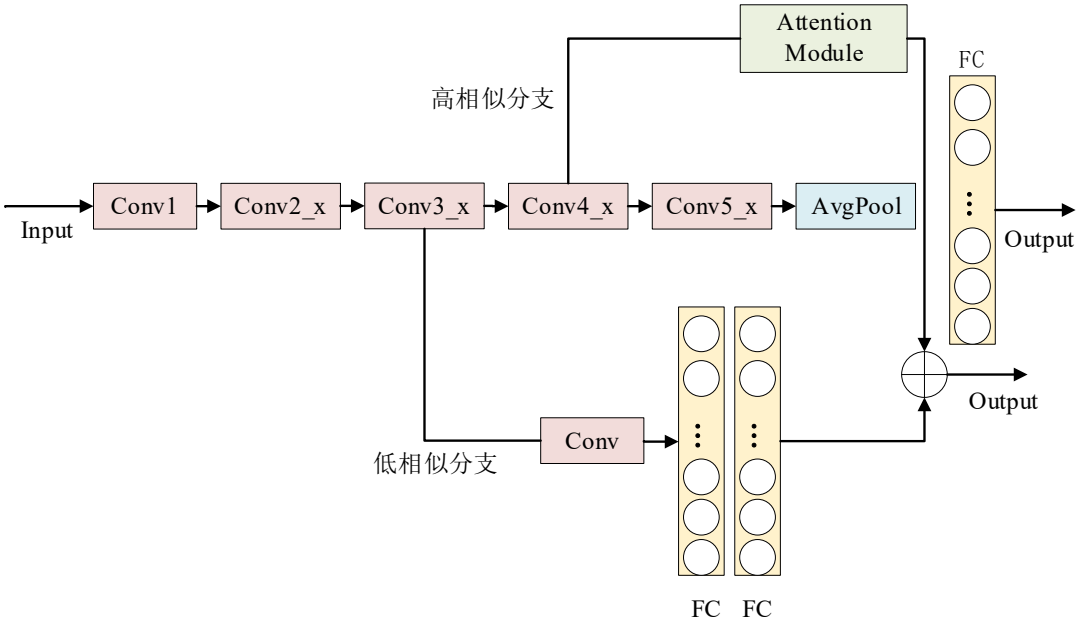


图 4-7 高低相似分支网络模块

(1) 划分高低相似集

相似性度量是度量两个样本在特征空间中相似度的一种方法。通过计算样本之间的相似性度量，可以帮助我们判断它们是否属于相似的类别。在面部表情识别中，为了有效区分不同类别的表情，需要借助相似性度量方法来评估它们之间的相似程度。在信息检索、文本分析以及机器学习等多个领域，余弦相似度(Cosine Similarity)和余弦距离(Cosine Distance)是作为衡量向量间相似性的重要指标。其中，余弦相似度通过计算两个向量在空间中的夹角余弦值来量化其相似程度，值越接近 1 表示向量越一致，相似度越高；反之表示向量相似度越低。具体计算方式如公式 4-5 所示。本设计使用余弦距离来计算每一对样本的相似性，当任意两个样本的余弦距离趋近于 0，意味着它们的相似度越高，具体计算方式如公式 4-6 所示。

$$cos(\theta)=\frac{\vec{a}\cdot\vec{b}}{|\vec{a}||\vec{b}|}=\frac{\sum_{i=1}^{n}a_i b_i}{\sqrt{\sum_{i=1}^{n}a_i^2}\sqrt{\sum_{i=1}^{n}b_i^2}} \tag{4-5}$$

$$d=1-cos(\theta) \tag{4-6}$$

其中，$\vec{a}$和$\vec{b}$分别表示两个 n 维向量，$|\vec{a}|$和$|\vec{b}|$分别表示向量$\vec{a}$和$\vec{b}$的模，a_i和b_i分别表示向量的第 i 个分量，$\cos(\theta)$是两个向量之间的余弦相似度，d 是任意两个向量的余弦距离，$\cos(\theta)\in[-1,1]$，$d\in[0,2]$，$\vec{a}\cdot\vec{b}$表示两个向量的点积即$\sum_{i=1}^{n}a_i b_i$。

图 4-8 展示了通过计算后各表情类别间的余弦距离，其中 D1～D7 分别对应愤怒、

厌恶、恐惧、快乐、悲伤、惊讶与中性这七种基本面部表情类别。通过计算每对表情特征向量之间的余弦距离d，并设定阈值为 0.3，如图 4-8 中红色虚线所示，可将表情划为两个集合，当余弦距离 d≤0.3 时，判定为高相似集，表明这些表情在特征空间中具有较高的相似性，若余弦距离 d≥0.3 时，判定为低相似集，表明这些表情之间存在较大的差异性。因此，表情数据集被划分成高相似集(愤怒、恐惧、惊讶)、低相似集(厌恶、快乐、悲伤、中性)。

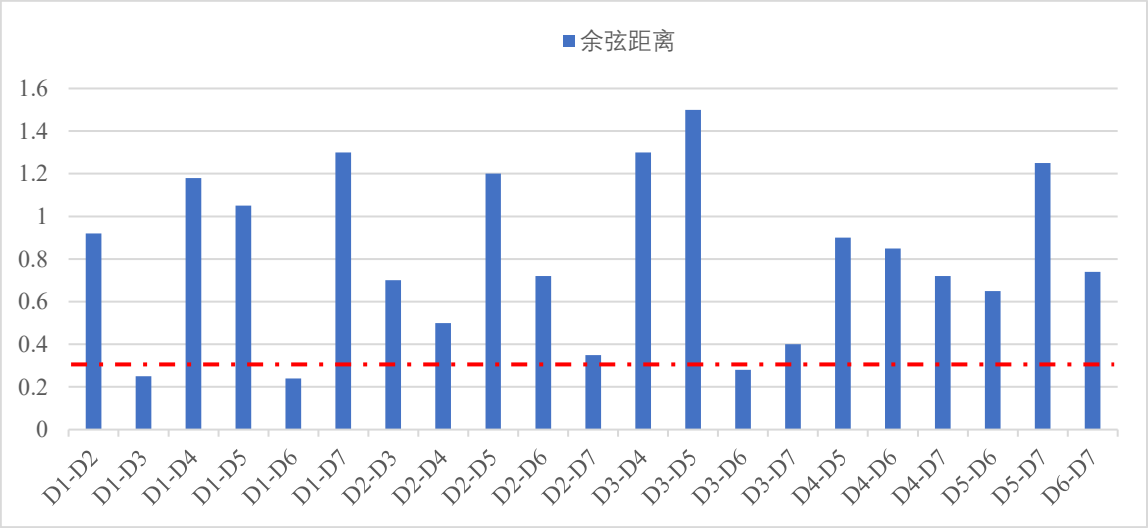


图 4-8 不同表情类别的余弦距离图

(2) 高低相似分支网络

当余弦距离 d≤0.3 时，属于高相似集，当余弦距离 d≥0.3 时，属于低相似集。然后将两个相似集在双分支的网络结构中分别输入到高低相似两个分支网络中进行处理。高相似分支通过深度通道注意力机制增强特征表示能力，重点提取高相似类别间的细微差异，从而显著降低了高相似度表情的误分类率，与此同时低相似分支在保持类别间区分度的基础上，通过浅层网络结构维持了整体模型的平衡性。

