电 子 科 技 大 学

UNIVERSITY OF ELECTRONIC SCIENCE AND TECHNOLOGY OF CHINA

硕士学位论文

DOCTORAL DISSERTATION

MASTER THESIS

论文题目 轻量级人体检测与行为识别算法的研究

学科专业 计算机科学与技术

学　　号 _____

作者姓名 _____

指导教师 _____

分类号 _____ 密级 _____

UDC [注1] _____

学 位 论 文

轻量级人体检测与行为识别算法的研究

（题名和副题名）

_____

（作者姓名）

指导教师 _____

电子科技大学 成 都

_____

_____

（姓名、职称、单位名称）

申请学位级别 硕士 学科专业 计算机科学与技术

提交论文日期 _____ 论文答辩日期 _____

学位授予单位和日期 电子科技大学

答辩委员会主席 _____

评阅人 _____

注 1：注明《国际十进分类法 UDC》的类号。

Research on Lightweight Human Detection and Behaviour Recognition Algorithm

A Doctoral Dissertation Submitted to

University of Electronic Science and Technology of China

Discipline: Computer Science and Technology

Author:

Supervisor:

School: School of Computer Science & Engineering

独创性声明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。据我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得电子科技大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

作者签名：_____ 日期：　　年　　月　　日

论文使用授权

本学位论文作者完全了解电子科技大学有关保留、使用学位论文的规定，有权保留并向国家有关部门或机构送交论文的复印件和磁盘，允许论文被查阅和借阅。本人授权电子科技大学可以将学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。

（保密的学位论文在解密后应遵守此规定）

作者签名：_____ 导师签名：_____

日期：　　年　　月　　日

摘　要

人体检测是当前机器学习领域研究的热点，该技术在虚拟现实、自动驾驶等领域有非常重要的意义。随着深度学习的快速发展，人体检测技术已经取得了不错的进展。但当前人体检测算法在复杂场景下仍然面临目标多尺寸、遮挡等问题，检测精度和速度往往顾此失彼，使得该技术在现实应用时受到一定约束。此外，智能安防、人机交互等领域迫切需要对人体提取更高级的语义信息，即人体行为识别。人体行为与人体姿态联系紧密，尽管基于人体姿态的行为识别已经有不少研究成果被提出，但现有的算法总将两者分开处理，算法的复杂性较高。针对上述问题，本文在人体检测与行为识别领域进行更深入的研究。具体工作主要包括以下几点：

1、针对当前轻量级网络的通用性不强、鲁棒性不高的问题，本文基于 ResNets 设计了一种轻量级网络 LimitNet，并使用网络 op 融合的方法在推理阶段对模型加速。该网络在三种测试集下的性能表现比当前常用的轻量级网络更突出，并且将 LimitNet 应用到人体检测与行为识别领域也能表现出突出的性能。

2、在人体检测技术的研究中，针对现有的算法精度和速度失衡的问题，在 CenterNet 的基础上提出 Refine_CenterNet，主要在上采样模块和预测模块上进行改进。在上采样模块中设计了一种基于注意力机制的特征融合模块来对浅层特征和深层特征进行融合，改进后的算法在小尺寸目标上的 AP 提高了 2%。在预测模块，设计了一种多阶段预测模型——先用输入图像尺寸下采样 8 倍的 Feature map 作为预测分支对网络优化，然后将下采样 4 倍的 Feature map 作为分支预测对网络进行微调。最终算法在测试数据集上的 AP 有显著提升，且推理延时没有明显增加。

3、在人体行为识别算法的研究中，本文设计了一种多任务网络模型，可以同时进行人体行为识别与姿态估计两种任务。其中在姿态估计中，提出 S-Soft-Argmax 方法直接从 heatmap 中回归对应的人体坐标，使得模型可以端到端的训练，提升姿态估计的准确率。在行为识别模块，基于姿态与外观特征，设计一种自适应特征加权的方法对两者进行融合，显著提升了模型的准确率。该模型在 Penn action 数据集中能够达到了 96%的识别精度，且推理时延只有 60ms。

本文在人体检测与行为识别的深入研究中，对算法的各个模块都进行了大量的性能对比实验。整体的检测算法与行为识别算法不仅能够达到高准确率，而且在 1080TI 下总体时延只有 67ms，在现有算法中很有竞争力。

关键词：人体检测，行为识别，轻量级网络，网络 op 融合，特征融合

ABSTRACT

Human detection is a hot spot in the study of machine learning, which is of great significance in fields such as virtual reality and autonomous driving. Human detection technology has made good progress with the rapid development of deep learning. However, in the real complex scenarios, current human detection algorithms are still difficult to deal with the problems of multiple target sizes and occlusions. There is usually the trade off between the detection precision and speed, which makes this technology subject to some constraints in real-world applications. In addition, intelligent security, human-computer interaction and other fields urgently need a technology to extract higher-level semantic information from the human body, i.e., human behavior recognition. Human body behavior is closely related to human pose estimation. Although many research results have been proposed for behavior recognition based on human pose estimation, the existing algorithms always deal with the two separately, which increases the complexity of the algorithm.this thesis conducts more in-depth research in the human detection and behavior recognition. The specific work of this thesis is summarized as following:

1. Aiming at the problem of low versatility and low robustness of the current lightweight network, this thesis designs a lightweight network named LimitNet based on ResNets, and uses the network op fusion method to accelerate the model in the inference stage. The overall performance of the network under the constructed complex test set is much better than the commonly used lightweight networks, and LimitNet also shows outstanding performance for the human detection and behavior recognition.

2. In the investigation of human detection, Refine-CenterNet is proposed on the basis of the CenterNet detection algorithm to solve the problem of the accuracy and speed imbalance of the existing detection algorithm, which is mainly improved on the up-sampling module and training method. In the up-sampling module, a feature fusion module based on the attention mechanism is designed to fuse the shallow features and deep features. The improved algorithm improves the AP on small-scale targets by 2%. In the training, a multi-stage training method is proposed—first use down-sampling 8 times the central point heatmap to optimize the training of the network, and then down-sampling 4 times the heatmap to fine-tune the model. The final detection algorithm has a significant

improvement in AP on the test data set.

3. In the investigation of human behavior recognition, a multi-task network model is designed, which can perform both human behavior recognition and pose estimation tasks at the same time. In the pose estimation branch, the S-Soft-Argmax method is proposed to directly regress the corresponding human coordinates from the heatmap, so that the model can be trained end-to-end. In the Feature fusion module, an adaptive Feature weighting method is designed to fuse the action Features of the two branches, which significantly improves the accuracy of the model. The model can achieve a recognition precision of 96% in the Penn action data set, and the reasoning delay is only 60ms.

In the in-depth study of human detection and behavior recognition, this thesis has conducted a lot of performance comparison experiments on each module of the algorithm. The overall detection algorithm and behavior recognition algorithm can not only achieve high accuracy, but the overall delay is only 67ms under 1080TI, which is very competitive in the current application algorithms.

Keywords: human detection，behavior recognition，lightweight network, feature fusion module，network op fusion

目 录

第一章 绪论

1.1 研究背景及意义

随着大数据和硬件计算能力的快速发展，人工智能技术迅速发展起来，深度学习在计算机视觉领域开始应用广泛。研究学者们都将卷积神经网络作为必不可少的模块应用到各个领域的算法研究中，包括目标检测、行为识别、目标分割、跟踪、语音识别等。目前人工智能领域的相关产品大量被应用到我们的实际生活中，如一些人脸识别软件、娱乐软件、虚拟现实等。其中人体检测技术作为当前最受学者们关注的计算机视觉方向之一，近几年已经取得了很好的进展。人体检测的定义是在图像或者视频中回归人体的位置和大小信息。随着智能监控、自动驾驶等领域的快速发展，人体检测技术必须更进一步、突破瓶颈快速发展。该技术不仅可以提升我们的生活效率，还可以在国家安全领域发挥重要作用。下面对人体检测在视频监控等领域阐述人体检测的意义。

在视频监控领域，利用人体检测技术可以统计教室内学生的数量，监控是否出现学生旷课现象，甚至通过人体检测技术实现监控教室内学生的动向。通过人体检测技术，我们可以实现人的智能化管理，在商场、地铁口、小区门口等有监控设施的公共场所，通过该技术可以统计街道上的人流流量，进而高效地进行疏导，避免发生人员聚集现象。在必要时甚至可以有效掌握人员动向，实现精准隔离，这在新冠疫情期间能发挥重要作用。在自动/辅助驾驶方向，通过实时检测车辆周围的行人情况从而引发车内报警功能启动制动，可以极大地避免交通事故的发生。在智能机器人方面，该技术同样发挥着重要的作用，机器人需要通过人体检测技术检测人体，从而与人进行交流或者避开人等。

人工智能的快速发展和视频监控的智能化，提高了从图像或者监控视频中提取有效信息的效率，使得其他计算机领域较难的技术也得以发展起来，这其中就包括人体行为识别技术。人体行为识别是在人体检测基础上再进行的一项行为语义的任务，通常人体检测是人体行为识别的前提或者先验信息，两者不可分割。行为识别的主要流程如图 1-1 所示。

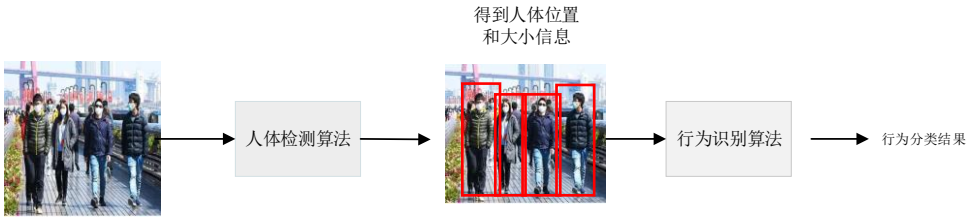


图 1-1 行为识别流程

行为识别任务更加抽象，其广泛应用于视频监控、虚拟现实和人机交互等领域。如通过该技术可以识别教室内学生的行为，从而观察学生是否有认真上课；在自动驾驶领域，可以时刻监督驾驶员的驾驶行为状态，判断是否有醉驾、疲劳驾驶等。在运动领域，人体行为识别可以作为一种辅助手段，用于分析运动员的动作，继而通过针对性指导纠正运动员的错误动作，减少运动员的受伤机率，从而能够极大提升运动员的竞技水平。同样在医疗领域，通过该技术医生可以时刻监控病人的行为，分析病人的健康状况、监控病人的安全，甚至可以辅助康复病人，这将会给医疗领域带来革命性的改变，极大的提升医疗效率。这些都使得此项研究具有很高的实践价值和研究意义。

尽管目前人体行为识别的发展很迅速，研究学者们在该领域已经取得了重要进展，但人体运动的多变性以及高复杂性使得人体行为识别的准确度较低、实时性较差，并不能满足行业的应用需求。目前来说，该任务仍然很具有挑战性。

综上所述，人体检测技术与行为识别技术不仅能够给人们的日常生活带来更加便捷舒适的体验，同时在国家安全、地区管理及人类智能化管理等方面都发挥着重要作用，是一项极具有研究意义和挑战性的研究工作。

1.2 国内外研究历史及现状

人体检测技术与动作识别技术是智能社会的重要辅助手段，在老人看护、人机交互、智能安防、视频搜索和敏感场所异常监控等领域都有着重要应用价值。人体检测作为目标检测的子任务，目前已经取得了阶段性的成果，但设计一种鲁棒性强的人体检测算法仍是当前学者们研究的难点。人体行为识别作为目标检测任务的下一阶段任务，由于图像或视频中存在着人体、物体、事件和场景等更复杂、更抽象的行为信息，设计一种高效的算法来快速的对图像或视频中多种行为信息进行分析和识别是当前人体行为识别研究中的热点和难点。

1.2.1 人体检测算法研究历史及现状

人体检测技术是当前计算机视觉任务的研究热点，目前已经取得了很大的成

果。研究人体检测的方法主要有两类：基于传统手工特征的方法和基于深度学习的方法。一般人体检测算法的主要包括以下三个部分有如下：候选窗口（候选目标框）的生成、特征提取和人体分类。其中传统手工特征的方法在早期的发展中发挥着重要作用，其特点是设计人工特征方法来提取候选区域的特征，然后对特征进行分类。其中常用的人工特征提取方法有 HOG-LBP[1]特征、Haar[2]特征、HOG[3]特征、SIFT[4]特征等等，但传统手工设计特征的方法存在灵活性低，扩展性弱等缺点，在解决人体光照、姿势变化、遮挡等视觉问题时，鲁棒性太差，只能在极少特殊的场景下才有效。

随着近十年来深度学习的快速发展，人体检测技术取得了重大突破。2012 在 ImageNet 图像分类比赛，深度神经网络 AlexNet[5]取得了的巨大成功，大幅度提升了图像分类的准确性，引发了广大学者对深度学习的关注和探讨。深度卷积神经网络的提出在计算机视觉领域是划时代的，它不仅可以提取图像中抽象的局部特征和全局特征，而且整体结构简单，可以进行端到端的训练和测试。从 2014 年，随着神经网络发展成熟，提出了 RCNN 检测方法，该检测方法使用了神经网络提取特征，在数据集上的表现远超基于人工特征的方法。深度卷积神经网络在目标检测领域广泛应用和发展，随后大量的学者投入到该领域的研究中。

图 1-2 展示了自深度学习成为主要的检测任务的研究方法后，在目标检测领域出现的一些优秀具有重要意义的方法。

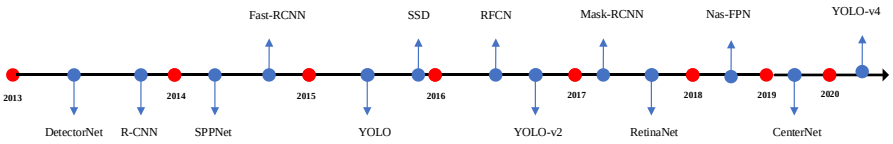


图 1-2 目标检测发展历史

2013 年 Ross Girshick[6] 基于神经网络的检测算法提出了 RCNN 检测方法，吸引了大量的学者的关注，但该方法存在只能固定输入分辨率、计算量大、端到端不可训练等缺陷。2014 年，在 RCNN 的基础上，Kaiming He[7]提出了一种特征池化的模块 SPPNet，该方法检测方法不可以适应任意大小的输入分辨率，同年在 RCNN 的基础上，提出了 SPPNet，使得检测网络能够支持任意大小的图片输入，并大量简化了网络模型的计算量，目标检测算法更进一步。同年，Ross Girshick[8]提出了速度更快、性能更好的 Fast-RCNN[8]，但该方法的检测速度与对难样本的检测还是有很大的不足。直到 2015，Faster-RCNN[9]、YOLO[10]以及 SSD[11]相继发表，该检测算法大幅度的提升了准确率，其中 YOLO 和 SSD 更是抛弃了候选区域生成的复杂工作，直接通过特征点回归的方式生成目标的位置和大小信息，一阶段的检

测框架正式受到广大学者们的追捧和研究。在 2015 年到 2017 年期间，检测方法主要都是基于 YOLO 和 Fast-RCNN 改进的版本，比如通过添加特征多尺度模块 FPN 进而提升多尺寸目标的精度和召回率的 RetinaNet 检测方法，比如 MTCNN 人脸检测算法通过多尺度输入、多阶段训练的方式提升算法的性能。随后，Joseph Redmon[12][13]等人的系列文章，不断将 YOLO 网络进行改进，继 YOLO-v3 后又相继提出 YOLO-v4、YOLO-v5 等。后续方法的主要思路基于 YOLO 等架构提出增强模块和训练方法提升准确性，如特征融合、多尺度输入、多阶段训练等方式。

从上述检测方法的发展来看，基于是否进行候选区域的提取主要将检测算法分为两大类：一阶段检测算法和两阶段检测算法。例如最开始出现的 RCNN 以及改进的系列 RCNN 方法，如 Fast-RCNN 等都是二阶段检测算法，而后设计的 SSD[11]、YOLO 系列和 RetinaNet 等算法都属于一阶段检测算法。两者之间的区别在于是否会生成候选框区域。二阶段检测算法优点通常是精确度较高，缺点是检测速度较慢，一阶段算法则相反。基于是否 Anchor 的先验信息，通常也把检测算法分为基于 Anchor 和 Anchor-Free 的检测算法。2018 年开始，基于 Anchor-Free 的方法开始流行起来，2019 年陆续提出一系列基于 Anchor free 的方法将目标检测领域有提升一个等级。如 CenterNet[14]等，该类算法简化了检测架构，不仅具有高准确性的特点，而且前向推理速度通常比基于 Anchor 的方法更快。

1.2.2 人体行为识别算法研究历史及现状

人体行为识别的定义简单来说就是对人的某些动作（如吃饭、打球、打电话等）进行分类，由于该分类任务更加抽象，所以目前还未取得比较瞩目的研究成果，但该方向依然是当前学者们研究热点。在应用方面，人体行为识别在视频监控、医疗辅助、机器人等领域发挥着重要的作用。

通常人体检测技术是人体行为识别的先验，它可以从采集的视频或者图像中来定位人体的位置和大小信息，从而生成只包含人体的视觉信息，然后再进行人体行为的分析。最开始，人体行为识别算法都是基于传感器的方法，即研究者们首先通过传感器来获取人体的空间信息，其次对其进行数据分析，最后得到人体的动作行为分类。但由于 20 世纪受限于硬件设备的落后且方法存在比较复杂、不灵活等缺点，使得无法获取到足够关键的人体信息，导致人体行为识别在当时未能得到较快的发展。进入 21 世纪后，随着大数据的发展、硬件资源的计算能力的提升，神经网络逐渐开始流行起来。人体行为识别从基于单帧的方法进一步提升到基于视频或者图像序列的行为识别。由于基于单帧的行为识别中图像的有效信息较少，且人体动作往往是持续性变化的，导致准确率始终达不到应用要求。但对于这方向的

研究依然吸引了有大量的研究学者们关注。而基于视频的行为识别方法在当前取得了一定的研究进展，由于该方法输入的是视频，动作信息更多，在准确率上的提升显著，例如当前基于 3D 卷积的方法、双流网络的方法、基于骨架提取方法等都取得了不错的效果。但目前研究方法的瓶颈在于基于视频的行为识别方法计算量大，且在复杂环境下行为容易受到很大干扰导致人体姿态发生巨大变化，通常会造成模型的误判。虽然行为识别的精确度取得了很大的提升，但计算代价较大、训练难度高，同时还有比较复杂的数据集的处理阶段。但由于基于视频的行为识别方法灵活度更高、扩展性更强、准确度更高的优点，获得了研究学者们的广泛研究。

基于深度学习的人体行为方法结构简单，只需要通过深度神经网络模型自动学习视频的行为表征，从而完成自动分类，该方法可以进行端到端的训练，相较于传统方法效率和准确率都有大幅度的提升。一般而言，人类行为识别有着多种模态，例如外观、光流、深度和身体骨骼等，这些模态相辅相成。基于这些特征，研究学者们在行为识别领域提出了不同的方向解法。当前，基于深度学习的行为识别方法主要有基于 3D 卷积网络[15][16][17]、双流网络[18][19]和骨架提取、LSTM[20][21] 等。

尽管目前人体行为识别的发展很迅速，研究学者们在该领域已经取得了重要进展，但人体运动的多变性以及高复杂性使得人体行为识别的准确度较低、实时性较差，并不能满足行业的应用需求。主要面临的难点在于空间复杂性和时间复杂性两方面。

空间复杂性是指人体在不同环境以及不同时刻可能呈现的姿态、视角和人体背景等状态千变万化。对于任何任务而言，输入太多的变量对于网络要求会更加严格。即使在相同的环境下，人体的行为动作也是千变万化、不可预知，设计一种神经网络结构能够对其具有强鲁棒性显然不太现实。

时间差异性也是该方向研究学者们研究的难点，其定义是人体在某个时间点的状态是不可预知的，且该状态会持续多长时间也是未知的，不是都一致的。依照人们的常识，人体发生动作时的一段时间内，可能存在空白期，即没有该行为动作发生，甚至会干扰到该行为的识别，这会严重影响精确性。所以在算法模型设计时需要考虑这些因素，从而提升模型的鲁棒性，但想要达到这些目标非常困难。

尽管目前基于深度学习的方法已经取得了阶段性的进展，但在大部分场景下行为识别的准确性和实时性还不能达到实际的应用要求，该领域研究道路依然未取得突破。目前人体行为识别的研究分为静态图像的人体识别和视频序列的人体行为识别。最初在静态图像上的行为识别学者都将其当作一种分类任务进行，但人体的结构很特殊，结合人体姿态估计技术，主要出现后面陆续出现对人体骨架的研究，从这开始就出现人体姿态估计的研究热点。利用人体关键点的信息，学者们使

用图卷积等方法对该图像的关节点进行分析，能够大幅度提升研究精度。随着深度学习和大数据的快速发展，视频人体行为识别也成为当前的热点。但这两种任务原理很相似，只是视频序列更关注图像序列中的时空变化。在视频研究领域，一个重要的难点在于时序，由于视频是具有时序的单帧图像组成的，如何去表示或者提取这些时序特征也是当前研究的难点和热点。很多通过单帧的图像无法判断的行为却可以通过视频准确识别。通常需要 3D 的卷积或者将视频序列进行处理后输入，算法的复杂度提升一个数量级。随着科技的发展，行为识别方法必将成为许多领域不可或缺的技术，例如视频监控、自动驾驶，甚至国家安全方面都有较高的应用前景。但目前人体行为的多样性、视频的视角变化和非刚性运动的复杂性等因素，导致人体行为识别依然存在巨大挑战。