由于高相似集中的表情类别在特征空间中分布更为密集，类间差异更为细微，因此需要更深层次的网络结构来捕捉其间的微小变化，而低相似集的表情类别的分类难度相对较低，采用浅层的网络即可实现有效区分。为此，本节构建了高相似分支网络，如图 4-9 所示，由全局平均池化层、两个全连接层以及特征重标定模块构成。其中，全局平均池化层用于提取全局上下文信息，全连接层实现特征变换，而重标定模块则通过通道注意力机制对特征进行动态调整，从而有效放大高相似度表情类别间的差异

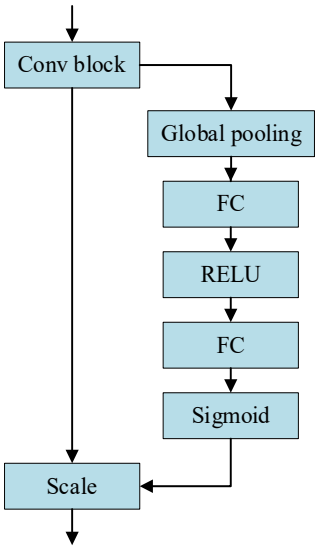


图 4-9 高相似分支网络

高相似分支网络输出的部分表情特征向量通过加权融合机制反馈至主干网络，可能会对模型的预测结果产生影响，如果不加入低相似分支，网络只针对高相似表情进行分支处理，而不为低相似表情分配任何分支或特征通道，那么对于那些被忽略的低相似类别，网络的特征表示将不再被学习或更新，可能在复杂场景中引入误差。为在一定程度上实现对低相似集的适度平衡。本小节提出含有卷积块的低相似分支网络，其结构如图 4-10 所示。

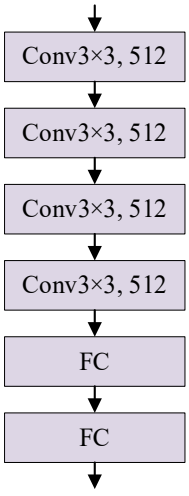


图 4-10 低相似分支网络

4.2 实验数据集及相关配置

4.2.1 实验环境配置及训练细节

本研究所有实验均配置在 GPU 显卡型号为 NVIDIA GeForce RTX 4060 和 CPU 处理

器型号为 Intel Core i7-14650HX@2.20GHz 的计算机上完成。其详细的系统配置环境如表 4-2 所示，完全满足深度学习实验的计算要求。实验环境基于 Anaconda3 构建，采用 Python 3.8.3 为编程语言，并使用 PyCharm 作为集成开发环境，在 Keras 深度学习框架下实现了卷积神经网络的搭建与训练。Keras 提供了便捷的神经网络构建方式，并支持在 CPU 与 GPU 之间切换运行，以提高计算效率并实现并行处理。

表 4-2 系统环境配置

环境配置	型号版本
CPU 处理器	Intel Core i7-14650HX@2.20GHz
内存	32GB
GPU 显卡	NVIDIA GeForce RTX 4060
CUDA 版本	11.1
CUDNN 版本	11.1
Python 版本	3.8.3
Keras 版本	2.2.5

本章基于 Python 的 Keras 库搭建了所提出的网络模型，在训练过程中设定迭代次数(epoch)为 100，设定训练批量大小(batch_size)为 48，为了优化模型的收敛性能，采用 Adam 优化器进行参数更新，其中设定动量系数(momentum)为 0.9，权值衰减率(weight decay)调整为 0.0005，设定初始学习率为 0.001，以优化模型收敛效果。为优化模型训练过程并防止过拟合，采用了一种动态学习率调整策略，当评估指标未表现出上升趋势时，将自动将学习率降低至初始值的 10％。在训练过程中，每个 epoch 都会遍历全部训练样本，当训练损失值低于验证损失时，立即停止训练，否则，模型将继续训练直至达到预设的最大迭代。

4.2.2 实验数据集

(1) 数据集介绍

人脸表情识别已经成为当前计算机视觉技术的热门研究课题，目前已经公开的数据集包括 FER2013(Facial Expression Recognition 2013)、AffectNet、CohnKanade(CK+)、JAFFE(Japanese Female Facial Expression)等。为了更好的训练本章的神经网络模型，本章选用 FER2013 和 CK+人脸表情数据集，本章在上述两个数据集上进行了大量的对比实验，在两个数据集上完成了对人脸表情的识别。与其他方法在预测准确率、预测精度

等方面进行综合比较，下面将详细介绍本章使用的两个公开数据集。

FER2013 是一个常见的面部表情识别标准数据集，最初由 Kaggle 提供，并在 2013 年的 ICML 竞赛上公开。该数据集共包含 35887 张训练图像、4489 张验证图像以及 3589 张测试图像，涵盖愤怒、厌恶、恐惧、快乐、悲伤、惊讶和中性 7 种表情类别。每个样本均为固定尺寸的 48×48 灰度图像，且每张图像都被标注了一个情绪标签，表示该图像所表现的表情，标签信息使得数据集可以用于训练和评估面部表情识别模型，是情感计算和面部表情分析领域的重要资源。图 4-11 展示了 FER2013 数据集中的部分表情示例图像。

图 4-11 FER2013 数据集

CK+数据集是一个广泛用于表情识别的情感数据集，是目前流行度高易用性极强的宏表情数据库，原始数据集 Cohn-Kanade 由 Pittsburgh 大学在 2000 年代初创建，CK+是其扩展版本，包含了更多的样本和更丰富的标注，是面部表情研究中的经典数据集之一。CK+数据集是当前使用最广的公开人脸表情识别测试库，包含 123 名受试者的 593 条表情序列，从自然状态逐步演变至峰值情绪。图 4-12 展示了 CK+数据集中部分表情示例图像。

图 4-12 CK+数据集

(2) 数据增强

由于人脸表情样本较小在训练过程中可能会导致模型的过拟合问题，本小节对 FER2013 和 CK+人脸数据集的训练集样本进行了多种数据增强策略，包括随机旋转、平移、水平翻转以及多尺度缩放等操作。为了评估数据增强对模型性能的影响，分别在数据增强前后对 DFENet 网络模型进行了对比实验，实验结果如表 4-3 所示。其中，Acc1 表示未进行数据增强时模型的识别准确率，Acc2 则代表采用数据增强之后模型的识别准确率。

表 4-3 数据增强前后模型的性能对比

数据集	初始数量	增强后数量	Acc1(%)	Acc2(%)
FER2013	28708	114863	86.58	94.81
CK+	593	2965	84.51	96.75

由表 4-3 可知，数据增强操作有效提升了模型的泛化能力，缓解了过拟合现象，两种人脸表情数据集在经过数据增强处理后，训练集的人脸图像数量增加了近 5 倍。通过数据增强后，DFENet 在这两个数据集的识别准确率都有所提升，尤其对于 CK+数据集而言，由于其原有的训练样本较少，增强处理有效抑制了过拟合，使 DFENet 在该数据集上达到 96.75 %的准确率。实验结果表明，数据增强可以有效地改善人脸表情识别的效果。

4.2.3 高低相似集划分

为了验证高低相似集的划分是否能有效提升识别的精度，将划分高低相似数据集前后进行了对比实验，其一的模型我们使用完整的 DFENet 网络，根据图像间的相似度将数据集分为高相似集和低相似集，并分别训练两个分支。其二的 DFENet-A 的模型未进行高低相似分支网络进行数据集相似的高低划分，而是直接将所有数据输入网络进行训练。高低相似集划分前后的实验结果如表 4-4 所示：

表 4-4 高低相似集划分对比实验

数据集	模型	Acc(%)	Sens(%)	F1(%)
FER2013	DFENet	94.81	94.78	94.84
	DFENet-A	91.53	91.79	91.28
CK+	DFENet	96.75	96.76	96.84
	DFENet-A	92.53	92.79	92.28

由表 4-4 可知，在划分高低相似集后，表情识别的准确率提升了近 3.2%，这一结果表明，高低相似集划分能够帮助模型更好地处理类别间的相似性，减少误分类。对于相似度较高的表情(如“愤怒”和“恐惧”)，高相似分支能够专注于细粒度的表情差异，从而提高模型的辨识能力。而对于差异较大的表情(如“快乐”和“悲伤”)，低相似分支则能够更好地捕捉到这些表情的全局特征，提高分类的准确性。

4.2.4 评价指标

为了通过计算每个阶段的性能参数来全方面评估网络的性能，本章采用准确率(Accuracy)、精确率(Precision)、灵敏度(Sensitivity)和 F1_Score、作为实验结果的评价指标。各评价指标的计算公式定义如式(4-7)至式(4-10)所示：

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4-7}$$

$$Precision = \frac{TP}{TP + FP} \tag{4-8}$$

$$Sensitivity(Recall) = \frac{TP}{TP + FN} \tag{4-9}$$

$$F1\_Score = 2 \times \frac{Precision \times Sensiticity}{Precision + Sensiticity} \tag{4-10}$$

4.3 实验结果及分析

4.3.1 DFENet 识别性能分析

通过采用本章提出的方法，在 FER2013 数据集上获得了 94.81 %的准确率，在 CK+数据集上获得了 96.75 %的准确率，图 4-13 是分别为在 FER2013 和 CK+数据集上每种面部表情的准确率。

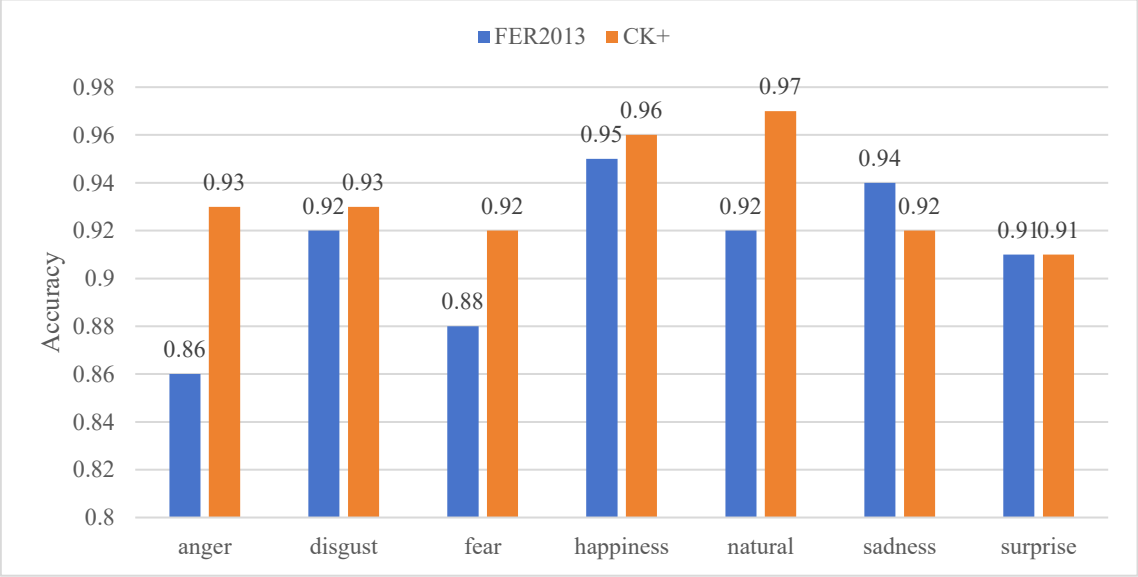
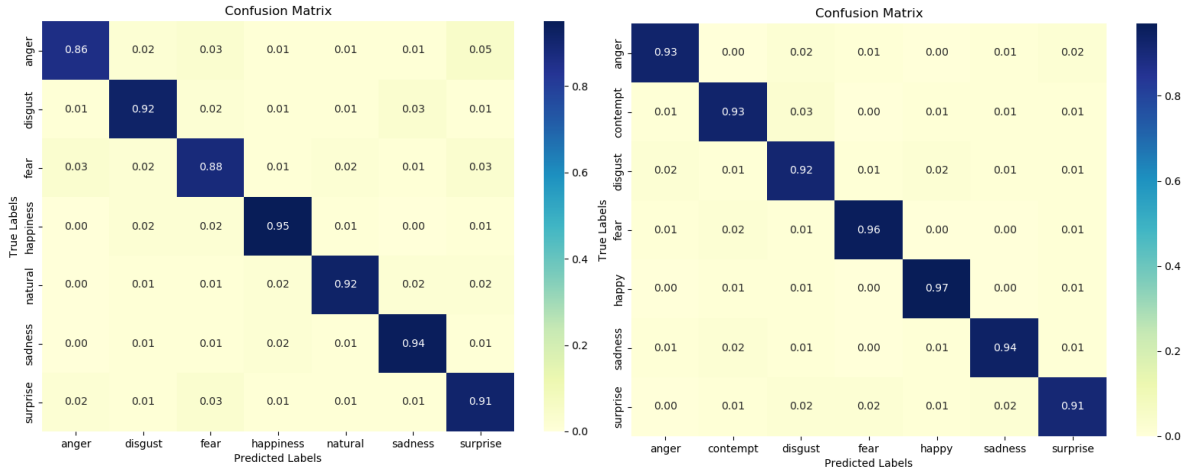