1.3 论文研究内容及创新

本文的主要研究目的是设计一种轻量级的人体检测与行为识别算法，并保持算法的准确率和推理速度的良好平衡，最终让其能够应用到现实场景中。结合当前技术的研究现状以及未来发展趋势，本文主要在以下几个方面进行了深入研究。

针对人体检测与行为识别的速度和精度不可兼得的问题，本文基于当前经典网络 ResNets 设计了一种高准确率低时延的轻量级网络 LimitNet，在输入分辨率为 112×96时，单帧推理时延只有 4.1ms，且在几种测试集上的性能比当前常用的网络更突出。

在人体检测技术的研究中，针对当前检测多尺寸目标的网络推理延迟大、计算量大的问题，对 CenterNet 的检测算法进行改进，提出 Refine_CenterNet 检测算法，主要在上采样模块以及预测模块改进。最后相较于 CenterNet，该算法的推理速度更快，且 AP 提升约 2%，尤其对小目标的检测效果显著提升。

在人体行为识别技术的研究中，基于人体的姿态和外观的特性，设计了一种姿态与行为识别多任务网络模型。其中在姿态模块，提出 S-Soft-Argmax 方法直接从 heatmap 中回归对应的人体坐标，使得模型可以端到端的训练。在行为识别模块，提出一种基于注意机制的特征加权融合方法对人体行为特征进行融合，最后模型的准确率显著提升，且推理速度在现有的算法中更具优势。

本文在人体检测与行为识别任务中进行了大量的对比实验，最终在 1080TI 下整体延时只有 67ms，且在现实复杂场景下人体检测算法能够达到 43.9%的 AP。在行为识别中，在 Penn action 数据集下能够达到 96%的准确率。本文设计的模型的整体性能与现有的方法相比更有优势。

1.4 本论文的结构安排

本文将研究内容分为六章，其中各章的主要内容简介如下所述：

第一章主要介绍了人体检测技术与人体行为识别技术的研究背景和意义，并对其发展历史以及研究现状进行了概述，其中重点阐述了人体检测和行为识别领域研究的技术难点。

第二章首先介绍了传统人体检测与行为识别算法的主要研究方法与相关基础理论，其次分析了传统方法与基于深度学习方法的优缺点，最后对当前轻量级网络做了简要的介绍。

第三章主要内容为轻量级特征提取网络的设计，详细介绍了数据集的构建方法以及轻量级网络的实验结果对比。

第四章主要对人体检测算法做深入的研究，包括人体检测算法的难点分析、算法的详细设计，最后对该算法做了对比实验分析。

第五章对人体行为识别算法展开深入研究，针对现有方法的难点，设计一种多任务的网络模型。并对设计的模型做了大量的对比实验验证算法的性能。

第六章对本文主要研究内容进行了总结，并分析不足之处以及对未来工作展望。

第二章 理论基础及相关技术

近年来，随着人工智能技术的快速发展以及科技的不断进步，人工智能相关产业也越来越引起世界的关注。人体检测技术和行为识别技术作为计算机视觉中研究的热点，受到广大社会人士的关注，在智能监控领域发挥着重要的作用，尤其是在公共安全方面。本章从人体检测和行为识别的传统方法到深度学习方法，以及本文涉及的其他相关知识进行详细的介绍。

2.1 传统机器学习

在 20 世纪，计算机技术刚开始发展起来，深度学习方法由于硬件计算能力不足，还没有成为主流的特征提取方法，传统机器学习方法在当时广受研究学者们的关注。其中在人体检测与行为识别等领域，只能通过设计手工特征的方式实现轮廓等信息的提取，当时比较主流的特征提取方法提取方法主要有以下四种。

LBP 特征常用人脸分析等领域。它的主要流程如图 2-1 所示：

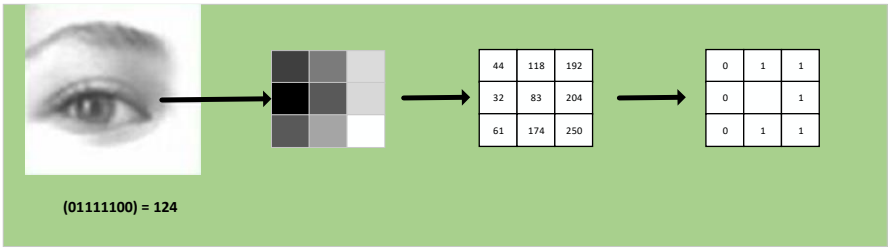


图 2-1 LBP 特征提取流程

从图中可以看出，LBP 特征的计算主要包括三个步骤：首先通过一种 3×3 的矩阵算子，然后第二步以矩阵中心点的像素值为阈值，将其与周围八个像素点进行比较，如果该像素点的值比阈值更大，则将该点的值置为 1；如果比该点的值更小，则置为 0，通过循环比较后，会得到 8 个 01 值，将其按照顺时针方向排列得到一个 8 位的二进制数，将其转化为十进制数则代表该矩阵的 LBP 码，因为总计有 8 位，所以易知该码有 256 种。LBP 码值的大小代表该矩阵区域的局部特征信息。该特征提取比较简单，由于其在不同光照、不同尺度与不同角度的图像中具有一定的鲁棒性等特性，可能用来做纹理识别、对象识别或者检测等任务。

Haar 特征主要通过像素模块法求差值的一种特征。图 2-2 展示了模板块的形式，其中包括黑色矩形和白色的矩形，从图中可以看出其融合方式有四种。首先组成特征模板后，将黑色矩形区域和白色区域的值分别进行求和并相减，最后得到的

值表示 Haar 的特征值。该特征提取的方式简单，对有明显像素变化的图像很有效，但仅限于一些简单的结构，如常规物体以及人脸识别等。

图 2-2 Haar 特征模板

SIFT 特征是由 David G.Lowe 提出的的局部特征提取方法，其特征更具表达性，对于视角变化、光照具有一定的鲁棒性的特点使得即使在深度学习发展成熟的当前在很多领域也发挥着重要的作用。但缺点是特征构建过程比较复杂。特征的生成过程如图 2-3 所示：1）首先在搜索图像空间，使用极值检测方法获得兴趣点；2）第二步获取特征的位置尺度信息；3）第三步在特征方向赋值，主要是通过计算像领域的梯度的方法进行赋值；4）最后进行特征点描述。

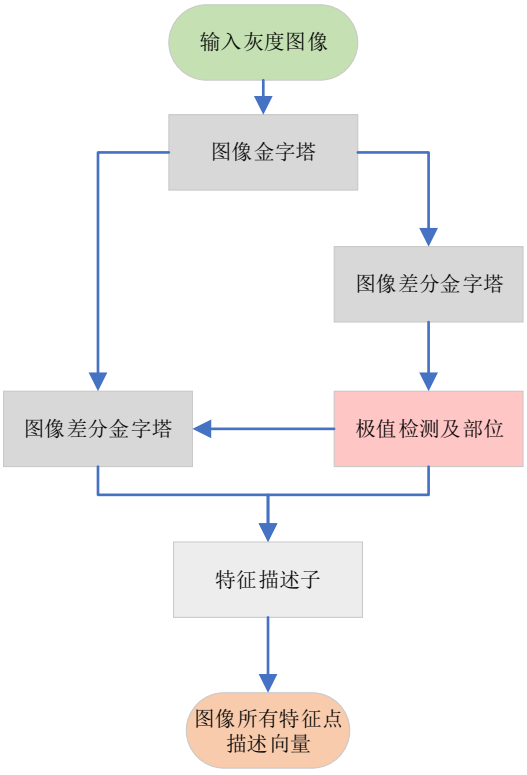


图 2-3 SIFT 特征提取流程

HOG 特征是一种通过计算图像的梯度信息生成的特征。其特征提取流程如下：首先需要将图像做归一化操作，然后计算出其中每个像素的梯度值并统计梯度的直方图，将改直方图作为像素的描述子组成一个 3×3 的区间，最后将输入图像生成的所有描述子整合起来即可得到 HOG 特征。在早期，由于 HOG 特征计算简单，

且在背景单一的人体检测任务上上取得了很好的效果，受到广大学者们的关注，后期很多研究者也将 HOG 特征应用到其他的领域，并取得了不错的效果。

以上几种传统的特征提取方法的缺陷是通用性不高、鲁棒性不好，容易受到复杂环境的影响。这是传统机器学习的弊端。但针对某种特殊的任务，基于传统机器学习的方法比基于深度学习的方法会有更好的效果，比如目前的特征工程方向就是研究手工特征的设计，然后结合深度学习方法可能会有更好的效果。但缺点是效率不高，通常一种任务的特征设计需要耗费很长的时间。

2.2 深度学习技术

深度学习技术是当前计算机技术领域应用最为广泛的技术之一，基于深度学习，我们可以训练出比人类顶尖选手更优秀的围棋手，可以设计出一种在某些领域比人类还具有识别能力的机器人。可以说深度学习就是计算机视觉任务的大脑，在人工智能领域扮演着不可或缺的角色。本节对深度学习领域相关一些理论基础知识做简单的介绍。

2.2.1 深度学习概述

深度学习最开始是源于对神经网络的研究，是一种比较复杂的机器学习算法，在计算机视觉任务和自然语言处理等领域都取得了很好的效果，性能远超过先前基于传统手工特征的方法。其原理是通过搭建深度神经网络模型让机器具有和人脑相同的学习系统，然后根据外界的刺激去学习升级，最终获得和人类几乎相同的视听或者判断能力。目前，深度学习在各个领域的研究中都取得了不错的进展，未来发展的势头也不可阻挡。

2.2.2 神经网络

神经网络(Neual Networks)在人工智能领域扮演着重要作用，目前各领域的优秀算法大部分都是基于神经网络模型。神经网络是由各种神经元连接组成，每个神经元可以通过反向传播方法学习其权重和偏置，从而完成自我能力的提升，其中的原理和人脑很相似。

在卷积神经网络中主要包括两种类型的层，一种是线性层，例如卷积层和全连接层等，其中卷积层是所有层中参数量最多的部分，是卷积神经网络的核心层，通常在网络中都只使用3×3或者1×1的大小的卷积，其余尺寸的卷积都可以由3×3卷积得到，且3×3卷积的参数更少。另一种则是非线性层，例如激活函数 Relu、Sigmod 等。

神经网络中一个卷积层中一般包含很多个滤波器，其中每个滤波器在空间（宽度和高度）上很小，如3×3，1×1，但是在深度维度上却是不确定的，需要根据任务的要求进行设计。二维卷积和一维卷积的方式一致，区别在于二维卷积的滤波器的维度多了一维，滑动窗口是一个矩形，而一维卷积则是向量。其中在卷积过程中，窗口在输入特征上进行滑动，其中设置的超参数通常有 padding、stride、输出宽度以及卷积核的大小等，其他特殊的卷积还有其他参数，如组卷积还需要设置组数。一个滤波器可以生成输出特征中的一个通道，最后将所有滤波器的结果叠加，即可得到最后的输出特征。卷积其实就是一种矩阵乘法，可以将一种特征变化为另一种特征，最后直接可以得到我们为他设定的特征，如分类的概率等。

全连接层其实是特殊的卷积层，只是它的卷积和大小和输入的分辨率一致，该部分主要放到最后一层进行特征的分类，如果分类数太多，全连接层的参数相对较多，容易成为某些分类任务的瓶颈。

池化层的原理和卷积层相似，是卷积层的一种特例，一般来说池化都会将输入的尺寸进行下采样，通常在卷积网络的最浅层和最深层使用。目前常用的池化操作主要有平均池化(Average pooling)和最大池化（Max pooling）。最大池化操作就是将矩形窗口内（一般3×3或者1×1的空间）的元素取最大值保留，其余的元素直接丢弃；平均池化则是将矩形窗口内的元素求和取平均值后保留。

激活函数是非线性层，是一种非线性的函数，该函数可以将整个线性模型转化为非线性模型，从而提高模型的表征能力。由于全连接层、卷积层等都是线性函数，如果整个网络中只有线性层，则该网络只具备线性函数的能力，在较难任务中缺乏表示能力，而引入激活函数的会增强模型的表示能力。目前常用的激活函数有 sigmod、Relu、pRelu 等。

以上几个模块组成神经网络的基本结构，但在反向传播时还需要为整个损失函数设计损失函数以及优化器。当前常用的损失函数有 Softmax、Arcface 等，优化器主要有 Adam 等，通常这两者也会很大程度的影响整个网络的训练效果。

2.2.3 深度学习框架概述

从 2014 年开始，深度学习成为人工智能领域研究的重要方法，为了使该领域的学者们完全精力放在算法的研究上，很多深度学习的框架应运而生，去提高算法的实现效率。目前广受研究学者们欢迎的深度学习框架有 Pytroch、Tensorflow、Mxnet 和 Caffe 等。这些框架将多种功能的接口封装提供给用户，可以很方便的帮助我们进行网络的搭建，提高模型算法的实现效率。同时在运行时将搭建好的网络进行前向计算和反向传播，并且还提供进行参数模型的存储和计算的优化等，这很

大程度上的解放了研究学者，让我们能更好的专注于算法的设计层面，促进了近年来人工智能领域的快速发展。本文主要使用 PyTorch 框架，接下来对其做简要的介绍。

2.2.4 PyTorch 概述

PyTorch 是由 Facebook 研发的深度学习框架，在 2017 年对用户开放其接口。它的底层是 Torch，只是 PyTorch 利用 Python 封装重写了很多内容。PyTorch 的最突出的特点就是可用性，简洁高效的特点受到广大研究学者的青睐。从最开始的 0.4 版本，到现在的 1.7 版本，PyTorch 这四年新增了更多的功能，同时 PyTorch 社区也越来越活跃，俨然已经成为当前最受学者青睐的深度学习框架，在开源代码上绝大多数顶会论文都使用 PyTorch 框架实现。

2.3 目标检测技术

目标检测技术应用广泛，在视频监控、目标跟踪、人数统计等任务中扮演着重要的角色，在各个领域中的得到了广泛的应用。目标检测的关键问题是确定给定图像中是否存在给定的类别的目标，从而返回目标的定位以及对应的类别。下面几个小节主要以传统目标检测和当前流行的目标检测方法进行简要描述。

2.3.1 传统目标检测方法概述

在大数据与硬件计算能力还比较落后的时代，传统目标检测算法受到了广大研究学者们的关注，其主要流程包括候选区域框的生成、特征提取和目标分类三步。传统目标检测算法和深度学习算法的流程一样，都是先提取图像特征，然后使用分类器对其进行分类，不同的是传统算法是基于人工方法设计的特征，例如前面讲的经典的四种方法。但是人工特征不灵活、特征表示能力很弱、不具有很强鲁棒性，在某个特定的数据集上表现良好，但整体的泛化性能并不好。为了提升模型的性能，常常用一些复杂的分类器来进行分类，比如 SVM 分类器[3]等。

其中候选框的定义是：图像中所有可能包含的目标的区域；传统方法使用很朴素的滑动窗口的方法。使用不同大小的框从左上方到右下角从左到右从上到下依次扫过整个图像，这样产生成千上万的框即候选框。很显然，使用这些候选框进行特征提取计算量太大，每一张候选框都要通过相同的方式提取特征，这样会造成大量的重复计算，同时这些候选框绝大多数是多余的，所以这样的方法费时费力。为了改进这个问题，J.RR.Uijlings 等人[22]提出了 Selective Search 算法来生成候选框，不同于先前枚举的方法，该方法是通过相似性对图像进行分割与合并，不断的将相

似的区域进行合并，将明显不同的区域进行分割，最后将合并的块作为候选框保留。该方法极大的提高了生成候选框的效率，同时产生较少了候选框的数量，很大程度的减小了后面分类任务的计算量。

2.3.2 基于深度学习的目标检测技术

随着大数据的发展和 GPU 硬件等资源的提升，基于深度学习的目标检测技术在近几年来得到飞速的发展。从 Two-Stage、One-Stage 检测框架，从基于 Anchor 到现在基于 Anchor-Free 的简易框架，目标检测技术的性能得到稳步的提升，甚至很多算法模型都已经广泛应用到实际场景中了，在我们的生活中发挥着重要的作用。本节主要对当前目标检测方向研究的主要方法进行简要概括。

2.3.2.1 Two-Stage 检测算法

从 2013 年 R-CNN 检测的出现，标志着目标检测正式进入到深度学习的时代，Two-Stage 检测框架应运而生。顾名思义，这种检测框架主要包含两个阶段的处理过程，和传统的检测流程相似：该框架检测流程首先先对整张图片粗略扫描生成候选框区域，然后在对候选区域进行分类和回归，得到目标的位置和大小信息。二阶段检测最具代表性的算法主要有 Fast R-CNN 和 Faster R-CNN 等，下面将以时间顺序对其进行简要概述。

2014 年，R-CNN 一提出就引起了广大学者的关注。R-CNN 采用多任务的方式同时做分类和 b-box 的回归，检测流程就是先使用 SS 方法产生候选区域，然后对候选区域进行特征提取，获取深层特征表示，最后将其再经过一个全连接层和分类器后分别输入到两个分支中分别进行 b-box 预测和分类。该方法的检测流程和传统的相同，唯一区别是该检测算法使用卷积神经网络模型提取图像的区域特征，不仅提升特征提取的效果，而且简化了操作难度；但过于耗时的问题依旧没有解决。

Fast R-CNN[23]较于 R-CNN 不同的是，将整张输入图片经过卷积网络生成特征图，同时为了让模型可以适用不同的图像输入的分辨率和减少候选框区域的数量，其使用 ROI pooling 操作从每个先前 SS 方法生成的候选框中提取一个固定长度的特征向量，其余的流程不变。对比 R-CNN，简化了检测流程，不仅提升算法的推理速，而且提升了算法的检测精度，提高了参数共享率。

Faster R-CNN[24]提出了 RPN 方法来生成候选框，该检测网络放弃了所以传统的模块，而是采用全卷积形式。不仅提升模型的准确率，而且可以进行端到端的训练。其中该检测方法中的候选框直接通过神经网络模型来获取，并且对候选框使用权重共享，这很大程度上的提升了模型的性能以及推理速度。近几年许多 Two-

Stage 检测算法主要都是基于 RCNN 等框架改进得到，主要在其基础上添加了多尺度输入、多尺度特征融合等模块，或者替换特征网络的 方法提升精度，Two-Stage 框架的提出加速目标检测领域的发展，在目标检测领域有划时代的作用。到了 2016 年 YOLO-v1 的出现，标志着 One-Stage 的方法为目标检测打开了一扇新世纪的大门，该 One-Stage 的框架不仅可以端到端的进行训练，在和 Two-Stage 框架的对比中，不仅没有丢失太多精度，而且在检测速度更是比 Two-Stage 的快许多，下一节对其 One-Stage 检测方法进行详细的介绍。

2.3.2.2 One-Stage 检测算法

One-Stage 检测算法直接通过在特征点上进行回归分类，得到类别信息以及目标的位置和大小信息，其整个检测过程只需要一个阶段的处理，即可完成目标检测任务。其与 Two-Stage 检测算法的主要差异是，它不需要生成候选框区域，而是直接对目标框回归，很大程度的简化了检测流程。具体来说，YOLO-v1[25]算法首先将图像输入到头部特征网络中生成一个 Feature map，然后将其进行划分为 S×S 的小格，然后直接通过几层网络对其进行回归任务，生成目标的坐标以及类别。该算法相较于 Two-Stage 检测算法，推理速度更快、整体结构更简单，但模型的准确率较低，很多一阶段检测算法在嵌入式设备中可以实时运行，这很符合现实场景中的应用要求。其中一阶段检测算法中最具代表性的算法有 YOLO 系列[25][26][27]算法、SSD[28]以及 RetinaNet[29]等。

2.3.2.3 Anchor-Free 检测方法

以上两种检测方法是根据是否需要生成候选区域来区别。而在 2019 年，以在训练中是否需要通过预先设置的 Anchor 获取候选框，将检测方法分为 Anchor-Base 和 Anchor-Free 的两种方法。其中上述介绍的大多数都是基于 Anchor-Base 的方法，即都需要通过聚类算法对训练集中的目标框大小聚类，然后在训练中设置，这样的好处是可以提升模型的训练效果、提升模型的准确率。而基于 Anchor-Free 的方法则不需要这一处理过程，而是直接回归目标框的位置和大小信息。该方法设置不需要通过 NMS 方法去除重复框，这很大程度的优化了检测模型的结构，对于算法的应用部署有很大的好处。其中基于 Anchor-Free 最具代表性的算法有 CenterNet 等，该方法直接基于大分辨率的 heatmap 来回归中心点的坐标，每一个网络最多回归一个中心点，同时对中心点的偏移量和目标框的长和框进行预测，最后结合三种预测结果就可以得到目标框的位置和大小信息。在准确率上可以与基于 Anchor 的高精度算法相媲美，且操作流程更简单。

2.3.3 目标检测评价标准

在目标检测领域，比较流行的公开训练测试数据集包括 COCO、VOC 等，目前大多数通用检测算法的性能测试都是基于 COCO 数据集。基于 COCO 数据集，就提出了一种准确率评价指标 Map 以及各领域常用的速度评价指标 FPS（推理速度评价指标，即处理每张图片所需的时间，且该性能指标需要在同一硬件下进行比较），其中 FPS 通常和模型的参数量，FLOPs（浮点运算次数，理解为计算量）、FLOPS(硬件设备计算能力)等相关，计算较简单，在这里就不详细介绍。下面对 map 评价指标仔细介绍介绍。

在目标检测的评价标准中，map 是最重要的准确率评价指标之一，根据数据集的发展，map 计算方式在随着改变，主要分为三个阶段：1）VOC2012 数据集之前；2）VOC2012 数据集之后；3）coco 数据集。在这里我们对当前的 COCO 数据集的评价计算方式进行详细的介绍。

map 是数据集中所有类别 AP（准确率）值的平均值，计算 AP 的如公式$(2-1)$和$(2-2)$所示：

$$\text{Precision} = \text{TP}/(\text{TP}+\text{FP}) \tag{2-1}$$

$$\text{Recall} = \text{TP}/(\text{TP}+\text{FN}) \tag{2-2}$$

在式（2-1）中，Precision 表示准确率，Recall 表示召回率，分别表示被检测出来的实例的正样本正确个数占检测出的总个数的比例和分类准确检测正样本的实例个数占总正样本个数的比例。

公式组中 FP、TP 和 FN 分别表示真实为负样本但被模型分为正样本的实例个数、真实为正样本但被网络检测算法划分为正样本的实例个数以及真实为正样本但被网络模型划分为负样本的实例个数。

上述描述的性能评价标准是当前最通用的标准之一，针对不同数据集的特点可以设定不同的标准，如某些任务只将 AP 的最大值或者指定 IOU 下的召回率作为目标检测的标准，这也是合理的。

2.4 人体行为识别技术

近年来，随着科技的快速发展，一些难度更大、更加抽象的任务开始被广大学者们关注，这其中包括人体行为识别技术。人体行为识别其实就是一种分类任务，但与普通的分类任务（如人脸分类，猫狗分类等）不同，行为识别更加的抽象，受到环境干扰的可能性更大。作为研究意义以及应用前景巨大的邻域，人体行为识别在未来会成为视频监控甚至国防安全的重要应用的技术支撑。本章对该技术理论

做详细介绍，主要包括基于手工特征的方法和基于深度学习的行为识别方法以及人体姿态估计方法三方面。

2.4.1 基于手工特征的人体行为识别方法

基于手工特征的人体行为识别是一种比较传统的方法，该方法需要专家根据不同任务设计不同的特征提取方法，效率较低。下面对最具代表的手工特征方法做简单的介绍。

其中一种是基于轮廓剪影的特征提取方法，其最手工特征具代表的算法，是由 Bobick 等[30]提出的，该方法通过从图像中提取人体的轮廓信息，从而建立时序图分析进行行为识别。该方法的优点是特征具有很强的描述能力，对于简单背景的图像性能较高。但是对于复杂背景的场景（如相互遮挡、有较大噪声的场景）由于干扰太大很难提取到人体的轮廓信息。

基于人体关节点的特征提取方法是通过姿态估计得到人体各个关节点的坐标信息，然后通过这些坐标信息就可以对人的行为进行表示。该方法最开始与 Fujiyoshi 等[31]用人体的头和四肢的总共 25 个关节点来表示人体的姿态。由于关节点对人体行为有很大管关联性，所以该方法在某些任务上取得了不错的效果。目前很多基于深度学习的算法也是基于姿态的方法，下一小节会详细介绍。

2.4.2 基于深度学习的人体姿态估计方法

由于人体行为类别和人体的姿态的关联性强，人体姿态估计成为行为识别技术研究中必要的技术支撑。基于姿态的人体行为识别也是行为识别研究方法中主要的方法。目前，在人体姿态估计的算法中，主要包含两类算法：一种是基于 Top-Down 的研究方法；另一种是基于 Bottom-Up 的方法。下面对这两类方法做详细的介绍。

Top-Down 的方法[34][35][36]的主要流程包括：1）先用人体检测技术获取图像中每个人的位置和大小的信息。2）然后对图像中的人体关键点做预测。3）结合两者的输出结果，得到每一个人的关节点图。该方法的准确率相对更高，但复杂的流程导致此类算法的效率相对更低效。而在关节点估计的方法中，主要通过设计一种高分辨率的网络结构提升关节点估计的准确率，如 Hrnet[34]等网络结构，在数据集能够达到很不错的效果，但是由于计算量太大，该方法达不到实时的效果。

Bottom-Up 的方法不需要人体检测技术的支持可以直接回归出关节点的坐标，然后自动将图像中的各关节点聚类。该方法的主要流程包括：先生成图像中的所有的人体关节点，同时生成各关节点间的联系矩阵，然后结合两者的输出可以将各关

节点分配到属于其人体的位置。该研究方法的效率更高，可以在嵌入式设备上实时运行。但准确率相对较低，且目前算法获取各关节点的联系矩阵的办法不多，准确率还有待提高。该方向最具代表性的算法有 OpenPose[37]等。

2.4.3 基于深度学习的人体行为识别方法

传统手工特征的方法对于复杂场景中具有光照、遮挡、视角变化等问题的任务不具有普适性，随着大数据和计算机硬件的发展，深度神经网络迅速发展壮大，从 2014 年开始，深度学习成为了各种计算机视觉领域的主流方法，由于其在计算机视觉任务中效率更高、准确率更高，深度学习方法几乎完全替代了传统的手工特征提取方法，成为主流的研究方法。

基于深度学习的人体行为识别方法的流程如图 2-4 所示，主要包括两个步骤，只需要通过深度神经网络模型自动学习视频的行为表征，从而完成自动分类，该方法可以进行端到端的训练，且操作简单、效率更高，在计算机视觉任务中应用广泛。

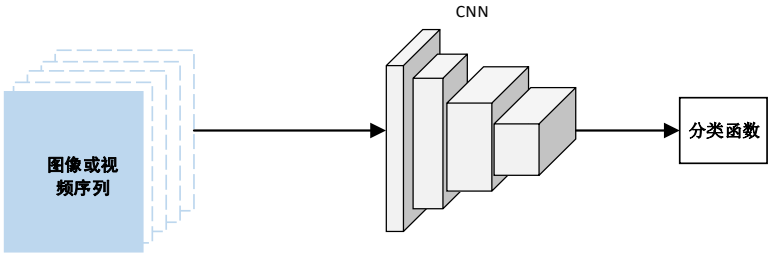


图 2-4 深度学习识别方法流程

一般而言，人体行为和许多因素相关，如人体的外观、姿势、光流等。这在学术研究上被称之为人体的模态，即构成人体行为的表达的特征。在当前的研究方法中，对应不同的模态可以提出了不同的研究方法。目前，比较主流的人体行为识别方法主要有下述几种。

Simonyan[38]在 2014 首先提出基于双流网络的行为识别方法，其基本流程如下图 2-5 所示。该模型分为两个分支，一个分支使用卷积网络对图像提取空间特征，该分支和普通的分类网络结构类似；而另一个分支首先将相邻图像转化为光流图像，然后利用神经网络从中提取时序信息，该方法可以显著提升精度。但其缺点也很明显，但需要预先对视频提取光流图像，且两个分支的训练是独立的，计算量相对较大，难以达到实时性的要求。

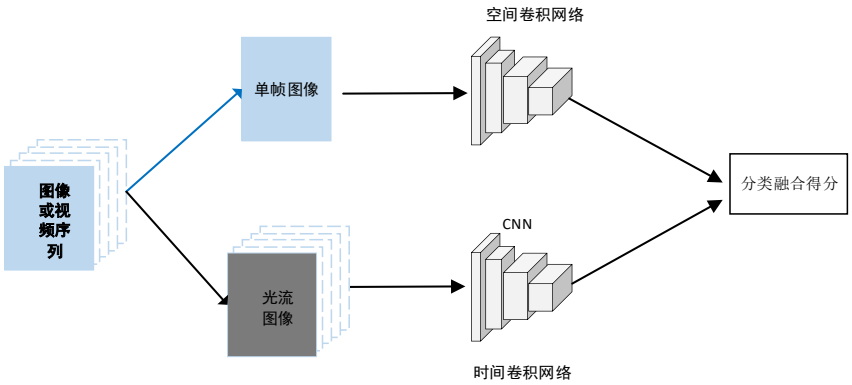


图 2-5 双流网络框架

由于视频相对于图像多了一维时间序列，所以容易想到将卷积操作变化成 3 维的方法是有效的。Ji[39]在 2010 年首次将 3D 卷积网络应用到计算机视觉任务中，和 2D 卷积的操作方式相同，只是相比于 2D 卷积，三维卷积多了一维时间维。但计算复杂度却高一个量级，对硬件性能要求更高，在实际场景的应用时很受限制。并且在人体行为识别的精度整体上比双流网络方法要低。因此，目前最先进的方法通常是结合了双流网络的思想，然后对此做了一些改进从而提升行为识别方法的性能。

由于人体的结构一致，人体骨骼模态对人体行为动作的关系密切，在某些任务中一种人体行为会对应人体特定的姿态，很显然人体骨架的状态可以代表人体行为的高级抽象的语义。人体相同关节点时间位置变化以及不同关节点的相对变化对于行为识别的分类有重要意义，通过以上方法就可以进行人体行为识别分析，整个骨架图序列如图 2-6 所示。基于骨骼的人体动作识别方法最开始只是在时间序列上直接生成关节点 heatmap，然后生成关节点的坐标信息，进而对其进行时序特征分析[40]，这类方法结构简单，实时性能较好，结合外观特征能够有效的提升行为识别算法的性能，但并没有考虑人体关节点之间的空间位置联系，而这种关系对于分析人体行为有着重要的作用。最近，研究者开发了试图利用关节间自然连接的新方法[41]，实验结果表明骨骼连通性在人体行为识别分析时序信息是多么重要，但是目前大多行为识别算法过于依赖通过手动设定的规则去分析关节点空间位置关系，导致算法难以应用到其他任务或者领域中。后面，在 STN[42]方法中提出了一种能自动捕捉关节的空间构型、时间动态的方法，在准确率上有比较大的提升。

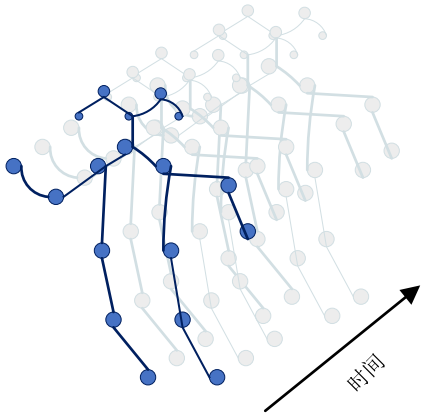


图 2-6 骨架图

2.5 轻量级网络概述

轻量级网络对算法的加速有明显的作用，性能良好的轻量级网络不仅可以提高算法模型的性能，而且能够减少算法的计算量，让模型能够在各个平台上都能保持优秀的实时性能。目前经过学者们的深入研究，常用的轻量级网络包括 SqueezeNet、ShuffleNet 系列和 MobilleNet 系列等。下面对其进行详细介绍。

2.5.1 SqueezeNet 系列

SqueezeNet 是深度学习领域最早设计的一批轻量级网络之一，该网络的核心是 Fire 模块，其结构如图 2-7 所示。该模块总计有三层卷积层，首先第一层 1×1 卷积对输入的通道数进行压缩，然后再进行 3×3 卷积操作，最后使用 1×1 的卷积对模型的通道数进行扩张。与直接将输入通过 3×3 卷积相比，Fire 模块的计算量更低。其中该模块一般遵守如下基本原则：如果第一层 1×1 卷积核数s_1，3×3 的卷积核数为e_3，1×1 扩张层的卷积核数为e_1，一般$s_1 < e_1 + e_3$。

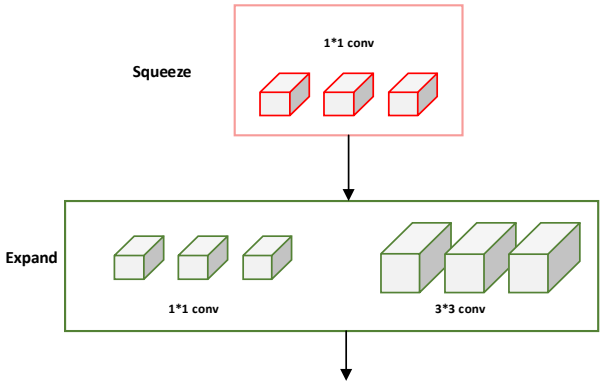


图 2-7 Fire-Module

基于 SqueezeNet 网络的升级版本 SqueezeNext 则是全部使用卷积网络，通过对于网络整体结构的优化，更进一步提升模型的对速度和表征能力。该网络系列的思想广泛应用到各个领域中。

2.5.2 MobilleNet 系列

MobilleNet 系列[45][46]是轻量级网络家族中比较常用的网络结构，因为该网络针对嵌入式设备，对网络结构和卷积方法做了相应优化， 如 MobilleNet-v1 使用的卷积方式为深度可分离卷积，如图 2-8 所示。这种卷积方式可以减小网络的计算量，且在 cpu 硬件上有特殊加速效果，所以其常用于嵌入设备中。MobilleNet-v2 在 MobilleNet-v1 的基础上，提出了某些创新结构，如 Swish 激活函数、自适应学习模块等，但使用最广泛的还是 MobilleNet-v1 网络。

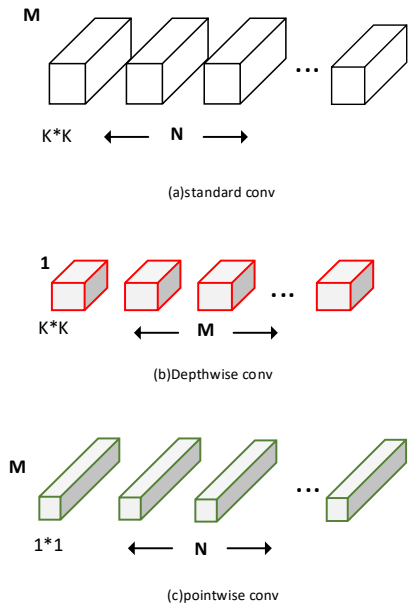


图 2-8 卷积方式

2.5.3 ShuffleNet 系列

MobileNet 系列网络中使用深度可分离卷积，由于这种卷积方式通道之间信息交流不足，容易丢失通道信息。导致最后生成的特征表示能力有限。ShuffleNet 系列[43][44]是则是基于这一点对其进行改进， ShuffleNet-v1 提出了通道 shuffle 操作，这弥补了分组卷积的缺点，在减少网络的部分计算量的同时，提升网络表征能力。在嵌入式设备中，MobileNet-v1 是最广泛使用的网络之一。

ShuffleNet-v1 则是最先提出了通道 shuffle 的操作，使用分组卷积的方式减少了模型的计算量。而 ShuffleNet-v2[43]则几乎完全推倒了 shuffleNet-v1 网络的设计，该网络首先提出 channel split 的方法对通道进行操作，该方法进一步提升网络的性能。

2.6 本章小结

本章主要介绍了和本文相关的理论技术。首先对传统的机器学习方法和深度学习理论进行了简要的介绍，其中包括两者的区别与联系，并分析了两者的优点和缺点。接下来介绍了目标检测技术和人体行为识别技术相关的研究方法理论，其中在目标检测领域，对当前方法做了总结；在人体行为识别技术中，对当前主流的研究算法进行了简单的介绍。最后对当前流行的轻量级网络做了简要的介绍。基于以上相关的理论知识，本文在后面章节对人体检测与行为识别技术做更深度的研究与讨论。

第三章 轻量级网络的设计与实现

随着计算机技术的发展以及计算机硬件设备的计算能力不断提高，人工智能技术快速发展成熟。基于深度学习的人体检测与行为识别技术也取得了阶段性进展。在人体检测领域，当前已有的算法在复杂的现实场景下的推理时延和准确率往往顾此失彼。在人体行为识别领域，由于行为样本具有时序性，导致设计的人体行为识别算法的太过复杂、计算量太大，目前在某些 GPU 硬件平台也达不到很高的准确率和实时性能。因此，在保持高准确率的情况下，如何优化人体检测与行为识别算法，提高算法的准确率与推理速度是本章的关键问题。

本章的主要研究目的是设计一种轻量级的通用特征提取网络，为后续章节的人体检测与行为识别算法提供网络支持。由于目前大多数的高性能网络都是通过裁剪、量化等方法实现模型计算量减小，网络结构通常比较复杂，在模型的内存消耗、时间消耗等方面并没有针对性的优化，且通常有精度损失。本文先进行了大量的对比实验，设计了一种有效的网络基础模块，然后基于该模块设计了一种更轻量化、更具鲁棒性的特征提取网络。并利用网络 op 融合的方法简化模型的结构，提升模型的推理速度。最后，构建几种复杂的测试数据集对算法模型做大量的对比试验。该网络在 1080TI 硬件条件下的前向推理速度和准确度优于当前的一些流行的轻量级网络，如 ResNet-18、EfficientNet-B0、MobilleNet-v1 等。

3.1. 数据集准备

为了验证特征提取网络的性能，通常需要构建一种高质量的训练集与测试集，该数据集要具备类别多、数据量大、且场景复杂多变等特点。为了达到这个目的，本文收集、制作了一些人脸数据，然后结合当前公开的人脸数据集，构建出一种数据量大、场景复杂的高质量数据集。下面几个小节对其进行详细的介绍。

3.1.1 训练数据集

最开始本文基于构建的数据集的全部样本作为训练集做一阶段训练并测试，发现结果并不理想。因此设计了一种两阶段的训练方法，最终将使用整个训练数据集拆分为二个分别进行各自阶段的训练，具体构建和拆分细节如下：

1、Msra_Celeb 数据集：该数据集用于第一阶段的训练，由 MS-Celeb-1M-v1c 和亚裔名人数据集整合而来。其中 MS-Celeb-1M-v1c 具有 86,876 个 id /3,923,399 个从 MS-Celeb-1M 数据集中清理的对齐图像。该数据集已被排除在 LFW 和 Asian-

Celeb 之外。亚裔名人数据集包含 93,979 个 id /2,830,146 张对齐的图片。LFW 和 MS-Celeb-1M-v1c 均不包括此数据集。

2、Msra_Celeb_id40：用于第二阶段的微调，包含了 idcard 数据集的部分图片和 msra_celeb 中的部分图片。其中 idcard 数据集是人脸身份证 id 数据，包含超过千万张图像和上万个类别。图像采样的流程是：先设置图片的上限，然后通过随机采样的方式获得图片中的一部分作为数据集。

从以上描述中，两种训练集有大量的重复样本，训练细节在本章测试中会详细介绍。

3.1.2 测试数据集

为了验证网络模型的鲁棒性，本章中使用的测试数据集一共有三个。第一个是公开的人脸测试数据集 LFW，该数据集在人脸测试中普遍使用；为了更好的测试特征网络的性能，本章自研了两个更难的测试数据集，其中一个测试数据集是实验室构建了一个人脸动态抓拍的数据集，另一个是室外人脸抓拍数据集；此外，在进行测试时还需要自建一种干扰数据集作为干扰图像，表 3-1 中对这几种数据进行简单的介绍。下面对这几种数据集做详细介绍。

表 3-1 数据集介绍

数据集	类别数	样本数量	样本来源
LFW 数据集	-	6000 对人脸	互联网
动态人脸数据	40	27139	动态抓拍
室外人脸数据	60	22055	摄像头抓拍
干扰数据集	-	30 万张	摄像头抓拍

LFW 数据集是当前人脸识别测试精度最主要的几个数据集之一，该数据集是从数据库中随机选取的 6000 对人脸图像，其中 3000 对人脸数据中的每一对都属于同一个人，另外 3000 对人脸则属于不同的人。在模型上的测试流程是：首先通过 LFW 数据集给出一对照片，通过模型验证是否该对图像属于同一个人（一般根据特征相似率来判断是否属于同一个人，本文使用的是 L2 相似度），然后与正确结果进行比较，最终所有人脸对测试完成即可得到该数据集下的测试准确率。

动态人脸数据集为真实的动态人脸抓拍数据，共包含 60 类，总计 22055 张图像样例。该数据集的采集方式是在运动状态下进行抓拍，这导致数据中有大量的运动模糊的图像，从而加大了人脸识别的难度。

室外抓拍数据集共包含 40 类，总计 27139 张测试图像，是由室外摄像头采集而来。其中存在大量遮挡、大角度人脸的样本图像。

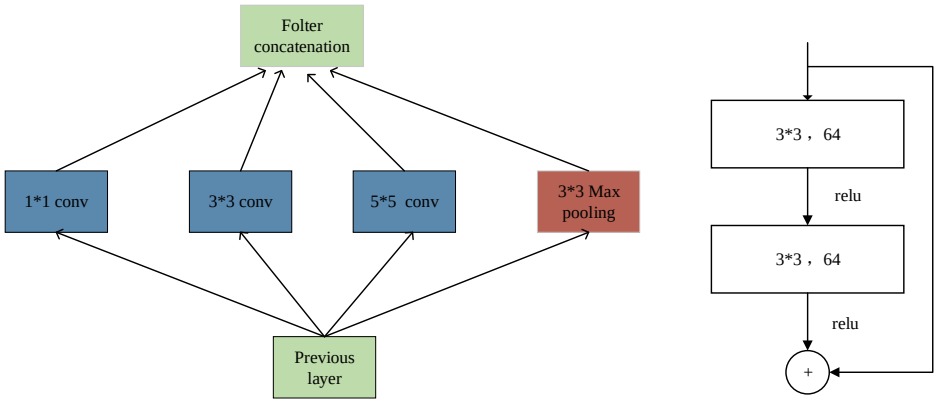
干扰数据集包含 30 万张由室内外摄像头采集的人脸数据，其中包括遮挡、运动模糊等难样本，用于测试时作为干扰图像。通常干扰图像越多，误检率就越高。