图 4-13 每种面部表情的准确率

由 FER2013 数据集实验结果可知，一些面部表情如“厌恶”、“快乐”、“中性”、“悲

伤”、“惊讶”这五种表情识别准确率较高，准确率均高于 90 %，其余面部表情如“愤怒”和“恐惧”的识别准确率也高于 86 %；由 CK+数据集实验结果可知，所有面部表情的识别准确率均高于 90 %。为了更加细致地比较识别结果的差异和不同表情类别的准确率，本节在 FER2013 和 CK+数据集上绘制了混淆矩阵，如图 4-14 所示。



(a) FER2013 数据集　　(b) CK+ 数据集

图 4-14 两个数据集 7 种面部表情的混淆矩阵

从图中两张混淆矩阵可见，本章所提出的模型在 FER2013 与 CK+数据集上整体准确率均较高，尤其在 CK+数据集上的表现更为优异。模型在大多数表情类别上均表现出较高的准确率，同时，模型对于某些区分度较小的表情如愤怒与恐惧等也取得了较高的准确率，说明改进网络确实增强了特征表达与判别能力，使模型在复杂场景中依旧具备较强的鲁棒性与泛化性。

4.3.2 消融实验结果分析

为了验证各个模块设计的合理性以及 DFENet 网络的识别效果，本节设计了消融实验。在 DenseNet 网络的基础上，依次加入多尺度卷积核和特征融合模块。在保证实验条件相同的状态下，将每个模块与原始网络结合后在 FER2013 和 CK+数据集上进行实验验证。在 DenseNet 网络中引入 FFM 模块得到 DF-Net 网络，在 DenseNet 网络中采用多尺度的卷积核 K 得到 DE-Net 网络。通过比较不同实验组的性能，我们可以有效地分析每个组件对 DFENet 模型的影响。表 4-5 和表 4-6 代表在两个数据集上的消融实验结果。表 4-5 展示了不同卷积核大小以及 FFM 在数据集 FER2013 上的实验性能对比，表 4-6 展示了不同卷积核大小以及 FFM 在数据集 CK+上的实验性能对比。

表 4-5 FER2013 数据集不同卷积核大小以及 FFM 的实验结果

数据集	模型名称	FFM	卷积核大小	Acc(%)	Sens(%)	F1(%)
FER2013	DFENet	√	K=3,5,7	94.81	94.78	94.84
	DE-Net	×	K=3,5,7	90.51	89.65	90.34
	DF-Net-A	√	K=3,3,3	86.21	85.45	85.58
	DF-Net-B	√	K=5,5,5	86.89	86.02	85.63
	DF-Net-C	√	K=7,7,7	85.65	84.87	84.44
	DenseNet	×	K=3,3,3	85.15	84.35	85.01

表 4-6 CK+数据集不同卷积核大小以及 FFM 的实验结果

数据集	模型名称	FFM	卷积核大小	Acc(%)	Sens(%)	F1(%)
CK+	DFENet	√	K=3,5,7	96.75	96.76	96.84
	DE-Net	×	K=3,5,7	93.11	92.36	92.96
	DF-Net-A	√	K=3,3,3	92.61	91.81	92.45
	DF-Net-B	√	K=5,5,5	92.79	92.10	92.64
	DF-Net-C	√	K=7,7,7	92.35	91.55	92.12
	DenseNet	×	K=3,3,3	88.75	89.51	88.65

由表 4-5 和表 4-6 可知，通过对不同模型配置进行比较，分析了卷积核大小和特征融合模块(FFM)对模型性能的影响。实验结果表明，加入 FFM 和使用多尺度卷积核的 DFENet 模型在所有配置中表现最佳，准确率在 FER2013 数据集中达到了 94.81 %，在 CK+数据集中达到了 96.75 %，显示出多尺度卷积核和特征融合对提高模型准确性具有显著作用。相比之下，未使用 FFM 的模型准确率较低，在两个数据集中准确率分别为 90.51 %和 93.11 %，表明 FFM 的缺失导致了性能的下降。关于卷积核的选择，实验体现了卷积核的大小对网络识别性能的影响，较小的卷积核(如 K=3、3、3)通常能获得较好的结果，而较大的卷积核(如 K=7、7、7)则导致了准确率的下降，这可能是因为较大卷积核会过度平滑图像信息，影响细节的捕捉。综上，FFM 与适中的卷积核大小(如 K=3、5、7)的结合能够有效提升模型性能，而卷积核过大或过小都可能对准确率产生负面影响。

4.3.3 经典算法对比实验

本章提出了 DFENet 网络，通过引入多尺度卷积核和特征融合模块的设计，显著提升了面部表情识别的准确性与鲁棒性。为了全面评估 DFENet 对人脸表情识别性能，验证网络的有效性和先进性，将该网络模型在 FER2013 和 CK+两个数据集上与深度学习模型(VTFF、ViT、ResNet50、Inception-ResNet-V1、Inception-ResNet-V2 和 AlexNet)完成对比试验，全面评估其在 FER2013 和 CK+数据集上的表现。表 4-6 展示了本模型及对比模型在数据集 FER2013 上的识别精度及各项性能参数，表 4-7 展示了在数据集 CK+的各项性能指标。如表 4-7 所示，本章提出的 DFENet 网络在 FER2013 数据集中的人脸表情识别精度最高，达到了 94.81 %。相较于检测性能较差的 AlexNet 模型提升了 5.75 %。如表 4-8 所示，本章提出的 DFENet 网络在 CK+数据集中的人脸表情识别精度最高，达到了 96.75 %。相较于检测性能较差的 AlexNet 模型提升了 5.59 %。本章的方法在相较于其他六个主流模型，识别精度处于领先位置，具备较大优势。