3.2 轻量级网络的详细设计

特征提取网络作为计算机视觉领域中最重要的模块，其性能的优劣影响着整个任务的性能好坏。通过调研发现，一种好的主干网络在人工智能各个领域的应用通常都是优秀的，例如 ResNets、MobilleNets，EfficientNets 等，这些网络最开始都是基于人脸数据集进行设计和验证，但在目标检测领域、语音设别等领域也都表现良好，可见高性能的特征提取网络大体上在人工智能相关任务中都是通用的。轻量化网络通常意味着计算量更小（也可能计算量大，但前向推理时间短，后续我们会验证）、前向推理速度更快。本章构建一种多类别、多场景的复杂人脸训练集数据集和难度不一的测试集，基于 ResNets 设计了一种结构简单轻量化的特征提取网络，基于构建的数据集对网络模型的性能做评估，最终该轻量化网络的准确率和前向推理速度都要优于当前的一些轻量级的网络。下面几节对算法模型设计的原理做详细的介绍。

3.2.1 网络 block 设计

目前许多研究方法都通过设计复杂的卷积神经网络去提升算法模型的准确率，例如通过设计多分支结构（如 Inception[47]网络中的不同大小卷积的分支级联和 ResNets 中的残差结构等，如图 3-1 所示）、深度分离卷积、点卷积等方法去提升网络的性能指标。尽管复杂的卷积神经网络在准确率上往往都比大多数简单的卷积神经网络更好，但它的缺点也很明显，主要有以下几点缺陷：



(a) Xception block　　(b) ResNet block

图 3-1 分支结构

1）复杂多分支的网络结构设计不仅使模型难以解释和实现，而且会降低模型的推理速度、提升内存消耗，这对硬件的计算能力和内存有更高的要求。

2）网络中的一些模块（例如，MobilleNets[45][46]和 Xception[47] 中的深度转换以及 ShuffleNets[43][44]中的通道混洗）会增加了内存访问成本，降低模型的推理速度。

除了以上几点外，影响推理速度的因素还有很多，例如网络深度、宽度等，但网络的浮点运算（FLOP）的数量不能绝对准确的反映实际速度，推理速度也和硬件环境等其他因素有关。表 3-2 展示了各轻量级网络的 FLOPs 和推理速度的大小的测试结果，各项性能结果在 1080Ti 上测试，Batch_Size 的大小为 128，其中速度表示一秒钟可以推理多少张样本[51]。

表 3-2 各网络性能对比

网络	TOP-1 Acc	速度	Params(M)	Theo FLOPs(B)
ResNet-18	77.16	3256	8.30	1.4
ResNet-34	71.16	2442	11.68	1.8
ResNet-50	76.31	719	25.53	3.9
EfficientNet-B0	75.11	829	5.26	0.4
ResNext-32GF	77.98	671	15.26	3.2
ResNext-50	77.46	484	24.99	4.2
ResNet-101	77.21	430	44.49	7.6
VGG-16	72.21	415	138.35	15.5
ResNet-152	77.78	297	60.11	11.3
ResNext-101	78.42	295	44.10	8.0

可以发现，一些轻量级网络模型的 FLOPs（如 EfficientNet-B0，ResNet-152 等）比一些经典的网络模型（如 VGG 和 ResNet 系列 [49]）更小，但它们的推理速度并不一定比经典的网络模型快。这也是工业界和学术界仍然在大量应用 VGG 和 ResNets 等经典网络初始版本的原因。

为了证明上述结论，本节做了以下实验进行验证。本次实验环境是在 PyTorch 深度学习框架和 1080TI GPU 的硬件环境下进行的，首先我们在 1080TI 24 G 内存的硬件条件下测试各种大小卷积操作的相对时间后，发现 3×3 的卷积速度相对最快，如表 3-3 所示：

表 3-3 不同卷积下的 FLOPs 与时间开销

卷积核大小	理论 FLOPs(B)	时间开销	真实 FLOPs
1×1	420.9	84.5	9.96
3×3	3788.1	198.8	38.10
5×5	10522.6	2092.5	10.57
7×7	20624.4	4394.43	9.38

本次实验设计的输入通道和输出通道都为 2018，输入图像的分辨率是 56×56，执行的卷积操作中 stride 都为 1，其中推理时间的结果是在硬件预热后取 100 次中的平均值。可以看出 3×3 的卷积的耗时是相对最短的，因为 1080TI 的硬件对其做了计算优化。基于以上结论不难想到，如果设计一个完全由 3×3 的卷积所组成的网络，是不是能提升模型的前向推理速度呢？基于以上猜想，下面设计一个实验对其进行验证。

根据以有结论和以上猜想，设计了一个只包含简单 3×3 卷积的网络模型，其整体网络结构如图 3-2 所示：

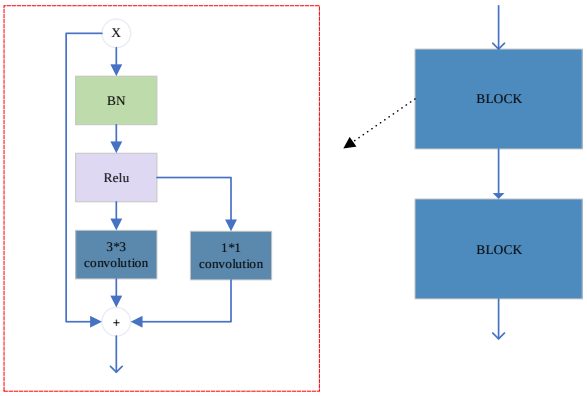


图 3-2 网络结构

可以看出，该网络一个 Block 中只有一层卷积，且该网络没有任何旁支结构。基于这种 Block 结构，设计了 18 层和 34 层的网络结构，其结构与其对应的 ResNet-18 和 ResNet-34 大致相同，唯一区别是没有残差分支。本章分别称之为 BaseNet-18 和 BaseNet-34。为了性能对比，在推理时使用特征融合的操作对其进行加速，然后分别与 ResNet-18 和 ResNet-34 网络模型进行对比实验。在训练中，输入图像的分辨率为 112×96，对所有网络都使用相同的数据处理和损失函数，整个分类网络只有主干网络不一致。训练完成后，在 1080TI 测试，Batch_Size 的大小为 1，在 LFW 测试集上的验证的结果如表 3-4 所示。

表 3-4 性能对比

model	内存消耗(M)	参数量(M)	推理速度(ms)	FLOPs	精度
ResNet-18	1329	11.2	3.2	1.5	99.10
ResNet-34	1389	21.3	5.4	3.2	99.30
BaseNet-18	1367	14.3	2.5	2.4	-
BaseNet-34	1533	16.7	4.3	3.6	-

其中 BaseNet 网络的推理速度是网络 op 融合优化后的推理速度，表 3-4 中与 ResNet-18 和 ResNet-34 进行对比结果表明。由于没有残差结构，导致 BaseNet 网络在训练中收敛效果不好，其在测试数据集上的准确率肯定不如 ResNets 网络。但 BaseNet 的优势是内存消耗更低，且前向推理速度更快。总结出现以上实验结果的原因主要有以下两点：

1、目前许多硬件或者推理引擎都会针对 3×3 卷积做了特定的速度优化，假如一个网络中的每一层卷积操作都可以节省几毫秒，对于几十甚至上百层的网络模型的加速效果不言而喻。

2、在内存消耗上，如图 3-3 所示。对于具有残差结构的分支，需要当所有的分支都完成各自的计算任务后，才能进行结果相加，而在这期间的中间结果都会保存在内存中，显然增大了内存的消耗，同时来回的内存操作会降低整个网络的推理速度，实验结果与理论分析相符。

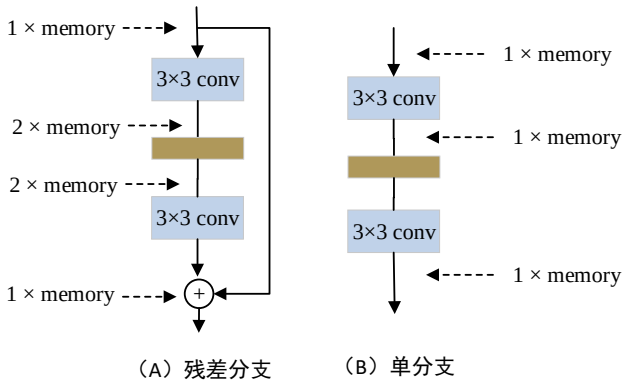


图 3-3 内存消耗对比图

以上实验结果验证了复杂多分支网络的推理速度慢和内存消耗大的缺陷。在训练中发现，BaseNet 由于没有分支，会出现最后无法完全收敛的情况，这也侧面证明有旁支结构的网络具有更好的抽象特征的能力。同时对模型提升训练效果有帮助，在训练的时候会防止梯度消失。

本文研究的目的不仅是提升神经网络模型的准确度，而且还需要提升模型的前向推理速度。从以上实验可以验证普通的单分支卷积网络有很多优点，如结构简

单、速度快等；但其却有一个致命的缺点：准确率太低、训练收敛慢。ResNets 成功的一个解释是，这种多分支架构使模型成为众多浅层模型的隐式集合。具体来说，由于每一个 Block 拥有两个分支，对于具有 n 个 Block 的网络模型，该网络模型可以解释成2^n个网络模型的集合，同时多分支似乎对训练有好处，显然这对于模型的性能提升有很大帮助[32]。上述实验也证明复杂的多分支结构确实可以提升网络的准确率和训练的效果，但多分支拓扑并不是完美的，在推理方面具有较大延时。在 RepVGG[51]中提出的推理时分支融合方法能够在推理阶段将多分支网络的结构转化为的单分支的结构，很好的解决了多分支的推理时问题。受 ResNets[49]和 RepVGG[51]的启发，本文设计了一种简单高效的轻量化网络 LimitNet，下面小节对该网络设计进行详细的介绍。

3.2.2 轻量级网络 LimitNet 设计

首先在网络的卷积核大小和激活函数选取上，由于在各个 GPU 硬件平台对一些普通的层计算都做了优化。为了网络的通用性和高效性，本章设计的网络 Block 都是基于最简单的 3×3 普通卷积核、1×1 卷积和 Relu 激活函数，网络的基本 Block 如图 3-4 所示：

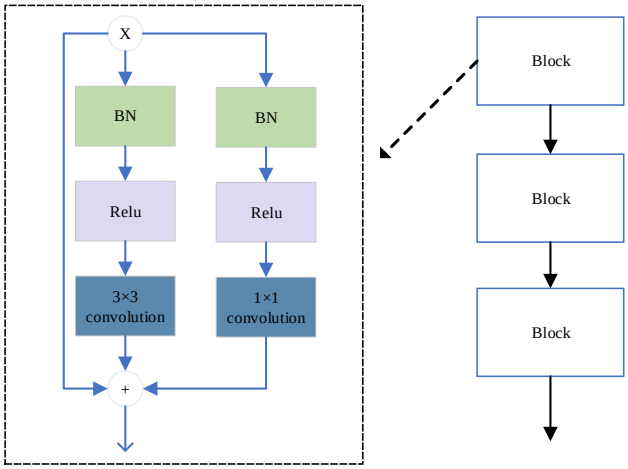


图 3-4 网络 Block

该结构和残差模块类似，区别是标准残差模块会进行跨层特征融合，而该网络的基本模块是所有相邻层之间进行特征融合，为了方便后续描述，称之为 LimitNet 网络。同时为了在模型推理时优化 LimitNe 的网络结构，使用网络 op 融合的方法进行操作的融合，下一节会对其详细介绍。

为了再更清晰的对比，图 3-5 展示了原始的 ResNet、LimitNet 训练网络和 LimitNet 推理时网络结构的对比。

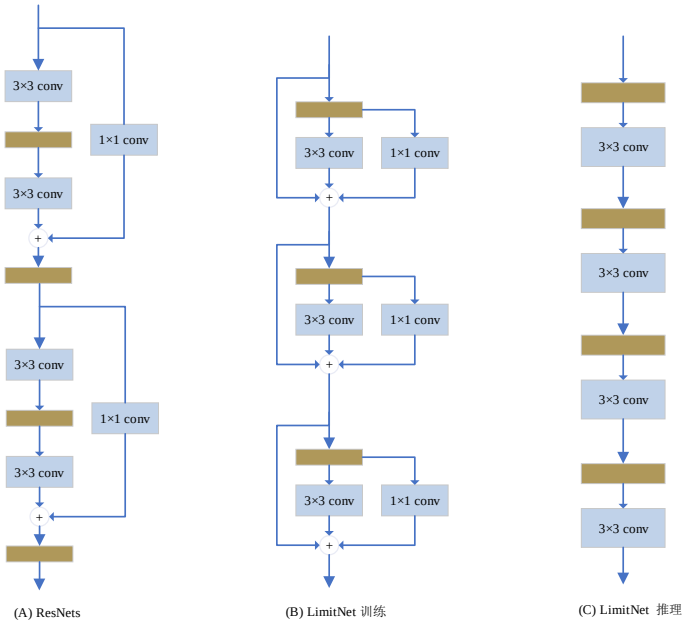


图 3-5 网络结构对比图

图 3-5（A）表示的是经典的 ResNets 网络的初始版本，该网络中包含着 1×1 卷积、3×3 卷积的残差结构以及 Identity 的残差结构（相邻 Bolck 的输出之间 add），ResNets 中的基础 Block 中每两层 3×3 卷积会采用 1×1 的卷积做了一个短连接层，如果输入和输出 tensor 的维度一样，就直接相加，不然就会进行 1×1 的卷积进行维度的改变，然后就只是简单 block 的叠加，在[32]中证明了残差结构能够有效地解决深度神经网络模型梯度消失问题，同时在训练时使得模型更容易收敛。图 B 则表示的是训练阶段的 LimitNet 的网络结构，对比图 A 和图 B，两者的网络主体结构大体一致，每一个 Block 中都包含残差结构；基于 pre_activate 的思想（训练时收敛更快）本章在 LimitNet 的网络中将 BN 层和 Relu 层前置，而且每一层都会进行一次 1×1 短链接和 x 的短链接，然后进行 block 的简单堆叠。；两种网络结构大体上一致，两个网络的主要差异有以下几点：1）LimitNet 网络中的残差块并没有跨层，而 ResNet 则存在跨层连接；2）LimitNet 中将 BN 层和 Relu 层前置；3）LimitNet 的残差模块采用三分支的形式，对比 ResNet 的 2n 分支，LimitNet 有 3n，分支越多模型的特征表达应该会更好。

图 3-5（C）则是 LimitNet 推理阶段的网络,通过网络 op 融合方法（下一节会详细介绍）融合 Bn 层和卷积层从而将多分支结构转化为单分支结构，该网络结构非常简单，整个网络都是由3×3卷积、BN 和 Relu 堆叠而成，这种方式有利于模型的推理加速，下面对网络设计的细节进行详细介绍。

与 ResNet 模型的架构类似，我们将卷积过程分成 5 个阶段，在每个阶段的卷

积层分配上，我们主要遵循以下三点基本准则：1）由于图像一般以高分辨率输入，如果在第一阶段设置多层，这很加大模型的延迟，因此在第一阶段仅使用一层卷积操作进行下采样两倍来降低时间延迟；2）通道数越多特征保存的信息越多，最后一个阶段应该具有更多通道，因此我们仅通过一层卷积操作保存参数；3）根据近年对于 ResNet 等模型的每阶段层数设计的经验（例如，大模型 ResNet-101 在其7×6的分辨率中使用了 69 层卷积操作），我们将最多的层放到倒数第二阶段（最终在训练数据上具有7×6的输出分辨率）。我们让这五个阶段分别具有 1、2、4、16、1 层，以构造一个名为 LimitNet-A 的实例。我们还构建了更深的 LimitNet-B 网络，其中在第二、三、又增加 2 层，z 在第四阶段增加四层，即每阶段具有 1、4、6、20、1。我们使用 LimitNet-A 与其他轻量级网络（ResNet-18、MobilleNet-v1 和 ShuffleNet-v1 等）和中等网络模型（包括 ResNet- 34、ResNet50）竞争性能的。

在网络通道数的设计上，根据经典网络（ResNets 等）的宽度设置从低层到高层分别为 64、128、 256 和 512，我们通过均匀缩放的方式设置（例如 VGG、ResNet 系列、ResNext 系列等）来确定 LimitNet 网络的图层通道深度。因为网络的大多数层在前四阶段，所以我们在前四个阶段使用乘数比例 α，而在最后一个阶段使用 β 的乘法系数，因为通常希望最后一层的输出对分类或其他任务（如检测任务）具有更丰富、更具表示能力的特征，而且由于 LimitNet 在第五阶段中只有一层卷积操作，设置较大的 β 值不会对模型产生显著的延时。最终我们设计的网络宽度可以表示为[64α; 128α; 256α; 512β]，由于第一层在比较大的分辨率的空间上卷积，为了减小计算量，所以在第一阶段的网络宽度则是 min(64; 64α)，这保证了网络从浅层到深层的层宽度保持非递减的特点。

基于以上基础原则，我们设计了 LimitNet-A、LimitNet-B 的两种网络的架构如表 3-5 所示,其中 LimitNet-A 中每阶段的网络宽度分别为[48，48，96，192，1280]，而 LimitNet-B 中为[64，128，256，256，2048]。

表 3-5 网络架构图

Stage	Output_size	LimitNet-A	LimitNet-B
Stage-1	56×48× $C1$	3×3,64 stride 2	
Stage-2	28×24× $C2$	[3×3] ×2	[3×3] ×4
Stage-3	14×12× $C3$	[3×3] ×4	[3×3] ×6
Stage-4	7×6× $C4$	[3×3] ×16	[3×3] ×20
Stage-5	4×3× $C5$	[3×3] ×1	[3×3] ×1
output	1×1	Average pooling，fc	
FLOPs		0.33	2.5

为了进一步减少参数和计算量，将网络的某些卷积层用其他卷积方法替换，如深度可分离卷积、组卷积等，这样可以针对不同的硬件平台使用不同的卷积方式（如 MobileNet 中采用的深度可分离卷积在 cpu 平台上加速效果比 GPU 硬件条件下的更好）。由于 ShuffleNet[43]中说明了对于可分离卷积核组卷积方法会造成通道信息丢失，从而对于提取抽象特征有一定的减弱。为了保持准确率与推理速度的平衡，本文在网络的奇数层使用普通卷积，而在偶数层使用深度可分离卷积，但在网络的最后一层和第一层会使用普通卷积，这是为了在最后保留到更好的信息。这种网络结构减弱了深度可分离卷积对网络整体的影响，而且该可以很大程度的减小网络的计算量，提升模型的推理速度。

在图像分类任务上，我们使用全局平均池化对其在 w 和 h 上做全局池化，最后使用全连接层输出所属类别的概率。对于其他任务，如目标检测任务，可以将任何阶段产生的特征用于特定任务的头部特征，在第四章人体检测和第五章人体行为识别中会详细介绍。

总体来看，我们的网络模型更加灵活，可以针对不同的任务调整超参数大小，让网络适应该任务的需求，是一种能适用于大多数计算机视觉任务的网络结构。

3.3 网络 op 融合

为了提升网络的推理速度，模型压缩和加速在目前成为广大学者研究的热点，由于某些设备计算能力弱，内存小，然而通常神经网络参数量巨大，想要让有数十亿参数量的大网络模型运行在这些设备上，显然不太现实。所以当前需要对模型进行压缩加速，本节会对常用的压缩加速方法做简单的介绍，同时详细介绍网络 op 融合的方法。

模型压缩和加速在目前成为广大学者研究的热点，由于某些设备计算能力弱，内存小，然而通常神经网络参数量巨大，想要让有数十亿参数量的大网络模型运行在这些设备上，显然不太现实。目前该研究方向主要有以下几种方法对模型进行压缩和加速，如图 3-6 所示：

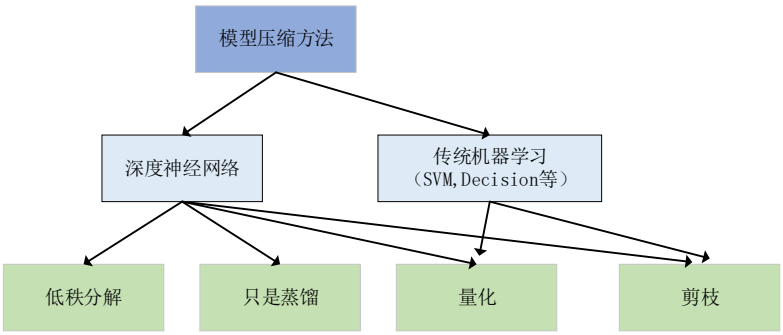


图 3-6 模型压缩方法

低秩分解的方式是去除模型的冗余信息和减小权值参数。设 W 是一个 n 行 m 列的数值矩阵，如果 W 的秩远小于 min(n,m)r,则称该矩阵为低秩矩阵。由于低秩矩阵的特点是每行或每列都可以用其他的行或列线性表出，所以其包含大量的冗余信息，而减小或消除冗余信息并不会让该特征失去表达性。

知识蒸馏方法是通过将大模型作为“老师网络”，小模型作为“学生网络”，然后利用“老师网络”监督“学生网络的“学习，从而提升小网络的性能，该方法需要优秀的”老师网络“，同时对”教学方法“要求严格，且训练更耗时。

量化方法主要就是在存储参数权重和计算时使用更小位数的类型，这样既能够减小模型的内存消耗，而且对于 8bit 整型而言硬件会对其做计算优化，计算复杂度与所需存储空间比浮点型 32bit 更小。一般而言，神经网络模型的参数都是用的 32bit 长度的浮点型数表示，实际上权重不需要保留那么高的精度，可以通过量化，比如用 0~255 表示原来 32 个 bit 所表示的精度，通过牺牲精度来降低每一个权值所需要占用的空间。其中最具代表性的两种量化方式为：一种是训练时量化，如谷歌提出的在训练中插入伪量化节点的方法；还有一种时训练后量化，如 ncnn 的量化方法等。该方法适合某些大网络模型的加速，对大多数小模型则会丢失太多精度。

剪枝方法顾名思义就是对已经训练完成的模型做模型裁剪，主要的裁剪方式有通道裁剪、层裁剪等，裁剪完成后，最后对模型做微调到原来的准确率。主要流程为：训练时通过层输出观察网络模型中该层是否占据重要的作用，进而对某些层进行删除的操作。由于大模型中存在大量的冗余操作，该方法可以很程度上的简化大模型，但在对小模型的作用却不突出，往往会丢失较大的精度，得不偿失。

通常以上几种模型压缩和加速方法或多或少都会有精度上的损失，特别是对轻量级的神经网络进行压缩和加速时精度损失明显。而本文使用 RepVGG[51]中使用的网络 op 融合的方法对模型的结构进行简化，该方法不仅可以对模型进行加速，而且不会损失任何精度。

网络 op 融合是一种在推理阶段简化模型结构的方法，前面我们已经论述了多分支结构的网络可以让模型训练更好的性能，但推理阶段我们希望模型的推理速度越快越好，所以需要用到网络 op 融合方法将模型结构进行简化。图 3-7 展示了模型推理阶段的重参数化过程，其中主要包括 Bn 层和卷积层的 op 融合以及不同卷积层之间的 op 融合。整个重参数化流程主要包括三个步骤。

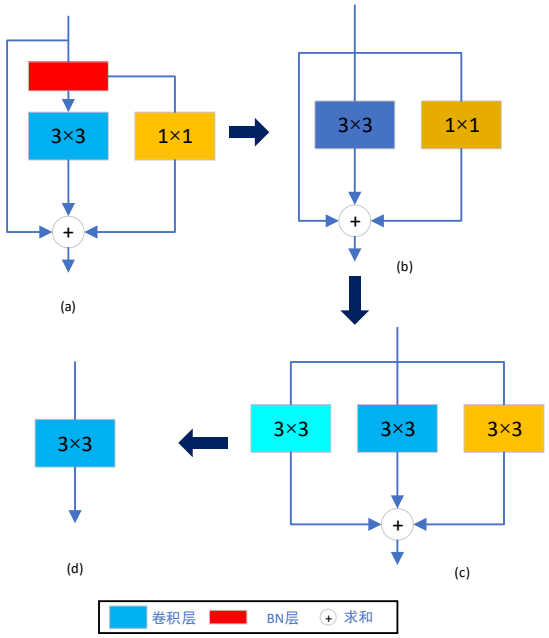


图 3-7 网络 op 融合步骤

这里用$w^3 \in R^{c_2*c_1*3*3}$ 表示 3×3 卷积的参数，c1 和 c2 分别表示输入、输出通道，其中 Wi 表示转换前的卷积层参数，μ^3、σ^3、γ^3、β^3 分别表示 3×3 卷积层后的 BN 层的均值、方差、尺度因子和偏移因子。我们令输入为$M^1 \in R^{N*C_1*H_1*W_1}$，$M^2 \in R^{N*C_2*H_2*W_2}$。

1、首先通过式 3 将残差块中的卷积层和 BN 层进行融合，将图中 a 中转化为 b 的样子，该操作在很多深度学习框架的推理阶段都会执行。

BN 函数的公式如下：

$$b_n(M,\mu,\sigma,\gamma,\beta)_{:,i,:,:} = (M_{:,i,:,:} - \mu_i)\frac{\gamma_i}{\sigma_i} \quad \forall\, 1 \le i \le C_2 \qquad (3-1)$$

将每个 BN 及其之前的 conv 层转换为具有偏差向量的 conv3×3 的卷积和 1×1 的卷积转换方式一样）。设$\{W`, b`\}$是从$(M,\mu,\sigma,\gamma,\beta)$转换而来的核和偏差，我们有

$$W`_{i,:,:,:} = \frac{\gamma_i}{\sigma_i}W_{i,:,:,:}, b`_i = -\frac{\mu_i\gamma_i}{\sigma_i} + \beta_i \qquad (3-2)$$

$$b_n(M*W,\mu,\sigma,\gamma,\beta)_{:,i,:,:} = (\mathrm{M}*\mathrm{W}^{`})_{i,:,:,:} + b_i^{`} \quad \forall 1 \le i \le C_2 \tag{3-3}$$

2、我们将融合后的 1×1 卷积和 identity 操作转化为 3×3 卷积。由于整个残差块中可能包含 1×1 卷积分支和 Identity 两种分支，对于 1×1 卷积分支而言，整个转换过程是先将 1×1 卷积核中的数值移动到 3×3 卷积核的中心点，然后直接用 3×3 卷积核替换 1×1 的卷积核；而对于 Identity 分支而言，由于该分支并没有改变输入的特征映射的数值，在这里直接设置一个 3×3 卷积核，然后将所有的卷积参数权重置为 1，那么它与输入的特征映射相乘之后，保持了原来的数值。

3、最后融合残差模块中各分支的 3×3 卷积，即将所有分支的权重 W 和偏置 B 叠加起来，从而该残差模块中只含一个 3×3 卷积。

综上所述，我们的网路 op 融合模块可以用与以下公式表示：当 $C_1 = C_2, H_1 = H_2, W_1$，则有公式$(3-4)$：

$$\mathrm{M}^2 = b_n(\mathrm{M}^1 * w^3, \mu^3, \sigma^3, \gamma^3, \beta^3) + b_n(\mathrm{M}^1 * w^1, \mu^1, \sigma^1, \gamma^1, \beta^1) + b_n(\mathrm{M}^1, \mu^0, \sigma^0, \gamma^0, \beta^0) \tag{3-4}$$

否则有公式$(3-5)$：

$$\mathrm{M}^2 = b_n(\mathrm{M}^1 * w^3, \mu^3, \sigma^3, \gamma^3, \beta^3) + b_n(\mathrm{M}^1 * w^1, \mu^1, \sigma^1, \gamma^1, \beta^1) \tag{3-5}$$

总之，网络 op 的融合可以减小模型的复杂度，在推理时，融合 Block 中的所有分支，最终可以将网络看成一个全 3×3 卷积的网络。在 1080TI 的平台下各项性能指标测试对比结果如表 3-6、表 3-7 所示：

表 3-6 网络 op 融合后的网络各项性能指标

网络	内存消耗(M)	参数量(M)	推理速度(ms)	FLOPs(B)
LimitNet-A	939	7.2	4.1	0.3
LimitNet-B	1023	50.2	5.4	2.5

表 3-7 原始网络的各项性能指标

网络	内存消耗(M)	参数量(M)	推理速度(ms)	FLOPs(B)
LimitNet-A	949	7.9	12.1	0.33
LimitNet-B	1035	55.1	15.1	2.8

上述结果表明网络经过网络加速后，前向推理速度大幅度下降，内存消耗、参数量以及 FLOPs 有小幅下降。由于只是该模型压缩方法的是对网络操作的参数的融合，所以该方法对精度没有任何影响，这就很符合本章开头对该网络轻量级网络的设定要求。

3.4 实验结果与分析

3.4.1 实验环境

本次实验基于 PyTorch 深度学习框架，在 1080TI 24GB 的显存下进行的，具体硬件说明如表 3-8 所示：

表 3-8 训练测试环境说明

系统版本	Ubuntu 16.04
CPU 版本	InteL(R) Xeon(R) CPU E5-2690 v2 @ 3.00GHZ，4 核,8 线程，32G 内存
GPU 及显存	1080TI 12GB 内存
训练框架	PyTorch1.3
其他软件版本	CUDA10.0 ，Cudnn7.6

3.4.2 训练细节

本章设计的网络模型的训练方法的具体流程如图 3-8 所示：

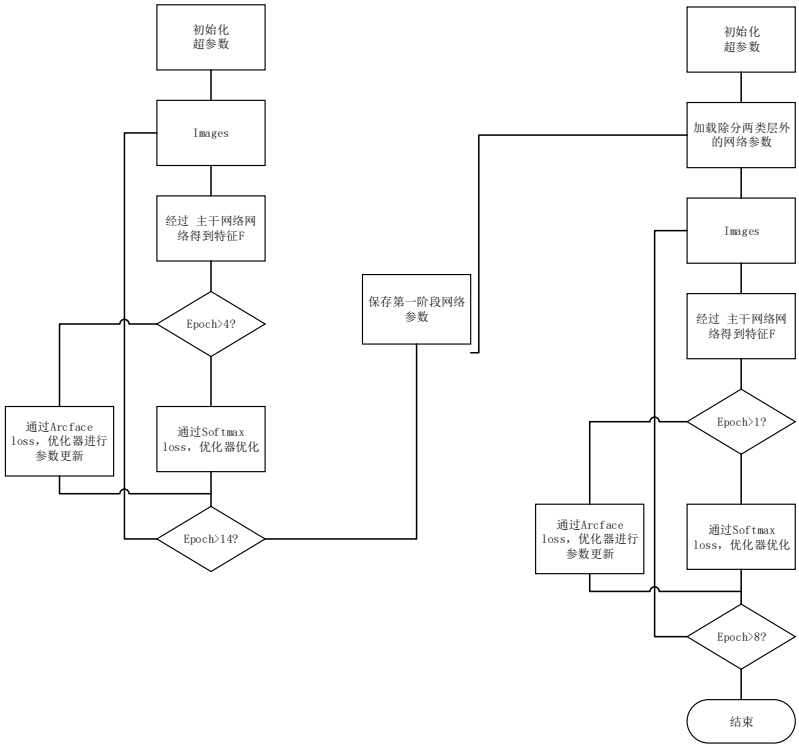


图 3-8 训练流程图

模型的训练包括两大步骤，第一步是先用 msra_celeb 训练集做第一阶段的训

练，该阶段总共训练 14 个 epoch。为了提升模型的训练效果，加速模型快速收敛，在训练中先用 Softmax 损失函数训练 2 个 epoch，然后用 ArcFace loss 进行训练。学习率的设置从 0.1，每 4 个 epoch 衰减 0.1。第二阶段中先加载第一阶段除分类层外的权重参数，然后用 msra_celeb_id40 数据集做 finetune（微调）训练。总共训练 8 个 epoch，和第一阶段的训练方式一样，先用 Softmax 训练一个 epoch，然后在用 ArcFace 训练，初始学习率设置为 0.01，每 4 个 epoch 衰减 0.1。两阶段的训练方法使得模型更能够保证模型的泛化能力，能够有效提升模型在各种测试集下的准确率。

3.4.3 损失函数

本章的分类损失函数共有两种，一种是 Softmax loss，另一种是基于 Softmax 改进的损失 ArcFace 损失。Softmax 的优点是训练收敛快，而 ArcFace Loss 的缺点则是收敛慢，但 ArcFace Loss 在训练时对模型的性能有明显的提升。基于以上两个原因，在训练早期会先使用 Softmax 训练，然后再使用 ArcFace Loss 训练。下面对这两种损失函数进行详细的介绍。

3.4.3.1 Softmax Loss

Softmax 损失函数是在计算机视觉任务特别实图像分类任务上被广泛使用，是 max 函数的改进版本；其计算过程如公式（3-6）与公式（3-7）所示：

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{N} e^{z_k}} \tag{3-6}$$

$$L_i = -log\left(\frac{e^{f_{y_i}}}{\sum e^{j}}\right) \tag{3-7}$$

首先第一步通过公式（3-6）计算概率值，z_j是输入，e^{z_j}的作用是增强该损失函数的训练，然后在进行归一化处理就得到属于各分类的概率值，改值表示输入的图像被分到各个类别的概率，然后进行公式（3-7）的操作计算交叉熵损失函数。

其中y_i=1 时，表示第 i 类是属于它的真实的类别，log 里对应的值就是正确分类的 Softmax 值，它的占比越大，这个网络的 loss 就越小，反之占比越小损失越大。

该损失函数的优点是训练收敛快，能够一定程度上区分样本，但对于较难分辨、不同类别相似性高的实例识别效果很差。

3.4.3.2 ArcFace Loss

ArcFace loss 是 Additive Angular Margin Loss（加性角度间隔损失函数）的缩写，是基于 Softmax loss 改进的损失函数，计算公式如下：

$$L = \frac{1}{N}\sum_{i} -log\frac{e^{s(cons(\theta_{y_i,i}+m))}}{e^{s(cons(\theta_{y_i,i}+m))} + \sum_{j\neq y_i} e^{s(cons(\theta_{y_i,i}))}} \qquad (3-8)$$

从公式可以看出该损失函数对输入做了权重的归一化，使用角度的大小来生成分类概率，同时对 θ 加上角度间隔 m，同时为了提升反向传播的效率使用了超参数 s，s 的值一般设置为 16 或者 32 等。该损失函数的优点是：性能高，能够提升模型对难样本的区分度，但该损失在训练中收敛较慢，所以通常配合 Softmax 损失进行训练。

3.4.4 实验结果对比分析

本节对几种性能较好、较常用的轻量级网络模型在不同数据集下的各项性能指标进行对比。训练和测试的输入图像尺寸都是112×96大小的分辨率。首先我们对除了准确率外的其他性能指标进行对比分析，测试结果如表 3-9 所示：

表 3-9 网络性能测试对比

网络	内存消耗(M)	参数量(M)	推理速度(ms)	FLOPs(B)
ResNet-18	1531	11.2	8.7	1.5
MobilleNet-v1	703	1.0	10.1	1.9
EfficientNet-B0	699	0.2	20.1	0.3
ShuffleNet-v1	937	1.1	13.9	0.12
LimitNet-A	939	7.2	4.1	0.3
LimitNet-B	1023	50.2	5.4	2.5
ResNet34	1591	21.3	12.1	3.0
ResNet-50	1695	23.7	16.7	3.4

从以上的测试结果可以发现，LimitNet 网络的推理速度最快，FLOPs 仅比 ShuffleNet 更大，这展现了 LimitNet 网络模型的优势；在参数量的对比上，LimitNet 的参数量比 MobilleNet 与 ShuffleNet 网路更多，因为它们采用了可分离卷积和组卷积的方式，这很大程度的减小了参数量，但我们对网络进行了分支融合和 op 融合，这提升了推理速度、减小了 FLOPs。在内存消耗对比上，分析与 ResNet 网络的对比，我们发现 LimitNet 的内存消耗更小，这是因为 ResNet 网络有多分支结构导致内存消耗大。以上两点都符合我前面的结论：1）多分支结构的模型会增大内存消耗和减小推理速度 Flops；2）FLOPs、参数量、内存消耗和模型推理速度并不

完全成正比。综上所述，我们网络在这几项性能评价指标的整体性能上要比其他几种轻量级网络更好。下面我们对各种轻量级网络的在各测试数据集上的准确做详细的对比分析。

3.4.4.1 LFW 数据集测试结果分析

LFW 数据集是当前人脸识别测试精度的主要数据集之一，该数据集由从从各种数据库中随机选取的 6000 对人脸图像组成，其中 3000 对人脸中每一对是属于同一个人，另外 3000 对人脸则属于不同人。在模型上的测试精度的主要流程是：首先通过 LFW 数据集给出一对照片，然后通过模型验证是否该对图像属于同一个人（一般根据特征相似率来判断是否属于同一个人），与正确结果进行比较，最后将所有人脸对测试完成即可得到该数据集上的测试准确率。各个网络模型在该数据集中的准确率如表 3-10 所示：

表 3-10 网络在 LFW 数据集下的测试精度

Model	准确率
ResNet-18	99.1
MobilleNet-v1	99.08
Shufflenet-v1	89.31
ResNet-34	99.30
ResNet-50	99.52
EfficientNet-B0	88.91
LimitNet-A	99.24
LimitNet-B	99.54

以上结果表明我们设计的 LimitNet-A 网络在 LFW 数据集上能够达到最高的准确率，比 ResNet-18 高 0.02，我们也测试了大模型的对比，可以看出我们模型的和 ResNet50 模型的准确率相当，但在其他方面的性能更突出。

3.4.4.2 动态人脸测试结果分析

该数据集为真实的人脸动态抓拍图像，包含 60 类，总计 22055 张测试图像。该测试数据集的测试方法和 LFW 数据集不一样，测试前先从每一个人的图像中选择其中一张入库（将图像放到 30 万张干扰数据库中），测试过程是从数据集中选择一张图像通过模型提取特征，然后选择库中与其最相似的 K(K=1,5,10)张图像的特征，判断 k 张图像中是否有和测试图像的 id 相同的图像。通过对所有测试数据的测试结果可以得到动态人脸测试的 Top-k 精度。各个模型的测试准确率如表 3-11

所示：

表 3-11 网络在动态人脸数据集下的测试精度

准确度	Top-1	Top-5	Top-10
ResNet-18	88.68	92.05	92.5
MobilleNet-v1	87.2	91.41	92.3
Shufflenet-v1	87.7	91.50	92.1
ResNet-34	90.31	93.20	93.6
ResNet-50	92.68	94.34	95.1
GhostNet	89.12	92.21	92.8
EfficientNet-B0	88.4	90.72	91.9
LimitNet-A	90.20	92.4	93.7
LimitNet-B	93.10	94.5	95.7

上述结果表明 LimitNet 网络在同量级的网络对比中准确率更好， LimitNet-A 比 ResNet-18 高 1.6%，比 MobileNet-v1 网络高 3%。同时也测试了中轻量级网络 LimitNet-B 的测试结果，可以看出 LimitNet-B 模型的和 ResNet-50 模型的准确率相当，但在其他方面的性能更好，速度比其快约 2ms。

3.4.4.3 室外抓拍数据集测试结果分析

该数据集包含 40 类，总计 27139 张测试图像，是由室外摄像头采集而来。其中部分人脸样本存在遮挡、偏角等现象。该测试数据集的测试方法和动态人脸数据集一样。通过对所有测试数据的测试结果可以得到该人脸测试的 Top-k 精度。各个模型的测试准确率如表 3-12 所示：

表 3-12 网络在室外数据集下的测试精度

准确度	Top-1	Top-5	Top-10
ResNet-18	93.65	95.06	97.1
MobilleNet-v1	92.68	94.55	96.3
Shufflenet-v1	92.08	94.09	96.3
ResNet-34	94.75	96.85	98.1
ResNet-50	96.68	97.85	98.7
EfficientNet-B0	92.04	94.19	95.97
LimitNet-A	93.90	96.5	98.3
LimitNet-B	96.98	98.1	98.8

同样，LimitNet 网络在该数据集上能够达到最高的准确率，但可以看出和其他

几种网络的准确率差距不大，这是因为该数据集的难度识别不高。这也体现了构建一种高质量的测试数据集对体现网络模型的性能指标有着非常重要的作用。

3.4.4.4 测试结果总结

以上实验表明，本章设计的 LimitNet 网络的总体性能在测试数据集上表现更好。相比于其他几种轻量级网络，LimitNet 虽然在内存占用和参数量没有特别突出，但模型的推理速度最快，且能够在各种类型的测试数据集中都表型出比较好的准确率。其中更小更轻量级的 LimitNet-A 网络，在轻量级网络中的性能最突出。LimitNet-B 网络可以和 ResNet-50 等中量级网络的准确率相媲美，且推理速度更快。以上测试结果表明 LimitNet 网络模型通用性更强，具有较强的鲁棒性。

3.5 本章小结

本章在人体检测与行为识别领域的研究中，提出一种更轻量级、更具鲁棒性的轻量级网络 LimitNet。在网络结构设计阶段，首先验证了各种卷积操作的对比试验，基于实验结论提出一种更好的网络 Block。然后基于该 Block 设计实现了一种轻量级的特征提取网络 LimitNet，并通过网络 op 融合的方法进行相邻层间融合、将多分支网络简化为单分支网络的结构。在不损失任何精度的情况下，能够显著提升多分支网络结构模型的推理速度。为了验证该网络的性能，构建了复杂的训练集和测试集，该网络在几种测试集上的准确率比当前流行的轻量级网络模型的更高，且内存消耗相对比较小、推理速度最快，具有较强的鲁棒性。

第四章 轻量级人体检测算法的研究

人体检测是当前机器学习领域研究的热点，人体检测技术的研究在虚拟现实、自动驾驶、人数统计等领域有着非常重要的意义。本文第二章中简要介绍了当前目标检测的思路，但当前大多数算法往往在精度和效率这两方面顾此失彼，要么精度高效率差，要么精度低实时性好，使得该技术在现实应用时受到一定约束。人体检测是目标检测领域下的子任务，不仅面临着和目标检测相同的问题（多尺度、复杂背景等），同时还有在实际场景下人体存在姿态变化、相互遮挡、运动模糊等问题，在解决一些实际场景问题上还略显乏力。