表 4-7 FER2013 数据集不同模型识别结果

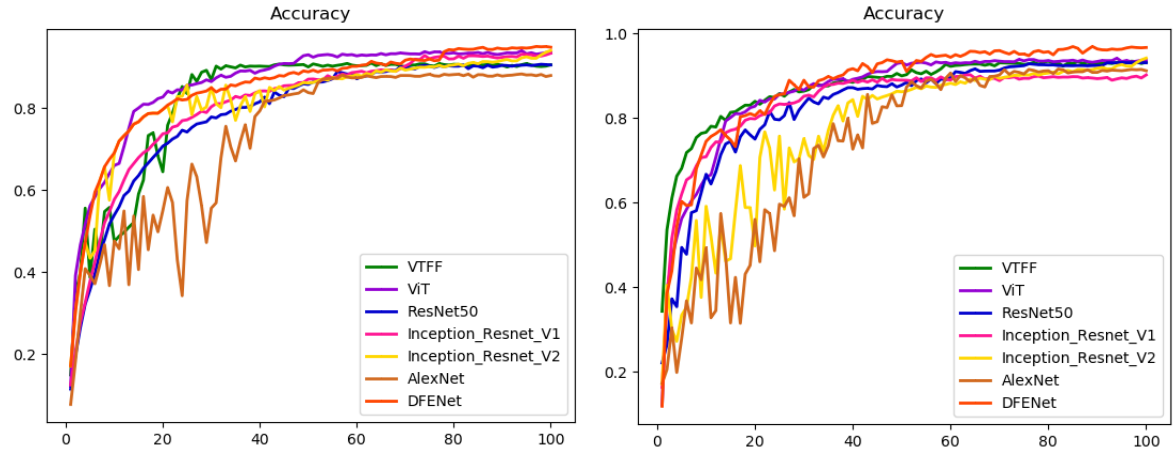
数据集	模型	Acc(%)	Prec(%)	Sens(%)	F1(%)
FER2013	VTFF	90.63	90.78	90.59	90.61
	ViT	93.56	93.61	93.45	93.55
	ResNet50	91.21	91.22	91.15	91.26
	Inception-ResNet-V1	91.87	91.92	91.68	91.85
	Inception-ResNet-V2	92.34	92.35	92.32	92.33
	AlexNet	89.06	89.01	88.94	89.03
	DFENet	94.81	94.84	94.78	94.84

表 4-8 CK+数据集不同模型识别结果

数据集	模型	Acc(%)	Prec(%)	Sens(%)	F1(%)
CK+	VTFF	93.25	93.34	93.22	93.33
	ViT	94.58	94.66	94.51	94.42
	ResNet50	93.39	93.42	93.32	93.84
	Inception-ResNet-V1	94.53	94.68	94.48	94.94
	Inception-ResNet-V2	95.17	95.53	95.15	95.61
	AlexNet	91.16	91.32	91.13	91.67

DFENet	96.75	96.96	96.76	96.84

根据 DFENet 模型和六种对比算法的实验结果，绘制了所有方法在 FER2013 和 CK+数据集上的准确率曲线，图 4-15 为 DFENet 与对比网络在 FER2013 和 CK+两个数据集上的准确率验证曲线图。跟其他网络相比，DFENet 网络对于特征的提取更加细致，验证曲线随着迭代次数的增加稳定上升，经过 100 个 epoch 的训练，DFENet 网络的检测精度高于对比的网络结构。

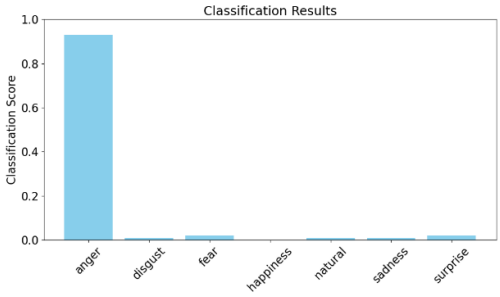


(a) FER2013 数据集　　(b) CK+数据集

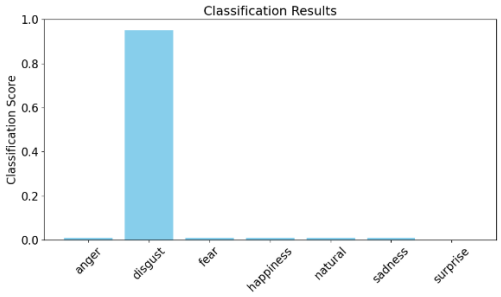
图 4-15 DFENet 与对比网络的准确率验证曲线图

4.3.4 人脸表情识别结果

最后将上述训练的模型用于人脸表情识别，本章通过训练后的模型准确预测了表情图像中的人脸表情识别情况，结果如图 4-16 所示。图(a)至图(g)分别展示了七种不同表情类别的识别结果，可以看到本章设计的模型在人脸表情识别方面取得了较好的效果，并且能够适应不同尺度、姿态和场景下的人脸，并且识别的准确率普遍较高。实验结果证明，本章方法具有良好的泛化能力和较强的鲁棒性。

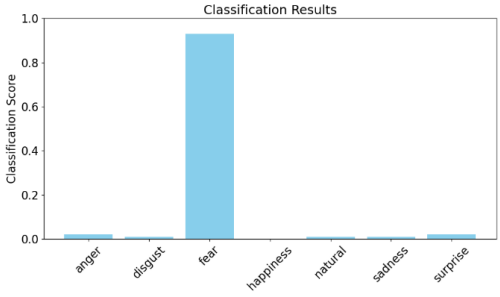


(a) 愤怒表情类别识别结果

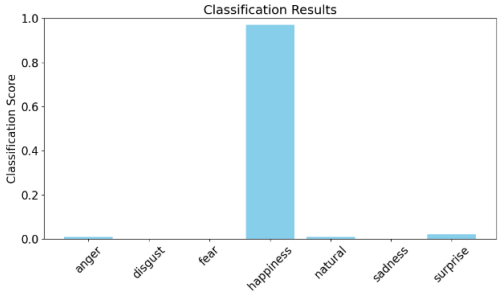


(b) 厌恶表情类别识别结果

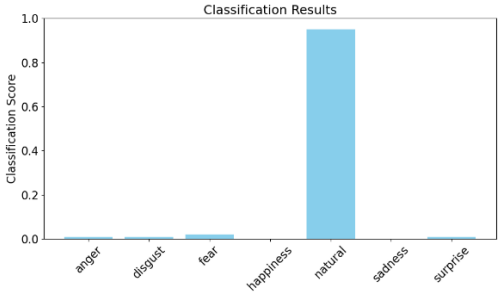
图 4-16 七种人脸表情识别结果图



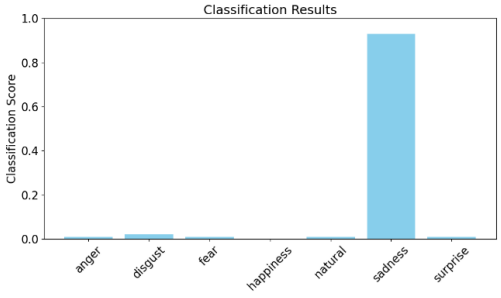
(c) 恐惧表情类别识别结果



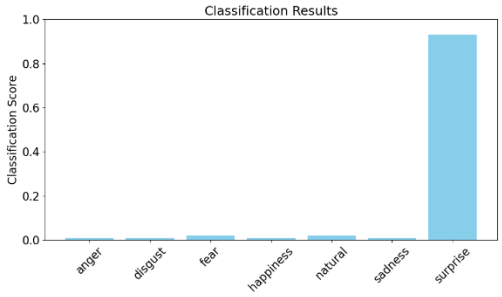
(d) 快乐表情类别识别结果



(e) 中性表情类别识别结果



(f) 悲伤表情类别识别结果



(g) 惊讶表情类别识别结果

续图 4-16 七种人脸表情识别结果图

4.4 本章小结

本章首先在数据集处理阶段，将人脸表情数据集划分为高相似集与低相似集，并构建由高低相似分支网络组成的多分支架构(RHLNet)。其中，高相似分支通过通道信息增强表情类间区分度，从而降低高相似类别的误分类率，而低相似分支在维持低相似集区分度的同时，确保整体网络的均衡性。其次构建了基于多级特征提取和融合的表情识别网络模型(DFENet)，使用 DenseNet 密集连接网络为主干网络，实现特征复用并提高特征提取效率，设计了一种特征提取模块 FEM，采用三个不同卷积核尺度的密集块分别提取低级、中级和高级面部表情特征，然后设计特征融合模块 FFM，先对局部特征进行独立提取，再对整体图像在输入前和提取全局特征后进行两个阶段的融合，进而将 FEM 与 FFM 结合为整体网络，通过 FFM 引导 FEM 提取并融合多层次面部表情特征。最后为检验本章所提模型的有效性，在 FER2013 与 CK+两个数据集上进行了多次消融和对比实验，获得了 94.81 %和 96.75 %的识别准确率，与现有多数表情识别方法相比，DFENet 在识别效果方面更具优势。结果表明，本章所提出的方法提高了表情的识别能力，验证了本章所提 DFENet 网络模型的有效性。

第五章 总结与展望

5.1 总结

随着科技的发展，人们对于从图像中获取清晰、准确信息的需求日益增加，深度学习在人脸图像处理领域得到广泛应用，人脸图像的检测和识别作为计算机视觉领域中的一个重要任务，对于众多人脸相关应用具有关键性的作用和重大的研究价值。为了减少人脸检测的网络参数，提高网络推理速度和增加多尺度表情特征的识别，提高识别精度，本文针对人脸检测和表情识别的方法进行了深入研究，主要工作如下：

(1) 针对目前的人脸检测算法往往存在结构复杂、模型参数和计算量大等问题，提出一种基于深度学习的轻量化人脸检测模型(MYDCFNet)。首先将 YOLOv4 的主干提取网络 CSPDarkNet53 替换为轻量型的 MobileNetV3 网络，同时在主干网络 MobileNetV3 的颈部设计了一种新的激活函数 NHS，通过调整α的大小，使新激活函数在计算效率和非线性特性之间取得平衡，然后设计一种新型深度可分离卷积 NESC，在特征融合过程中将 PANet 网络中的普通卷积替换为 NESC，通过动态调整深度卷积核的结构，更好地适应不同尺度的特征，提高泛化能力；提出一种采用多尺度空洞卷积并行结构的特征增强模块替代 SPP 模块，捕获更广泛的上下文信息；最后，设计一个新的卷积注意力模块 DCAM，通过动态调整卷积核大小融合多尺度特征，在每次上采样过程中，将特征输入卷积注意力模块加强特征提取。最终实验结果证明，本模型大小较原始模型显著减少了约 80 %，计算速度提升了近 67 %，检测精度提升了近 3 %，能够进行实时人脸检测。

(2) 针对目前的表情识别算法存在不能更好提取各层次的特征和细节以及部分表情类间相似度高导致误判的问题，提出一种基于多级特征提取和融合的表情识别网络模型(DFENet)。首先在数据集处理阶段，将人脸表情数据集划分为高相似集与低相似集，并构建由高低相似分支网络组成的多分支架构(RHLNet)。其中，高相似分支通过通道信息增强表情类间区分度，从而降低高相似类别的误分类率；而低相似分支在维持低相似集区分度的同时，确保整体网络的均衡性。其次构建了基于多级特征提取和融合的表情识别网络模型(DFENet)，使用 DenseNet 密集连接网络为主干网络，并设计了一种特征提取模块 FEM，采用三个不同卷积核尺度的密集块分别提取低级、中级和高级面部表情特征；设计特征融合模块 FFM，先对局部特征进行独立提取，再对整体图像在输入前和提取全局特征后进行两个阶段的融合，进而将 FEM 与 FFM 结合为整体网络，通过 FFM

引导 FEM 提取并融合多层次面部表情特征。最终实验结果证明，本方法在 FER2013 和 CK+数据集上的识别精度达到了 94.81 %和 96.75 %，有效提高了表情识别的准确度，能对人脸表情进行有效识别。

5.2 展望

在人机交互、安防监控、情感计算等诸多应用领域中，人脸检测与表情识别技术日益成为关键。人脸检测旨在从复杂场景的图像或视频中准确定位人脸位置并进行基本的特征提取，是后续识别与分析的前提；而表情识别则在此基础上进一步判别人物的情感状态，对于提升人机互动的自然性与智能化水平具有重要意义。虽然本文提出的方法获得了较好的效果，但在未来的研究工作中，还可以从以下几个方面继续研究：

(1) 虽然本文方法应用在人脸检测和表情识别上可以取得较好的效果，但本文针对人脸检测和表情识别的算法均采用的是静态图像作为输入，不能较好的应用于动态数据，例如动态视频、连续帧序列。后续研究可考虑引入循环神经网络，以应对表情动态序列数据的识别需求。

(2) 虽然本文所提出的 DFENet 模型在表情识别任务中取得了较好的识别性能，但仍存在网络结构复杂、模型参数较多、计算资源消耗较大的问题。后续研究可以引入轻量化网络，采用深度可分离卷积、双线性卷积等方式构建轻量化表情识别网络，在保持识别精度的前提下降低模型参数和计算复杂度。

参考文献

[1] Dastres R, Soori M. Advanced image processing systems[J]. International Journal of Imagining and Robotics, 2021, 21(1): 27-44.

[2] 曹家乐, 李亚利, 孙汉卿等. 基于深度学习的视觉目标检测技术综述[J]. 中国图象图形学报, 2022, 27(06): 1697-1722.

[3] Li Y. Research and application of deep learning in image recognition[C]//2022 International Conference on Power, Electronics and Computer Applications (ICPECA). IEEE, 2022: 994-999.

[4] 杨育婷, 李玲玲, 刘旭, 等. 基于多尺度-多方向 Transformer 的图像识别[J]. 计算机学报, 2025, 48(02): 249-265.