本章首先简要概括了人体检测算法的现有的主要方法，然后详细分析了当前人体检测技术的难点问题。接下来基于 CenterNet[14]检测算法改进实现人体检测的任务，针对多尺度目标的难点，设计一种特征融合模块融合低级语义信息和高级语义信息，并减少细节信息在采样时丢失的问题，从而提高算法在多尺度目标下的精度。在模型的训练中，提出了一种多阶段训练方法训练该检测网络，进一步提升模型训练效果。下面将对人体检测算法的详细设计和具体细节进行详细的介绍。

4.1 人体检测技术难点

在过去几十年中，随着深度学习的发展，国内外学者们在人体检测领域上的研究上取得了非常大的成果。目前为止，在少遮挡、背景比较单一的简单场景中能够取得非常不错的效果，在 COCO 数据集上 AP 最高能够达到 50%以上，有的算法已经可以进行产业化和工业化的落地。在该领域，目前常用的都是基于深度学习的方法，包括 One-Stage 检测框架，代表的有 YOLO 系列，SSD 等，Two-Stage 检测算法最具代表性的算法有 Faster-RCNN 等，这些方法在一些公开简易数据集上，例如 voc 数据集、COCO 数据集都取得了不错的效果。但是在比较复杂的场景下，如图 4-1，例如地铁站，大型的商场等这类遮挡严重，背景复杂的场景下，往往达不到好的效果。人体的姿势不同、不同程度的相互遮挡、光照等因素都会给人体检测技术带来困难，下面对当前针对这些难点问题提出的研究方法进行简要概述。

图 4-1 难样本展示

目前针对以上人体检测的难点，许多研究学者也都提出了一些解决方案。例如针对人体多姿势的问题，通过增大数据集多样性来解决，如训练数据集需尽量包含尽可能多的人体姿势。针对人体背景复杂的问题，许多学者设计了更加优秀的特征提取网络，如 EfficientNet、ResNest 网络等，以及用图像增强的方式去解决复杂背景的难点。在多尺度目标的问题上，YOLOv3、RetinaNet、MTCNN 分别提出了特征融合、特征多尺度、输入多尺度等方式去检测各种大小的目标框的问题，同时优化 NMS 方法去解决候选框去重的问题。然而对于这些解决方法，大多数都是用速度换取准确率的提升，从而都不可避免增加算法的复杂度与推理时延，导致许多模型都不能应用到实际的视频监控等领域。所以说人体检测任务仍然面临很大的挑战，如何设计一个鲁棒性很高的网络模型是人体检测目前需要解决的问题。

本文基于 Anchor-Free 的检测框架，设计复杂度较低的卷积神经网络，不仅提高了算法的精度，同时降低了算法的推理时延，在算法的延时和精度之间达到一个很好的平衡。由于基于 Anchor-Free 的检测方法 CenterNet[14]与 R-CNN 相比不需要区域建议网络以及 ROI 等重要组件，与 YOLO 和 SSD 相比，CenterNet 不需要预先设定 Anchor 的大小，并且 CenterNet[14] 在推理阶段不需要 NMS（Non-MaximumSuppression）[54]也能达到很高的候选框的去重效果，所以在速度上得到了很大的提升。为了能够保持检测的速度和准确率的平衡，本文改进 CenterNet 提出 Refine-CenterNet 算法对人体进行检测。

4.2 数据集的准备

为了更好的验证人体检测算法在实际场景下的检测性能，本次实验所用的数据集都具有场景较复杂、人体较密集的特点，同时每一张图片中都会有不同尺寸大小的目标图像，且有大量相互遮挡的难样本。下面对数据集进行详细的介绍。

4.2.1 教室学生数据集

该数据集是实验室标注的教室内场景的图像，大部分训练集中每张图像都包含几十张人体样例，图片中的人体存在大量的桌子遮挡、人体相互遮挡、自我遮挡的情况，如图 4-2 所示：

图 4-2 示例图像

该数据集总共有 1000 张，包括 800 张训练数据以及 80 张测试数据。其中训练集中共计 35000 个目标示例，测试集中共有 4006 个目标实例。数据集中的样例尺寸大小分布如表 4-1 所示。

表 4-1 样本大小分布

尺寸（像素面积）	训练集实例数量	测试集实例数量
256	802	114
256-1024	17499	1987
1024 以上	16699	1905

4.2.2 PANDA 数据集

PANDA 数据集（gigaPixel-level humAN centric video Dataset）是首个在国际上公开的大场景多目标的动态数据库，该数据集是由清华大学团队参与构建。目前该数据集中共包含来自 21 个不同场景的 555 张图像样本，其中每一张图像分辨率接近 10 亿个像素，一张图像中可包含多达千个目标，场景平均覆盖平方千米级范围，百米外人脸清晰可识别，其中在原始训练集中包括 390 张静态图片样本，测试集包括 165 张测图像样本。在本文的研究中，我们将每张图像分割为多张样本图像，并对样本进行筛选，最后数据集中包含 10000 张训练集以及 2000 张测试集。样例图如图 4-3 所示，数据集中的尺寸大小分布如表 4-2 所示：

图 4-3 示例图像

表 4-2 样本大小分布

尺寸（像素面积）	训练集实例数量	测试集实例数量
256	2788	310
256-1024	8335	1200
1024 以上	41320	10234

4.3 基于 CenterNet 的人体检测网络详细设计

在基于 Anchor 的 One-Stage 和 Two-Stage 的检测方法中，在训练前通常需要通过聚类算法获取训练数据集中 Anchor 的大小分布信息，然后设置 Anchor 大小的超参数。而且这种 Anchor-Base 的检测方法通常在推理阶段还需要通过 NMS 方法（非极大值抑制）删除重复框，该方法的原理是通过计算 Bbox 间的 IOU（即两个候选框的重叠度）从而达到删除同个目标的重复候选框的目的。同时为了检测多尺寸的人体大小，需要增加 FPN 特征金字塔模块，这增大了前向推理的计算复杂度。目前基于 Anchor 的人体检测算法在业界都能够达到很高的精度，但算法的推理速度往往很受限制。而基于 Anchor-Free 的检测方法结构更简单，可以不用太多的数据后处理操作就能达到很高的性能。本章主要基于 Anchor-Free 的 CenterNet[14] 检测方法来实现人体的检测，该方法可以不使用 NMS 和先验框就能达到去除重复框的目的，能够很大程度的简化整个检测框架，同时设计特征融合模块对低层特征和高层特征进行融合提升算法在多尺寸目标下的精度，最终该算法保证了速度和

精度的平衡。

4.3.1 基于 CenterNet 的改进检测算法框架概述

首先介绍基于 Anchor-Free 的原始 CenterNet 检测方法的整体结构，如图 4-4 所示：

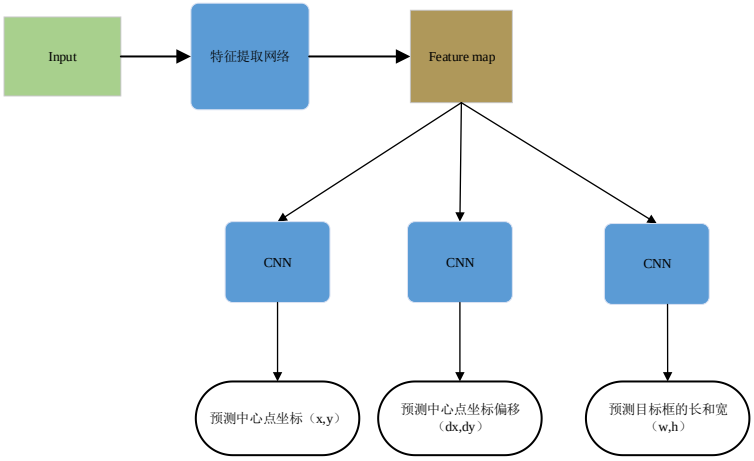


图 4-4 CenterNet 检测架构图

该检测框架的思想是利用编解码形式的网络生成 Feature map，然后通过三个分支分别生成 heatmap、中心点的偏移和中心点的长与宽，最后直接通过操作 heatmap 就可以得到人体的位置和大小信息。其中 heatmap 在姿态识别领域已经成为当前最有效的解决姿态关节点识别的方法，通常比直接回归关节点坐标更有效。

三个预测分支的综合结果可以获得最终人体目标框的位置和大小，从以上描述的检测框架可以发现 CenterNet 检测框架的几个优势：

1、该检测框架检测流程简单，并且各模块低耦合，可以进行各个模块的单独优化。

2、由于 heatmap 图分辨率比较大，每一个网络可以产生一个心中点，所对应的中心点能够包含大小目标，所以不需要多尺度预测和特征金字塔就能够很好的检测大小目标，一定程度上减小了模型的复杂度。

3、由于每一个目标框只会预测一个中心点，所以在卷积网络后可以不需要添加 NMS，进一步减小了模型的复杂度，提升了模型的推理速度。

本章在 CenterNet 检测方法上提出了一种更轻量化的特征提取网络，并设计了一种特征融合模块提高算法在多尺度目标下的精度，同时使用多阶段训练方法增强提升模型的训练效果。改进的 Refine_CenterNet 检测框架如图 4-5 所示：

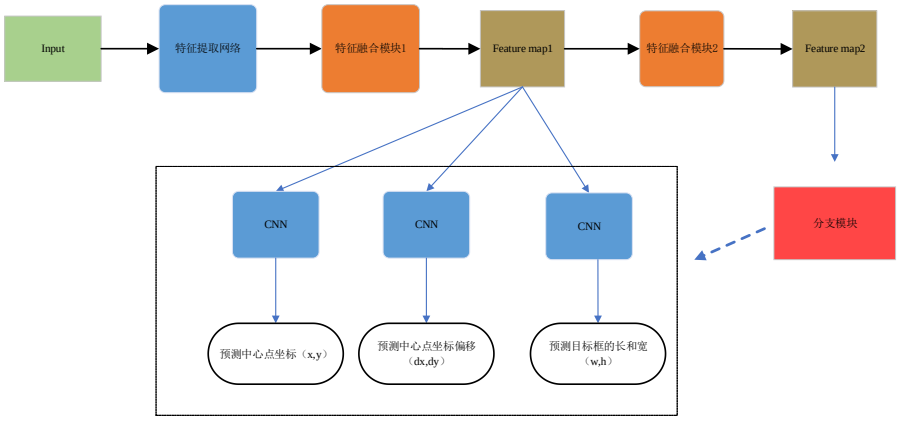


图 4-5 Refine_CenterNet 检测架构图

在图 4-5 中，我们令输入 RGB 图像为$I \in R^{w*h*3}$，其中宽是 w、高是 h、深度是 3，将I输入到头部特征提取网络中生成头部特征$Head \in R^{\frac{w}{l}*\frac{h}{l}*c}$，其中l是下采样倍数，然后使用特征融合模块对 Head 特征进行上采样 N 倍，最终生成 $Feature\ map \in R^{\frac{w}{k}*\frac{h}{k}*c}$，其中 k = l/N，最后对 Feature map 分别进行三个分支的预测。与原始检测框架 CenterNet 主要的差异有以下两点：1）设计了一种特征融合的方法将低层和高层的特征进行融合，该模块可以提升对小目标检测的精度；2）充分利用了小尺寸的 Feature map，第一阶段先对下采样 8 倍的 Feature map1 做分支预测，第二阶段则是对下采样 4 倍的 Feature map2 做分支预测，这种训练方法不会增加网络前向推理的复杂度，但通过这种阶段性的训练能够提升训练效果。下面对每一个模块的设计进行详细的介绍。

4.3.2 头部特征提取模块

在检测算法中，头部特征提取网络通常占据着前向推理的大部分时间，同时也是算法性能优劣的关键。在第三章已经验证了轻量化的 LimitNet 网络结构在分类任务的高效性与鲁棒性，本章检测算法中将 LimitNet-A 网络作为检测框架中的头部特征提取模块，其中将 LimitNet-A 中第 5 阶段的输出作为人体检测的头部特征。但与第三章中的 LimitNet-B 差异（如图 4-6 所示，左图是标准 LimitNet-A，右图是人体检测的头部特征提取网络）在于我们在第五阶段将 stride 设置为 1，使其输出的头部特征保持和第四阶段一样的尺寸大小，最终头部特征网络将输出 I 下采样了 16 倍。这么的设计的原因是在 LimitNet-A 网络的第五阶段中只有一层卷积，所以参数量没有明显增加，同时获得更高分辨率的特征表示，减小了上采样模块的计算量。

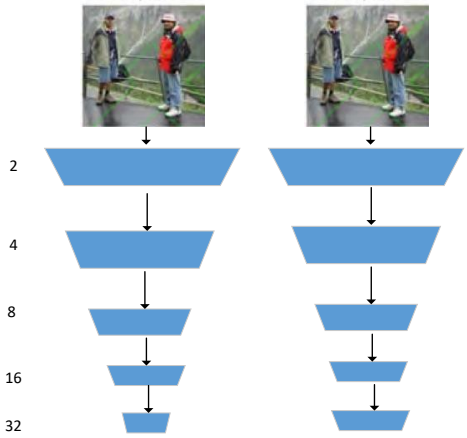


图 4-6 头部特征提取网络

4.3.3 特征融合模块

由于该检测框架是基于 Anchor-Free 的方法，所以最终只需要用一个大尺寸的 Feature 去预测目标框的位置以及大小，在这里我们没有用到 RetinaNet 等方法中通过设计 FPN 多特征的方式去提升人体检测在多尺寸目标下的准确率。因为从本质上看，FPN 是一种分而治之的方式，解决多尺度预测的问题，不同尺度相当于不同的焦距，从而聚焦于不同大小的物体上。而 CenterNet 的原理则相反，是采用合而治之的方式，其关键流程是首先获取到大尺寸的 Feature Map，然后通过该 Feature Map 直接回归坐标。由于特征中同时包含了小物体和大物体的信息，所以只需要一种特征尺寸即可完成对所有尺寸的目标的回归。而且使用 FPN 必然会引进来尺度分配的问题,还会增加了网络的计算量，让整体框架复杂化，增加检测算法推理时延与训练难度。

在检测任务中，深层的特征由于下采样倍数太大（通常 32 倍），导致小目标在深层特征中占据不到一个像素点，因此对图像中小目标的检测效果通常不会太好。为了解决这个问题，本章设计了一种特征融合的方法将浅层特征和深层特征融合，可以在不显著增大模型推理时延的情况下提高人体多尺寸目标的检测效果。

在上采样路径上，BiSeNet[52]中指出在特征表示的层面上，低层特征和高层的特征表示不同，仅仅以通道来连接低层和高层特征，则就会带来很多噪音，所以我们设计了一个基于注意力机制特征融合模块（AFFM）来融合低层丰富的空间信息和高层的语义信息，如图 4-7 所示。

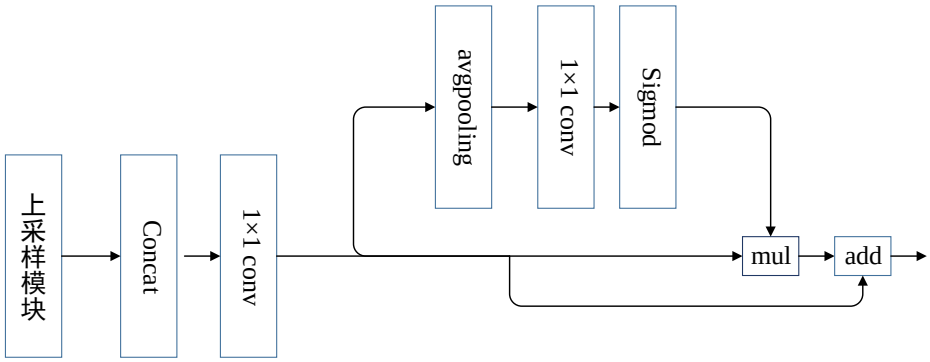


图 4-7 AFFM 特征融合模块

在 AFFM 特征融合模块中，首先将深层特征通过上采样模块生成和浅层特征相同尺寸的特征图，然后与浅层特征图在通道维度进行 Concat。然后经过 1×1 卷积将其转换为指定通道数的输出特征图 F（该通道数一般和高层的特征图的通道数相同），经过类似于 SENet[53]中注意力机制模块将特征图转化为 he 特征图 F 的通道数相同长度的特征向量。将该特征向量与特征图对应通道进行加权，再和对先前的特征 F 对应通道的值相加，即可得到最后的融合特征图。这种融合方法可有效减少特征融合的噪声，增强融合特征的表示能力。AFFM 模块可以有效提升小目标的检测性能，同时整个模型的计算量并没有增加太多。

4.3.4 DCN 模块

在计算机视觉任务中神经网络的卷积方式通常是固定的几何结构，这导致某些任务的特征提取受到限制，如何有效地提取到较难的特征（包括人体姿势变化、遮挡等）一直以来都是研究者们面临的挑战。为了得到更好的 Feature Map，我们使用 DCN（Deformable convolution）模块来自适应地学习感受野，与普通的卷积是规则的形状不同，DCN 的感受野的形状是可变化的，更容易得到了任意形状融合的特征信息。DCN 模块的结构如图 4-8 所示：

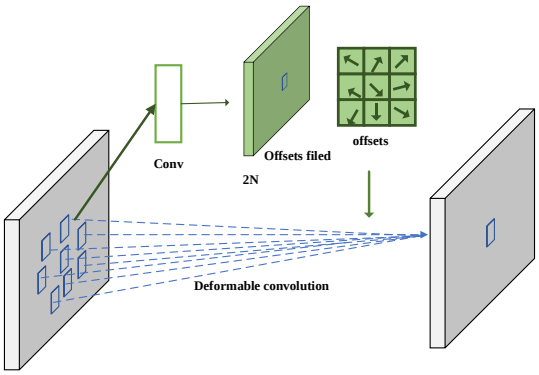


图 4-8 DCN 模块

其流程是首先由输入的特征映射得到 offset 卷积在各方向上的偏移，该 offset 的尺寸大小的原始输入一致，通道数是原始输入的 2 倍，表示在 x 和 y 方向上的偏移。然后基于原始输入特征以及 offset 通过可形变卷积获得输出的特征映射。

普通卷积操作可以用以下公式进行描述：

$$y(p_0) = \sum_{p_n \in R} w(p_n).x(p_0 + p_n) \qquad (4-1)$$

而 DCN 的公式则是：

$$y(p_0) = \sum_{p_n \in R} w(p_n).x(p_0 + p_n + \triangle p_n) \qquad (4-2)$$

区别在于 DCN 多了$\triangle p_n$，而该$\triangle p_n$就是通过图中上面一个分支生成的，表示 x 或者 y 的在空间的偏移量，这有助于目标检测生成更具表征能力的 Feature map，从而提升模型的准确率。

4.3.5 多阶段分支预测模块

与 CenterNet 检测方法只使用下采样 4 倍的 Feature map 作为分支预测不同，本文在训练将下采样 8 倍的 Feature map 也作为分支预测，联合下采样 4 倍的 Feature map 对网络优化，如图 4-5 中所示。在下采样 8 倍的分支预测模块的细节如下所述。

通过头部特征网络和第一个特征融合模块得到 Feature map 后，会分别进行三个分支预测。第一个分支预测目标框中心点的坐标（x,y），该分支具体的网络结构如图 4-9 所示。首先通过 3×3 的卷积操作，然后通过 1×1 的卷积操作得到和类别相同的通道数，因为只有一个类别，所以该通道数为 1。最后得到 64×64×1（第二阶段的分辨率加倍）的输出 heatmap，然后经过 3×3 的 max pooling 层就能得到把峰值的坐标点转化为目标框中心点的坐标。

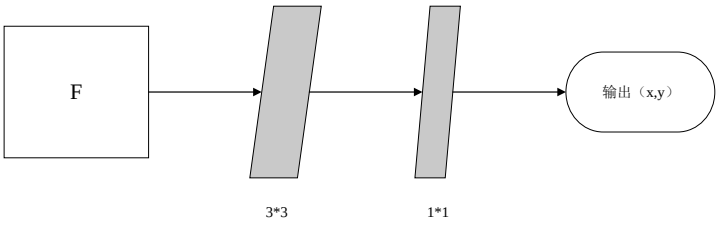


图 4-9 分支模块结构

由于网络在下采样后再还原到原始图像中的坐标会有一定的位置偏移，在第二个分支我们对该中心点坐标对应的偏移进行回归。Feature map 通过两层卷积后

得到 64×64×2 的输出，表示每一个中心点对应的 x 方向和 y 方向的偏移量。

第三个分支则是预测目标框的长和宽。和偏移量预测分支类似，通过两层卷积后直接得到 64×64×2 的输出，代表每一个中心点对应的长和宽。

最后整合三个分支的输出就能得到最后目标框的位置和大小信息，该阶段只是人体检测粗检阶段，由于 heatmap 尺寸较小，准确率可能达不到最好的性能，但针对大目标有较好的检测效果。

在下采样 4 倍的预测模块的结构与下采样 8 倍的预测模块完全相同，唯一区别是 Feature map 尺寸要大一倍为 128×128，导致预测分支中 heatmap 和对应的输出偏移预测矩阵、长宽矩阵的的尺寸要大一倍。由于先对下采样 8 倍的模块及逆行训练，所以对于下采样 4 倍的模块的训练收敛效果会更快、更好，这对于模型的提升训练效果很有帮助。在模型训练方法中会对其训练方法做详细介绍。