[5] Tejasree G, Agilandeeswari L. An extensive review of hyperspectral image classification and prediction: techniques and challenges[J]. Multimedia Tools and Applications, 2024, 83(34): 80941-81038.

[6] Zhao Z Q, Zheng P, Xu S, et al. Object detection with deep learning: A review[J]. IEEE transactions on neural networks and learning systems, 2019, 30(11): 3212-3232.

[7] Vijayakumar A, Vairavasundaram S. Yolo-based object detection models: A review and its applications[J]. Multimedia Tools and Applications, 2024, 83(35): 83535-83574.

[8] Zou Z, Chen K, Shi Z, et al. Object detection in 20 years: A survey[J]. Proceedings of the IEEE, 2023, 111(3): 257-276.

[9] 陈浩楠, 朱映映, 赵骏骐, 等. 基于多模态关系建模的三维形状识别方法[J]. 软件学报, 2024, 35(05): 2208-2219.

[10] 陆凯, 岳康, 胡昊辰, 等. 神经反馈训练中的虚拟现实技术综述[J]. 计算机辅助设计与图形学学报, 2023, 35(08): 1150-1161.

[11] Hsiang E L, Yang Z, Yang Q, et al. AR/VR light engines: perspectives and challenges[J]. Advances in Optics and Photonics, 2022, 14(4): 783-861.

[12] La Marca A, Capuzzo M, Paglia T, et al. Testing for SARS-CoV-2 (COVID-19): a systematic review and clinical guide to molecular and serological in-vitro diagnostic assays[J]. Reproductive biomedicine online, 2020, 41(3): 483-499.

[13] 王万良, 王铁军, 陈嘉诚, 等. 融合多尺度和多头注意力的医疗图像分割方法[J]. 浙

江大学学报(工学版), 2022, 56(09): 1796-1805.

[14] Alalwan N, Abozeid A, ElHabshy A A A, et al. Efficient 3D deep learning model for medical image semantic segmentation[J]. Alexandria Engineering Journal, 2021, 60(1): 1231-1239.

[15] Müller D, Kramer F. MIScnn: a framework for medical image segmentation with convolutional neural networks and deep learning[J]. BMC medical imaging, 2021, 21(1): 1-11.

[16] Akan S, Varlı S. Use of deep learning in soccer videos analysis: survey[J]. Multimedia Systems, 2023, 29(3): 897-915.

[17] 期治博, 杜磊, 霍如, 等. 基于边缘计算的多摄像头视频协同分析方法[J]. 通信学报, 2023, 44(08): 14-26.

[18] Lin W, Liu H, Liu S, et al. Hieve: A large-scale benchmark for human-centric video analysis in complex events[J]. International Journal of Computer Vision, 2023, 131(11): 2994-3018.

[19] Alalwan N, Abozeid A, ElHabshy A A A, et al. Efficient 3D deep learning model for medical image semantic segmentation[J]. Alexandria Engineering Journal, 2021, 60(1): 1231-1239.

[20] Wang W, Wang X, Yang W, et al. Unsupervised face detection in the dark[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(1): 1250-1266.

[21] C Mccullagh P. Face detection by using haar cascade classifier[J]. Wasit Journal of Computer and Mathematics Science, 2023, 2(1): 1-5.

[22] Mohammed M G, Melhum A I. Implementation of HOG feature extraction with tuned parameters for human face detection[J]. International Journal of Machine Learning and Computing, 2020, 10(5): 654-661.

[23] Viola P, Jones M J. Robust Real-Time Face Detection[J]. International Journal of Computer Vision, 2004, 57(2): 137-154.

[24] Ogunseye E O, Adenusi C A, Nwanakwaugwu A C, et al. Predictive analysis of mental health conditions using AdaBoost algorithm[J]. ParadigmPlus, 2022, 3(2): 11-26.

[25] Alsharekh M F. Facial emotion recognition in verbal communication based on deep learning[J]. Sensors, 2022, 22(16): 6105-6119.

[26] Mathias M, Benenson R, Pedersoli M, et al. Face detection without bells and

whistles[C]//European conference on computer vision. Cham: Springer International Publishing, 2014: 720-735.

[27] Sudhakar K, Nithyanandam P. Facial identification of twins based on fusion score method[J]. Journal of Ambient Intelligence and Humanized Computing, 2021, 3(1): 1-12.

[28] Banerjee A, Sarkar S, Nasipuri M, et al. Skin Diseases Detection Using LBP and WLD: An Ensembling Approach[J]. SN Computer Science, 2023, 5(1): 72-86.

[29] 郑向前. 轻量化人脸检测与微表情识别关键技术研究[D]. 甘肃: 兰州大学, 2023.

[30] Zahid S M, Najesh T N, Ameen S R, et al. A Multi Stage Approach for Object and Face Detection using CNN[C]//2023 8th International Conference on Communication and Electronics Systems (ICCES). IEEE, 2023: 798-803.

[31] 曾曦, 辛月兰, 谢琪琦. 基于性别约束的多分支网络人脸表情识别[J]. 计算机工程与应用, 2023, 59(09): 245-254.

[32] 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.

[33] Vimal C, Shirivastava N. Face and face-mask detection system using vgg-16 architecture based on convolutional neural network[J]. International Journal of Computer Applications, 2022, 183(50): 16-21.

[34] Sunnetci K M, Akben S B, Kara M M, et al. Face mask detection using GoogLeNet CNN-based SVM classifiers[J]. Gazi University Journal of Science, 2023, 36(2): 645-658.

[35] Chen B, Ju X, Xiao B, et al. Locally GAN-generated face detection based on an improved Xception[J]. Information Sciences, 2021, 572(1): 16-28.

[36] Yan H, Wang X, Liu Y, et al. A new face detection method based on Faster RCNN[C]//Journal of physics: conference series. IOP Publishing, 2021, 1754(1): 012209.

[37] Peng S, Huang H, Chen W, et al. More trainable inception-ResNet for face recognition[J]. Neurocomputing, 2020, 411(1): 9-19.

[38] 姚鸿勋, 邓伟洪, 刘洪海, 等. 情感计算与理解研究发展概述[J]. 中国图象图形学报, 2022, 27(06): 2008-2035.

[39] Ekman P, Friesen W V. Facial Action Coding System (FACS): a Technique for the Measurement of Facial Actions[J]. Rivista Di Psichiatria, 1978, 47(2): 126-38.

[40] Z M Hafed, M D Levine. Face Recognition Using the Discrete Cosine Transform[J].

International Journal of Computer Vision, 2001, 43(3): 167-188.

[41] Liu C J, Wechsler H. Gabor feature based classification using the enhanced Fisher linear discriminant model for face recognition[J]. IEEE Trans on Image processing, 2002, 11(4): 467-476.