4.3.6 损失函数

损失函数的设计对于目标检测算法训练至关重要，本次实验的损失函数大体上依然使用 CenterNet 的损失函数，主要分为三个部分的损失，下面将对其进行详细介绍。

中心点的预测是将主干网路提取获得的特征 F 输入到几层卷积网络中生成 heatmap 图$\mathrm{H} \in [0,1]^{\frac{\mathrm{w}}{\tau}*\frac{\mathrm{h}}{\tau}*1}$。在计算中心点的损失函数时，对于 GT（真实目标框）的中心点 ,其位置为$\mathrm{p} \in R^2$，计算得到低分辨率（经过下采样）上对应的关键点 $\tilde{P}$ 。然后将 GT 中心点通过高斯核公式$(4-3)$生成真实 hetmap。

$$Y_{Xyc} = \mathrm{e}xp\left(-\frac{(x-\tilde{p}_x)^2 + \left(\left(y-\tilde{p}_y\right)^2\right)}{2\sigma_p^2}\right) \qquad (4-3)$$

将信心分散到热力图$\mathrm{Y} \in [0,1]^{\frac{W}{\tau}*\frac{H}{\tau}*c}$ 上，其中σ_p是目标尺度-自适应的标准方差。如果有两个目标框的高斯函数发生重叠，取其中最大的数值作为该网格的值。在训练时，由于存在难样本的问题，我们使用 Focal Loss 函数进行模型的训练，如公式$(4-4)$和$(4-5)$下：

$$L_k = \frac{-1}{N}\sum_{xyc}\left(1-Y_{\mathrm{xyc}}\right)^{\alpha}\log\left(\bar{y}_{xyc}\right) \qquad if\ Y_{xyc}=1 \qquad (4-4)$$

$$L_k = \frac{-1}{N}\sum_{xyc}\left(1-Y_{\mathrm{xyc}}\right)^{\alpha}\left(\bar{Y}_{xyc}\right)^{\alpha}\log\left(1-\bar{Y}_{\mathrm{xyc}}\right) \qquad otherwise \qquad (4-5)$$

其中 α 和 β 则是该 Loss 函数的两个超参数，理论上可以设置为任意正整数。

但根据 CenterNet 的训练经验，在训练时将两个数分别设置为 2 和 4。公式中 N 是输入图像中的中心点个数，除以 N 主要为了将所有 Focal Loss 归一化。同时为中心点的预测设计损失函数。在这里令（$x_1^{(k)}, y_1^{(k)}, x_2^{(k)}, y_2^{(k)}$） 是目标的 bbox，则中心坐标为$p_k$，计算方式如公式$(4-6)$所示：

$$p_k = \left(\frac{x_1^k + x_2^k}{2}, \frac{y_1^k + y_2^k}{2} \right) \tag{4-6}$$

我们用关键点估计 $\hat{Y}$来得到所有的中心点。此外，为每个目标 k 回归出目标的尺寸s_k，如公式$(4-7)$所示：

$$s_k = \left(x_2^{(k)} - x_1^{(k)}, y_2^{(k)} - y_1^{(k)} \right) \tag{4-7}$$

为了减少计算负担，使用单一的尺寸预测$\hat{s} \in R^{\frac{W}{l}*\frac{H}{l}*2}$，我们在中心点位置添加了 L1 loss，如公式$(4-8)$所示：

$$L_{size} = \frac{1}{N} \sum_{k=1}^{N} \left| \hat{s}_{P_k} - s_k \right| \tag{4-8}$$

由于图像下采样时，GT 的关键点会因数据是离散的而产生偏差，我们在另一个分支中对每个中心点附加预测了个局部偏移$\hat{o} \in R^{\frac{n}{l}*\frac{H}{l}*2}$ ，这个偏移使用 L1 loss 函数来训练，如公式$(4-9)$所示：

$$L_{off} = \frac{1}{N} \sum_{p} \left| \hat{O}_{\tilde{p}} - \left(\frac{p}{l} - \tilde{P} \right) \right| \tag{4-9}$$

在训练时，我们直接使用原始像素坐标来进行损失函数的计算，实验显示此种方式更有效。同时为了调节各个损失对网络权重的影响，我们对各个损失进行了加权求和作为最后的总损失，整体的损失函数如公式$(4-10)$所示：

$$L_{det} = L_k + \lambda_{size} L_{size} + \lambda_{size} L_{off} \tag{4-10}$$

实验中，所有输出共享大部参数权重，我们将超参数λ_{size}设置为 0.1 , λ_{off}设置为 1，最终网络会在每个位置输出 5 个值(即中心点类别数 1 、中心点偏移量的两个参数 x,y 和中心点尺寸的两个参数 w,h)，最终可以得到目标框的位置和大小信息。

4.4 实验结果对比分析

4.4.1 实验细节

本章检测算法的训练与测试环境如表 4-3 所示：

表 4-3 实验环境说明

系统版本	Ubuntu 16.04
GPU 及显存	1080TI,12GB
训练框架	PyTorch0.4.1
其他软件版本	CUDA10.0,cudnn7.6

在训练前，与 YOLO 等基于 Anchor 的检测方法不同的是，我们的检测算法不需要用 K-means 等聚类方法获得数据集的先验框的信息，我们只对数据集的进行必要的数据增强操作，包括图像旋转、旋转等，将图像转化为 512×512 的大小。

在训练中，设计一种多阶段训练方法，具体流程如下：首先使用下采样 8 倍的 Feature Map 做分支预测，将其输出结果作为损失函数的输入对网络进行优化，其中超参数 Batch_Size 设置为 32、学习率为 1e-3，使用 Adam 优化器进行 100 个 epoch 的训练。然后进行第二阶段的训练，在该阶段使用下采样 4 倍的 Feature Map 在做损失函数对整个网络进行微调。由于在浅层特征中已经收敛，所以在深层特征进行微调操作时学习率设置为 1e-4，其他超参数保持不变，进行 60 个 epoch 的训练。该训练方法可以有效提升模型的训练效果，加速模型的收敛，对检测精度的提升有帮助。同时和原始 CenterNet 相比在推理阶段不会增加任何多余的操作。

在推理时，本次实验只在单尺度上进行测试，在推理时先提取 heatmap 中的峰值坐标，具体做法是在3×3的窗口以中心点为阈值，对该中心点周围的 8 个临近点进行对比，如果 8 个临近点的值都小于该阈值，则保留该中心点的坐标，保留满足要求的前 100 个峰值点。然后结合对应的中心点偏移和 w、h，就可以获得目标框的位置和大小信息。

基于以上训练和测试方法，本节将对各个算法模型的性能做全面的对比分析。

4.4.2 教室学生数据集实验结果分析

本小节的测试数据集为教室学生数据集，测试图像的输入分辨率统一为 512×512，测试时的 Batch_Size 为 1。在该数据集上测试结果如表 4-4 所示。

表 4-4 不同网络在 Refine_CenterNet 上的性能表现

网络	AP	AP50	AP75	FPS
ResNet-18	38.1	85.2	30.5	94
Dla-34	44.2	90.7	37.9	52
MobilleNet-v1	40.6	86.7	31.6	72
LimitNet-A	41.5	88.5	33.1	142
LimitNet-B	43.9	90.6	35.4	108

上述结果表明 LimitNet 在人体检测算法具有比较好的性能，准确率和推理速度都比同等量级的网络更好，虽然比复杂的网络结构 Dla-34 的准确率低，但推理速度却快很多。为了测试算法模型中各模块的性能，分别进行了如下几项测试实验结果对比，如表 4-5、表 4-6、表 4-7。

表 4-5 不同网络在 CenterNet 上的性能表现

网络	AP	AP50	AP75	FPS
ResNet-18	37.5	84.3	29.6	100
Dla-34	44.1	90.6	37.7	54
MobilleNet-v1	28.9	84.7	30.1	75
LimitNet-A	40.5	86.5	31.1	148
LimitNet-B	42.1	88.7	33.7	112

首先对比原始 CenterNet 检测框架在各个网络下的测试结果，我们发现 LimitNet 作为头部特征提取网络时相较于其他轻量级网络的准确率更高且推理时延更低，这验证了 LimitNet 网络不仅在分类任务中表现出突出的性能，在目标检测任务中也能表现出好的性能。另外，对于表 4-4 中的检测结果，发现 Refine_CenterNet 在 Dla-34 模型上的性能相较于 CenternNet 提升不明显，这是因为 Dla-34 是一种高分辨率结构的网络，本来就具有浅层特征与深层特征的融合。其它网络在精度上则都有明显。

为了验证特征融合模块的有效性，对比了是否使用该 AFFM 特征融合模块的检测算法在小目标实例的性能表现，小目标的定义是像素面积小于 256 的目标框。如表 4-6 所示：

表 4-6 小目标的检测结果

网络	AP	AP With AFFM	AR	AR with AFFM
ResNet18	15.4	16.7	24.1	25.5
Dla-34	18.4	18.9	26.3	26.6
MoblieNet-v1	15.9	17.8	24.3	26.4
LimitNet-A	16.7	18.7	25.8	26.7
LimitNet-B	18.3	19.6	25.6	27.1

以上结果表明在各个网络中使用 AFFM 特征融合模块后对小目标性能有显著提升。其中召回率有显著在轻量级网络上的性能提升尤为明显，提升了接近两个百分点。由于 Dla-34 网络结构本生就有复杂的特征融合模块的设计，所 AP 和 AR 提

升幅度相对较小。对此，可以完全使用基于 AFFM 的轻量级的网络替代 Dla-34 的复杂网络，保持精度和准确率的平衡。

表 4-7 是使用 NMS 后的检测算法的各项性能指标，可以发现 NMS 对准确度提升不明显，甚至可能会出现 AR（召回率）下降的情况。这是因为在密集场景下样本中，存在许多目标相互遮挡、重叠的情况，NMS 方法处理可能会删除真实的目标框，使得模型的召回率降低。

表 4-7 使用 MNS 后处理方法的检测结果

网络	AP	AP50	AP75
ResNet-18	38.1	85.0	30.4
Dla-34	44.1	90.8	37.8
MobilleNet-v1	40.3	86.7	31.7
LimitNet-A	41.3	88.6	33.2
LimitNet-B	43.8	90.6	35.4

上述实验的性能对比表明，该算法模型能够一定程度上的提升 CenterNet 对于小目标检测的准确率，同时使用 LimitNet 作为主干网络能够提升人体检测的速度，最高可达 142FPS。本文设计的检测算法在计算能力相对较弱的硬件条件下也能达到较好的实时性。检测算法在该数据集上的测试效果如图 4-10 所示。

图 4-10 测试效果图

4.4.3 PANDA 数据集

在 PANDA 数据集中测试中，数据输入和训练一样都是 512×512。和教室数据集的对比实验一致，在 PANDA 数据集中对检测模型的各模块的性能做测试对比实验。首先对各网络在 Refine_CenterNet 测试结果如表 4-8 所示：

表 4-8 各网络在 Refine_CenterNet 上的准确率

网络	AP	AP50	AP75	FPS
ResNet-18	42.1	88.1	34.1	94
Dla-34	48.2	92.2	39.1	52
MobilleNet-v1	42.3	88.7	34.6	72
LimitNet-A	44.5	90.4	38.3	142
LimitNet-B	45.7	91.6	39.4	108

从结果中可以看出，LimitNet 在 PANDA 数据集上同样具有更好的性能，模型更具鲁棒性。同时我们测试了各个轻量级网络在 CenterNet 框架下的检测性能，如表 4-9 所示：

表 4-9 不同网络在 CenterNet 的性能表现

网络	AP	AP50	AP75	FPS
ResNet-18	41.0	86.2	33.0	100
Dla-34	45.2	90.1	38.1	54
MobilleNet-v1	41.2	86.7	33.3	75
LimitNet-A	43.3	88.2	36.3	148
LimitNet-B	44.3	89.3	37.4	112

对比表 4-8 中的测试结果，我们发现改进的检测框架比原始 CenterNet 检测框架的 AP 高接近 2%，且推理速度几乎相同，这验证了该检测算法的有效性。另一方面，只从表 4-9 的对比结果来看，LimitNet 网络同样有更高的性能，这验证了该网络结构的鲁棒性和有效性。

表 4-10 展示了模型是否使用特征融合模块的在小目标上的准确率，分析结果可知，AFFM 特征融合模块对小目标 AP 和 AR（召回率）都一定程度的提升。

表 4-10 各个尺寸目标的性能指标

网络	AP	With AFFM	AR	AR with AFFM
ResNet18	19.5	20.3	21.1	23.2
Dla-34	23.8	24.1	26.1	26.6
MoblieNet-v1	19.8	20.6	22.3	24.1
LimitNet-A	20.2	22.5	22.9	24.1
LimitNet-B	22.5	24.6	24.2	25.8

表 4-11 是没有添加 NMS 后的各项性能指标，可以发现 NMS 对准确度提升不明显，几乎可以忽略。在该数据集上测试效果如图 4-11 所示:

表 4-11 使用 NMS 后的检测性能

网络	map	AP50	AP75
ResNet-18	41.1	86.4	33.4
Dla-34	45.1	90.3	37.9
MobilleNet-v1	40.1	86.0	33.1
LimitNet-A	43.2	88.1	36.4
LimitNet-B	44.4	89.3	37.2

图 4-11 测试效果图

4.4.4 实验测试总结

从以上两个数据集上的测试结果可以看出，本文设计的人体检测算法 Refine_CenterNet 的性能相较于 CenterNet 有明显的提升,特别是在小目标的检测效果上。在加入 AFFM 特征融合模块后，AP 提升约 2%，且对于小目标的检测性能提升明显。同时验证了 NMS 对于 CenterNet 检测方法的性能提升几乎没有作用，这也简化了整个网络的检测流程，提升模型的检测效率。同时在两种数据集下的测试对比中，发现在 PANDA 测试的性能表现更好，这是因为 PANDA 数据集的训练样本足够多。

4.5 本章小结

本章首先分析了当前人体检测技术面临的技难问题，然后详细的介绍了基于 CenterNet 改进的人体检测算法的设计，包括数据集的构建、头部特征网络、特征

融合和 DCN 模块等，最后该算法进行实验验证。经过实验对比，发现本文多阶段的训练方式对提升类似结构的检测模型是广泛有效的，改进的特征融合模块在多尺寸目标下的数据集中表现更好；同时本章的人体检测算法平衡了人体检测速度和精度的问题，在教室学生数据集上能够达到 43.9%的 AP，并且推理速度能够达到 67fps/s。

第五章 人体行为识别技术的研究

人体行为识别是当前机器学习领域研究的热点，在医疗辅助、国家智能安全、虚拟现实等领域有着非常重要的意义。与更加成熟的人体检测技术相比，人体行为识别技术的研究难度更大，目前在该领域的研究成果更少。在行为识别技术实际应用的过程中，通常需要先通过人体检测技术检测人体的位置和大小信息，从而进一步进行人体行为的识别。在第四章设计一种轻量化的人体检测算法，本章基于先前学者研究的基础上，先分析了人体识别技术的难点，然后针对这些问题，结合人体姿态和外观特征，设计了一种多分支网络模型对人体行为进行识别。该模型不仅可以图像样本数据集和视频数据集上达到较高准确率，而且算法计算量相对较小。

5.1 人体行为识别技术的概述与难点分析

人体行为识别的定义简单来说就是对人的某些动作（如吃饭、打球、打电话等）进行分类，由于该分类任务更加抽象，所以目前还未取得比较瞩目的研究成果，但该方向依然是近年来学者们研究热点。在应用方面，人体行为识别在视频监控、医疗辅助、机器人等领域发挥着重要的作用。目前在该领域根据主流研究内容可以将行为识别分为两种：基于单帧图像的行为识别和基于视频的行为识别研究。前者研究的瓶颈在于单帧图像包含的有效信息很少，对于比较复杂的动作分类任务，单从一张图像去识别复杂的人体行为是很艰难的，这需要研究更加复杂的网络结构或者针对特定行为手工设计特征的方法，显然这类方法效率不高。由于数据的发展，基于视频的人体行为识别方法研究逐渐增多，但始终没有非常鲁棒的解决方案，无法实现实用化和产业化。视频行为识别的主要难点有以下两个方面：

1）计算复杂度大：在视频研究领域，一个重要的难点在于时序，由于视频是由具有时序的单帧图像组成的，如何去表示或者提取这些时序特征也是当前研究的难点和热点。但很多单帧的图像是无法判断的行为通过视频可以准确识别，通常需要 3D 的卷积或者将视频序列进行处理后输入，算法的复杂度提升一个数量级。

2）数据问题：目前公开的人体行为识别的数据集很少，且大多数行为识别的数据集都是从视频网站上获取的；由于数据集不规范，视频片段时长不一致、光照、人体多尺度、摄像头位置等问题，使得学者们无法在人体行为识别技术上的研究更进一步。

本文针对上述行为识别的难点问题，基于姿态和外观特征设计一种高效轻量级的算法模型。

5.2 数据集的准备

本节对本文要用到的人体行为识别数据集做简单的介绍，主要包括人体姿态数据集 MPII，自研的人体行为教室学生数据集和公开的 Penn action 数据集。

5.2.1 MPII 数据集

MPII 人体姿态数据集是目前人姿势估计的性能评估比较流行的数据集，该数据集中总计包括大约 2.5 万张图像，包含超过 4 万个人体实例，且都带有 16 个关节点的标注信息。这些图像的收集来源是对人类日常活动中人体数据的合法收集。总体而言，其中该数据集能够涵盖大约 410 项人类活动，且每个图像都带有活动标签。每个图像的来源都是从 YouTube 视频中提取，并提供前后未注释的帧。此外，对于测试集，该数据集还有更丰富的注释，包括身体部分遮挡和 3D 躯干和头部方向等。

5.2.2 教室学生数据集

该数据数据集的场景是学生教室的拍摄图像或视频，包括 2 个学校的视频片段和 15 个学校的图片数据，对于视频数据，我们按时间间隔直接把它截帧得到图片数据，然后手工标注目标位置、大小以及行为分类信息。该数据集包括三类：站立、坐下、趴桌子。人体边框数量 31320 个，其中数据分布为坐下 24467 个、趴桌子 4186 个、站立 2667 个。

5.2.3 Penn action 数据集

宾夕法尼亚大学行动数据集（宾夕法尼亚大学）包含 2326 个视频序列，每个序列包含 15 个不同的动作和人体关节点标注。

5.3 人体行为识别网络的详细设计

行为识别在我们日常生活中应用广泛，在智能监控、人机交互、视频序列理解、医疗健康等众多领域扮演着越来越重要的角色。目前行为识别主要在两个方向的进展：一种是基于 3D 卷积神经网络，由于硬件计算能力快速提升，基于 3D 卷积神经网络的方法开始广受学者们的关注和研究，但该方法通常计算复量相对较大，以现有的普通 GPU 的能力，对于在监控视频中实时监控并分析人体行为的任务没有太大研究价值。所以本节主要研究基于姿态识别的行为识别方法。姿态估计在行为识别中扮演中重要的角色，通常后一个被用作前一个问题的先验，姿态估计和行为识别有着极大的相关性。目前许多方法都是用姿态模型提取人体的姿态特征（关

节点坐标），然后再将其用做行为识别网络的输入，从而提升行为识别的精度，但这种两阶段的方法计算复杂度比较高，通常都是分阶段进行的，目前将两者联合起来解决这两个问题的方法几乎很少，本章提出一个可以端到端多任务的框架来同时处理这两种任务。下面将对算法模型进行详细的介绍。

5.3.1 模型整体框架

本章的任务在于设计出一种轻量级、能够应用到实时监控系统中的人体行为识别算法模型，基于[58]等方法，设计了一种多分支网络模型，如图 5-1 所示：

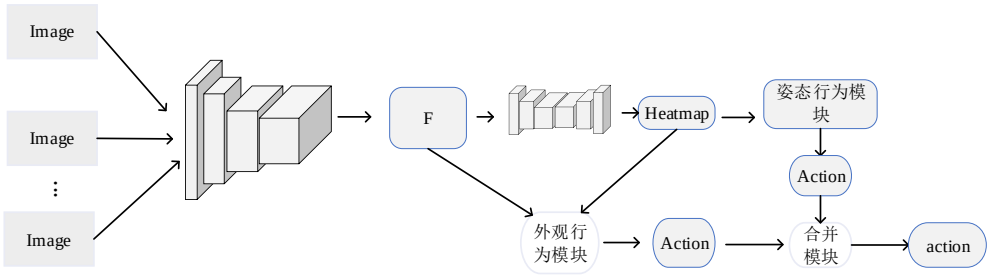


图 5-1 模型整体框架

其中一个分支基于人体姿态进行行为识别，另一个分支则是基于外观特征进行人体动作分析，最后将两者特征进行融合做最终的行为分类。具体来说，假设令输入 Image 为 RGB 视频或者图像为I $\in R^{w*h*3*t}$，其中宽是 w、高是 h、深度是 3，t 是视频的时间序列（如果是单帧图像的行为识别，则 t 为 1）。根据研究者们的经验，t 通常为 6、8、16，因为过小的帧数不能完全代表行为的信息，过多的帧数会过增加计算量。然后将I输入卷积神经网络中分别提取视频的外观特征F $\in R^{\frac{w}{l}*\frac{h}{l}*c*t}$ 和 heatmap$\in R^{\frac{w}{l}*\frac{h}{l}*J*t}$，其中l是下采样倍数，J 是人体关节点的个数。然后 heatmap 可以通过 S-Soft-Argmax 方法生成关键点的坐标矩阵，该方法在后面会详细介绍。另一个分支通过 F 和 heatmap 的操作可以得到增强的外观特征 FH，最后利用关键点坐标矩阵和 FH 特征可以得到最后的行为类别矩阵。该框架结构比较复杂，但可以同时进行姿态检测和行为识别两种任务，并利用姿态检测任务提升对行为识别的识别性能，在现有的行为识别方法中相对计算量更小。下面各小节会对各个模块进行详细的介绍。