[42] Wen Z, Huang T S. Capturing Subtle Facial Motions in 3D Face Tracking[C]. Proceeding of the 9th IEEE International Conference on Computer Vision-Nice, France: IEEE, 2003: 1343-1350.

[43] Li Y, Zeng J, Shan S, et al. Occlusion aware facial expression recognition using CNN with attention mechanism[J]. IEEE transactions on image processing, 2018, 28(5): 2439-2450.

[44] Krizhevsky A, Sutskever I, Hinton G. ImageNet Classification with Deep Convolutional Neural Networks[J]. Advances in Neural Information Processing Systems. 2012, 25(2): 45-55.

[45] 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.

[46] Cotter S. Low complexity deep learning for mobile face expression recognition[C]//Proceedings of the 3rd International Conference on Vision, Image and Signal Processing. 2019: 1-5.

[47] Mollahosseini A, Chan D, Mahoor M H. Going deeper in facial expression recognition using deep neural networks[C]//Winter Conference on Applications of Computer Vision (WACV). IEEE, 2016: 1-10.

[48] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30: 1-15.

[49] 周明, 段楠, 刘树杰, 等. 神经自然语言处理最新进展——模型、训练和推理[J]. Engineering, 2020, 6(03): 155-188.

[50] 杨韫韬, 聂勇伟, 张青, 等. 基于 RNN 和注意力机制的双向人体姿态补全方法[J]. 计算机辅助设计与图形学学报, 2022, 34(11): 1772-1783.

[51] F. Ma, B. Sun, S. Li, Facial Expression Recognition With Visual Transformers and Attentional Selective Fusion[J]. IEEE Transactions on Affective Computing, 2023, 14(2): 1236-1248.

[52] Zhang L, Hong X, Arandjelović O, et al. Short and long range relation based spatio-

temporal transformer for micro-expression recognition[J]. IEEE Transactions on Affective Computing, 2022, 13(4): 1973-1985.

[53] Asifullah K, Zunaira R, Anabia S, et al. A survey of the vision transformers and their CNN-transformer based variants[J]. Artificial Intelligence Review, 2023, 56(3): 2917-2970.

[54] Vijayakumar A, Vairavasundaram S. Yolo-based object detection models: A review and its applications[J]. Multimedia Tools and Applications, 2024, 83(35): 83535-83574.

[55] Gai R, Chen N, Yuan H. A detection algorithm for cherry fruits based on the improved YOLO-v4 model[J]. Neural computing and applications, 2023, 35(19): 13895-13906.

[56] Xu D, Wu Y. Improved YOLO-V3 with DenseNet for multi-scale remote sensing target detection[J]. Sensors, 2020, 20(15): 4276-4298.

[57] Dewi C, Chen R C, Jiang X, et al. Deep convolutional neural network for enhancing traffic sign recognition developed on YoloV4[J]. Multimedia Tools and Applications, 2022, 81(26): 37821-37845.

[58] Zhang X, Wang W, Zhao Y, et al. An improved YOLOv3 model based on skipping connections and spatial pyramid pooling[J]. Systems Science & Control Engineering, 2021, 9(1): 142-149.

[59] Piao Y, Jiang Y, Zhang M, et al. PANet: Patch-aware network for light field salient object detection[J]. IEEE transactions on cybernetics, 2021, 53(1): 379-391.

[60] Nan Yahui, Ju Jianguo, Hua Qingyi, Zhang Haoming, Wang Bo. A-MobileNet: An approach of facial expression recognition[J]. Alexandria Engineering Journal, 2022, 61(6): 4435-4444.

[61] Hu G, Wang K, Liu L. Underwater acoustic target recognition based on depthwise separable convolution neural networks[J]. Sensors, 2021, 21(4): 1429-1438.

[62] Yin X, Li W, Li Z, et al. Recognition of grape leaf diseases using MobileNetV3 and deep transfer learning[J]. International Journal of Agricultural and Biological Engineering, 2022, 15(3): 184-194.

[63] Prasetyo E, Purbaningtyas R, Adityo R D, et al. Combining MobileNetV1 and Depthwise Separable convolution bottleneck with Expansion for classifying the freshness of fish eyes[J]. Information Processing in Agriculture, 2022, 9(4): 485-496.

[64] Qiming Z, Hongwei Z, Mi Z, et al. A study on expression recognition based on improved

mobilenetV2 network[J]. Scientific Reports, 2024, 14(1): 8121-8121.

[65]徐沁, 梁玉莲, 王冬越, 等. 基于 SE-Res2Net 与多尺度空谱融合注意力机制的高光谱图像分类[J]. 计算机辅助设计与图形学学报, 2021, 33(11): 1726-1734.

[66]Zhao Y, Zhang H, Zhang X, et al. Fire smoke detection based on target-awareness and depthwise convolutions[J]. Multimedia Tools and Applications, 2021, 80(18): 27407-27421.

[67]Song W, Liu Z, Tian Y, et al. Pointwise CNN for 3d object classification on point cloud[J]. Journal of Information Processing Systems, 2021, 17(4): 787-800.

[68]Roy S K, Manna S, Song T, et al. Attention-based adaptive spectral–spatial kernel ResNet for hyperspectral image classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2020, 59(9): 7831-7843.

[69]Xie Y, Tian W, Zhang H, et al. Facial expression recognition through multi-level features extraction and fusion[J]. Soft Computing, 2023, 27(16): 11243-11258.

[70]田富有, 曹玉佩, 赵航, 等. 结合空间注意力机制与多任务学习的耕地地块分割模型[J]. 遥感学报, 2024, 28(11): 2850-2864.

[71]Xu H W, Qin W, Sun Y N, et al. Attention mechanism-based deep learning for heat load prediction in blast furnace ironmaking process[J]. Journal of Intelligent Manufacturing, 2024, 35(3): 1207-1220.

[72]Soydaner D. Attention mechanism in neural networks: where it comes and where it goes[J]. Neural Computing and Applications, 2022, 34(16): 13371-13385.

[73]Fu H, Song G, Wang Y. Improved YOLOv4 marine target detection combined with CBAM[J]. Symmetry, 2021, 13(4): 623-636.

[74]Choi M, Kim H, Han B, et al. Channel attention is all you need for video frame interpolation[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(07): 10663-10671.

[75]Lu E, Hu X. Image super-resolution via channel attention and spatial attention[J]. Applied Intelligence, 2022, 52(2): 2260-2268.

[76]史雨馨, 朱继杰, 凌志刚. 基于特征增强 YOLOv4 的无人机检测算法研究[J]. 电子测量与仪器学报, 2022, 36(07): 16-23.

[77]Liu M, Lin K, Huo W, et al. Feature enhancement modules applied to a feature pyramid

network for object detection[J]. Pattern Analysis and Applications, 2023, 26(2): 617-629.