5.3.2 姿态估计模块

由于直接回归关键点坐标任务难，所以我们的姿态估计模块是基于 heatmap 的方法来获取人体的关键点的坐标，该方法可以不采用后处理方法就能够得到关键

点的坐标，便于后面进行端到端的训练。

本文的研究特点在于设计一种具有高效实时性的模型算法，能够用较小的计算代价实现良好的行为识别准确率。从架构图中可以看到，该 heatmap 提取模块主要包含两个模块，头部特征提取和上采样模块。由于在第三章和第四章已经验证 LimitNet 在人脸识别以及目标检测任务有较好的性能，在头部特征提取网络的设计上依然采用轻量级的 LimitNet 网络。

根据研究表明，在上采样之前保持高分辨率表示是有益的。HRNet[60]在整个过程中都保持了高分辨率的表示，其优越的姿态估计结果从经验上证明了上述思想。然而，大多数现有的方法往往会产生低分辨率的表示(例如，68,44)，因为通常改分辨率的表示会严重增加算法的计算代价。为了在获得更高分辨率表示的同时保持轻量级的特点，与第三章中头部特征提取网相同，我们在第五阶段依然设置 stride 为 1，将输出尺寸保持和第四阶段结束相同，这样不仅可以提升关节点识别的准确率，而且在姿态估计模块可以删除了一组反卷积层操作，从而减小了一部分计算量，头部特征提取网络的整体结构如图 5-2 所示。

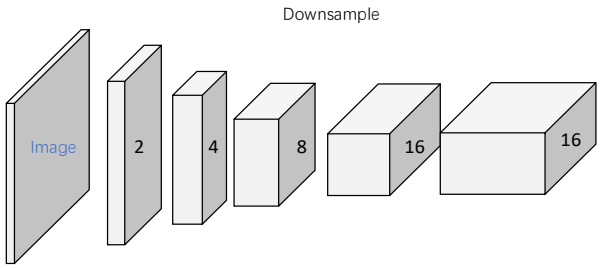


图 5-2 头部特征提取模块

通过下采样网络提取到深度外头部特征 F 后，我们将 F 输入上采样模块进行特征的上采样。该姿态模块包括上采样模块以及 heatmap 处理模块。

根据 HRnet[60]等工作的研究，heatmap 的形式比直接回归关键点的坐标更有效，且 heatmap 越大，最后关键点的提取效果就越好，但通常提取 heatmap 的网络也会比较复杂。为了简化网络同时减小网络的前向推理时间，与第三章人体检测任务不同，我们不需要进行特征融合操作，同时我们并没有使用像 HRnet[60]中比较复杂的网络结构，而是自己设计一种简单的编解码形式的网络结构，如图 5-3 所示，其中 Gdc 表示分组反卷积。

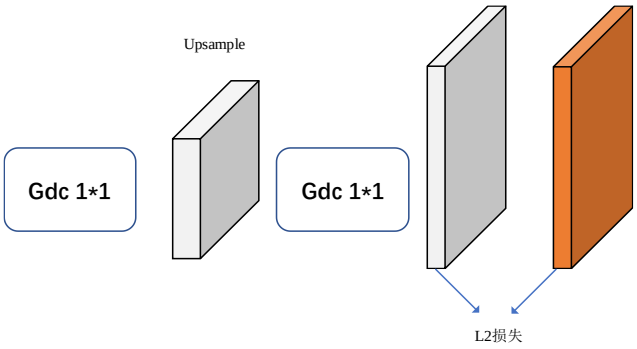


图 5-3 上采样模块

上采样模块包括两次反卷积的操作，通过这样的操作最后 heatmap 长和宽都只是原始图像的四分之一，最后生成的 heatmap 能够提取更精确的关键点坐标。为了减小模型的整体计算量，在该模块中的反卷积操作使用组反卷积。这里没有对其进行使用第四章中将小尺寸的 heatmap 用来训练，因为关键点个数太多，在小尺寸上训练效果不好，容易造成信息干扰。我们使用另外一种方式，在论文[58]中证明了多阶段的网络的有效性，我们也在上采样模块设计了多个 block，每一个 block 都是如图 5-3 所示的结构，从前往后可以精细化关键点的坐标。

5.3.3 S-Soft-Argmax 方法

在基于 heatmap 的姿态估计方法中，现有的方法通常使用 argmax 函数从 heatmap 中获取关节点的坐标信息，然后将坐标信息转换到原始的图像分辨率对应的位置；但通过 argmax 函数获得的坐标值是离散的，只能是正整数，在对其还原到原始分辨率的过程中会导致位置偏移，这限制了最终关节点坐标的准确度，而且由于 argmax 函数是不可导的函数，整个模型不能端到端的训练。Luvizon 等人[58]和 Sun 等人[59]基于 argmax 函数提出了 Soft-argmax 的技术回归人体关节点的坐标信息，整个过程可微，可以进行端到端训练。Soft-argmax 函数的原理如下所述。

首先，真实 heatmap 是通过在每个关节点上使用二维高斯分布生成的，如公式（5-1）所示：

$$H_j(x,y) = exp\left(-\frac{(x-x_j)^2 + (y-y_j)^2}{2\delta^2}\right) \qquad (5-1)$$

其中H_j表示 heatmap，σ 表示高斯函数的标准差，(x_j, y_j)表示第 j 个关节点的二维坐标。然后使用H_j通过 Softmax 函数直接生成概率$s_j(x,y)$，如公式（5-2）所示。Soft-Argmax 的整个过程可微，能够端到端的进行训练。

$$s_j(x,y) = \frac{e^{H_j(x,y)}}{\Sigma_x \Sigma_y e^{H_j}(x,y)} \quad (5-2)$$

但是通过高斯分布生成的 heatmap 中元素的值会被归一化到[0,1]，会存在大量的元素的值接近 0，即使真实值为 1，其他元素为 0，但因为$e^0 = 1$，$e^1 = e$，则概率 p 为：

$$\mathrm{p} = \frac{e}{n-1+e} \quad (5-3)$$

heatmap 中大量的元素会使 p 概率值变小，进而影响 Soft-Argmax 在空间应用 Softmax 时的精度。

为了解决这个问题，参考 Arcface loss 的原理对 Soft-argmax 方法改进，其中 Arcface 具体原理在第三章已经详细介绍，提出了一种名为 S-Soft-Argmax 的方法，公式如（5-4）所示。

$$s_j(x,y) = \frac{e^{sH_j(x,y)}}{\Sigma_x \Sigma_y e^{sH_j}(x,y)} (b > 1) \quad (5-4)$$

可以看出该方法和 Arcface loss 有相似之处，与 Soft-Argmax 相同，最终坐标通过公式（5-5）计算得到。

$$\hat{x}_j = \sum S_j \nabla W_x, \hat{y}_j = \sum S_j \nabla W_y \quad (5-5)$$

其中w_x和w_y为等权矩阵,如[0,1,2,…,r]，r 为对应的行号或者列号。∇为逐元乘法。证明了 S-Soft-Argmax 在后期处理中是一个模型无关的一般函数，它对我们的网络模型的准确率提升有好处。

5.3.4 动作识别模块

如图 5-1 所示，动作识别方法主要分为两部分：基于身体关节坐标和基于外观特征。基于身体关节点坐标的部分是通过提取人体的关节点坐标进行行为识别任务，而基于外观特征部分则更关注人体的外观特征以及背景等因素。最终对两部分动作特征进行融合操作，即可完成对整个人体行为的识别。在这一节中，我们详细的描述了这两个关键部分。

5.3.4.1 基于姿态的动作识别

前面姿态模块可以产生关节点的热图，然后通过 S-Soft-Argmax 操作可以得到关节点的坐标矩阵。这里我们使用关节点矩阵作为动作识别的输入，主要有以下两个原因：1）由于每一张 heatmap 图有多个通道，如果将图序列的 heatmap 进行整

合是一件比较麻烦的事；2）关节点的坐标信息比 heatmap 更具有关节点位置联系的表达能力，可以减轻行为识别模块的计算量。

对此，为了探索人体关节位置编码的高级信息，我们将图像序列的关节点坐标矩阵进行整合，我们选择将时间维度编码为纵轴，关节编码为横轴，每个点的坐标(2D 为(x, y)， 3D 为(x, y, z))编码为通道。利用该方法，我们可以直接用 2D 卷积网络直接提取有时序的关节点矩阵的特征，这会降低模型特别大的计算量。由于姿态估计方法是基于静止图像的，我们使用一个时间分布的抽象来处理视频剪辑，通过随机抽样视频得到，这是一种简单的方式来处理单个图像和视频序列。

由于人体行为可能受到多个关节点的影响（如学生坐着看书），也可能只受一个关节点的影响（如摇头动作等），这样的不确定性让该任务变得困难。如果针对特殊动作设计关节点的手工特征，这会是一种非常实用的算法，但其鲁棒性不高，只能在特定场景特定动作下使用，而且效率低下，泛化性不强。相反，直接使用二维卷积也可以加强关节之间的关联，而且相对更容易训练。基于上述的原因，在基于姿态的动作识别模块，我们使用一种全卷积神经网络来从输入的关节点坐标矩阵中提取特征，整体的架构如图 5-4 所示：

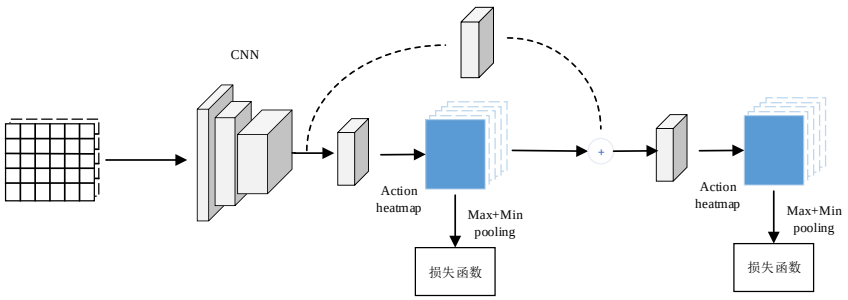


图 5-4 基于姿态的动作识别网络

首先将坐标矩阵通过卷积网络输出特征，然后生成的 Action map（动作热图），其中 Action map 的通道数为行为类别数。生成 Action map 的目的是这种结构会在随后的层中进行通道组合产生更具表达能力的特征。另外，在生成 Action map 时，不同的关节会产生不同坐标变化，从而进行关节关联响应，从而在深层产生更具区别性的激活。

最后通过概率函数将输入视频或图像产生每个动作的输出概率，在之前，我们使用 Action map 做池化操作，将其转化为 Softmax 需要的输入类型，动作地图上的池操作。为了对每个动作的最强反应更加敏感，我们使用了 Max + Min 池化。Max 池化是寻找最具影响力的关节点组合，而 Min 池化是最不影响力的关节点组合，两者相加能够让动作更具鲁棒性。

此外，受基于人体姿态的动作识别方法的启发，我们使用带有中间监督的堆叠

结构来细化动作识别，每个预测块的 Action map 被重新注入下一个动作识别块，该方法能够进一步提升模型的准确率。

5.3.4.2 基于外观特征的动作识别

基于外观的行为识别部分与基于姿态的部分相似，不同之处在于两者的输入特征不一致，前者是局部外观特征，而后者是关节点坐标矩阵。在 Luvizon 等[58]的研究中表明，通过外观特征和概率 map 进行融合能够提取局部化的外观特征表达能力，能够有效的提升行为识别的准确率。具体操作为：将在全局输入流的末尾得到的视觉特$F_t \in R^{w_f * H_f * N_f}$的张量乘以概率映射$M_t \in R^{w_f * H_f * N_J}$最后得到的位姿估计部分，其中$w_f * H_f$为 Feature map 的大小，$N_f$是特征的数量，$N_J$是关节点的数量，然后让 Feature map 和每个概率映射按通道相乘，得到一个大小为$R^{w_f * H_f * N_J * N_f}$的张量，这种方式可以加强某个关节点附近区域的外观影响力，使得特征更具表达性。最后，将空间维度按总和折叠，得到时间 t 的外观特征大小为$R^{N_J N_f}$。对于一个帧序列，我们有 t={0,1，…， T}的视频剪辑外观特征$V \in R^{T * N_J N_f}$。上述外观特征提取过程如图 5-5 所示。

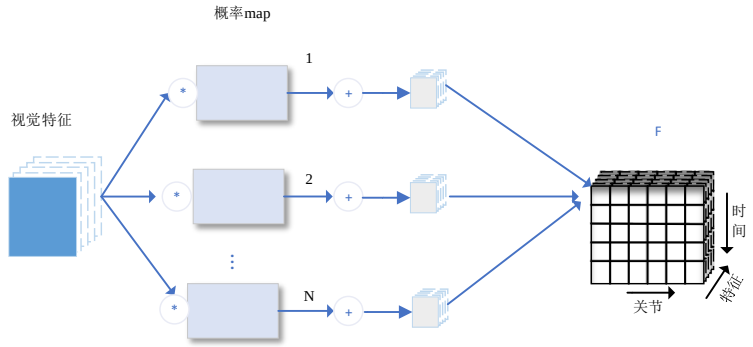


图 5-5 融合模块

获取到输出 F 后，需要将其进行分类，在这里我们采用和普通分类网络类似的结构，如图 5-6 所示。

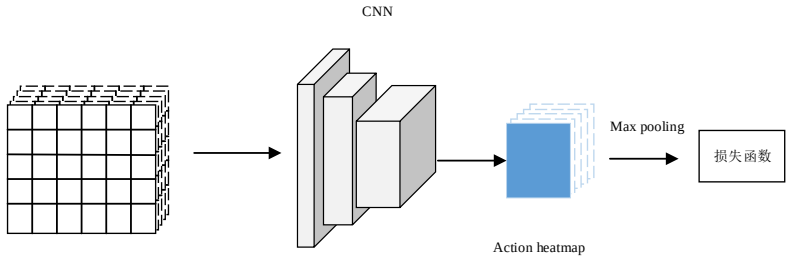


图 5-6 基于外观的动作识别模块

5.3.4.3 基于姿态与外观特征融合的行为识别

有些动作很难通过高级的姿势表现来区分，如果只是单纯的考虑关节点的坐标对某些动作是很难区分的，例如，摸头和打电话的动作；但是如果能够提取到手机的视觉信息，在分类输出中就很容易将不同行为动作分开。另一方面，某些动作与视觉信息的相关性较弱，却与身体关节点位置有极高的相关性，如敬礼、摸头等，在这种情况下，关节点位置信息对于行为的分类更加重要。因此，为了提升模型的行为区分能力，我们设计了 AFF（动作特征融合模块）将 action map 和外观 action map 进行融合，主要思想是通过对 Feature map 进行处理，生成一个和其通道数一样的一维向量，然后将该向量作为每个通道的加权的权重，最后将该权重分别与对应通道相乘输出融合后的特征。具体细节如图 5-7 所示。

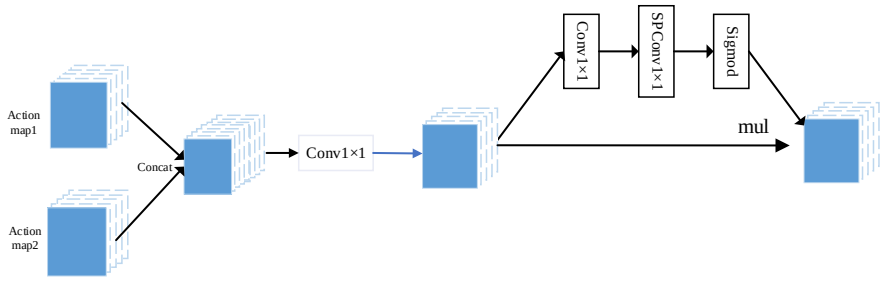


图 5-7 动作特征融合模块

AFF 模块的流程首先是先将两个 Action map 进行 Concat，此时通道数加倍；然后通过 1×1 卷积将特征通道融合，通道数减半，通过两层卷积、平均池化层和 sigmod 生成对应通道的权重，最后再与原始特征对应通道相乘既可得到最后的新的 Action map。该结构让两种 Action heatp 自动学习加权权重，方便训练和优化，提升模型准确率。

行为识别双流模型充分利用了人体行为的特点，将其分为外观和姿态的动作识别主要有两个好处，首先，这两部分可以共用大部分的参数，这比两阶段的模型的时延更少，计算效率更高。其次，提取的视觉特征任务和姿态识别任务之间耦合度低，可以独立运行，对于姿态估计能够独立运行，这对于每个模块的升级有很大好处。

5.3.5 损失函数

5.3.5.1 姿态识别损失函数

在姿态识别时，我们采用如下损失函数[58]进行训练，如公式（5-6）：

$$L_p = \frac{1}{N_J}\sum_{n=1}^{N_J}\left(\left|\left|\hat{p}_n - p_n\right|\right|_1 + \left|\left|\hat{p}_n - p_n\right|\right|_2\right) \quad (5-6)$$

其中$\hat{p}_n$和p_n分别为第 n 个关节的估计和真实框的坐标。该损失函数使用了坐标之间的 L1 损失与 L2 损失，然后对其进行求和、平均操作得到最后的损失函数值，使用这样的损失结构可以提高训练效果。。

5.3.5.2 动作识别损失函数

在本文第三章中介绍了交叉熵损失函数，从其公式可以看出其输出概率越大损失越小。在目标检测中，对于负样本而言，输出概率越小则损失越小。此时的损失函数在大量简单样本的迭代过程中比较缓慢且可能无法优化至最优，动作识别分类任务是一种比较难的任务，由于行为的相似性，难样本较多，同时样在类别不平衡数据集中，样本数量少的类别的样本天然就属于这种困难的样本。交叉熵损失在目标检测领域，2017 年，Kaiming He 在 RetinaNet 中提出 Focal Loss 来解决检测任务中的正负样本不均衡的问题与训练效率低下的问题。我们对其稍加修改即可应用到人体行为识别的任务中。

在第四章人体检测的损失函数中详细介绍了 Focal Loss 的形式，本章对其用在动作分类任务中，如公式（5-7）：

$$\mathrm{F(x,c)} = -\left(1-\frac{e^{x_c}}{\sum_j e^{x_j}}\right)^{\alpha} log\left(\frac{e^{x_c}}{\sum_j e^{x_j}}\right) \quad (5-7)$$

公式中，其中 x 是神经网络的输出，首先经过 Softmax 函数，获得相应属于某行为的概率，从公式中可以看出，Focal Loss 可以降低概率高的样本的损失，从而相对地提升概率低地样本的损失，这对学习提升难样本的训练效率更有帮助，其中α是一个正整数。

5.4 实验结果对比分析

本节将对人体行为识别算法的实验环境、训练方法以及各模型的测试结果等进行详细的介绍。首先在训练和测试的实验中，实验具体硬件和软件环境如表 5-1 所示：

表 5-1 硬件环境

系统版本	Ubuntu 16.04
CPU 版本	InteL(R) Xeon(R) CPU E5-2690 v2 @ 3.00GHZ，4 核,8 线程，32G 内存
GPU 及显存	1080TI,12GB
训练框架	Tensorflow1.9
其他软件版本	CUDA9.0,cudnn7.0

5.4.1 实验细节

首先在姿态模块中训练的数据集来自 MPII 和 Penn Action 的混合数据，其中每个人体包含 16 个关节点，测试集来自 MPII 数据集。在动作识别模块，我们使用 Penn action 数据集作为测试集。在行为识别的训练中，我们先对视频数据按固定间隔分为 N 块，然后对每一块视频进行随机采样一张图像，组合成 N 张图像作为一个样本输入，我们在训练中将 N 设置为 16。因此，在我们的训练样本数将会大量增加。在推理中，本文的测试方法使用[61]中的测试方法，包括两种测试方案：视频样本中单片段或多片段的结果。对于单片段，在视频的中间裁剪一个单独的视频片段，然后通过算法模型测试。对于多片段测试，我们裁剪多个 T/2 帧时间间隔剪切，多片段剪切的方式最后的分数是通过所有片段的平均结果来计算的，最后在 Penn 动作数据集的 2D 场景上评估我们的动作识别方法。

5.4.2 训练方法

在训练中，我们提出一种多模块-多阶段训练方法。该方法先对各个模块进行局部训练，再对每个模块训练时采用多阶段训练方法。在每个模块的训练中，进行多阶段的迭代更新，其中每个训练阶段结束时评估性能，然后选择最优参数进行下一阶段的训练，一直重复直到损失不下降为止。下面对各模块的训练设置进行详细介绍。

首先在姿态估计模块，对于姿态估计的初始化，使用 Adam 优化器进行模型的参数更新，初始学习率为 0.001，训练时使用 Batch_Size 为 16，初始学习率 1e-3、使用 Adam 优化器进行 120 个 epoch 的训练，其中学习率没 40 个 epoch 衰减 0.，其中 S-Soft-Argmax 中的 s 设置为 60。

在基于姿态的动作识别模块和基于外观的动作识别模块，使用相同的训练方法。使用 Adam 优化器进行模型的参数更新，损失函数使用 Focal Loss，初始率设置为 0.1，训练 100 个 epoch，在第 30、60、90 个 epoch 衰减 0.1。

在整体的动作识别训练时，主要包括两部分损失，姿态估计损失与动作识别损失。在训练前，需要加载先前训练的参数权重，采用 Adam 优化器，学习率初始值为 0.01，由于是对模型的少部分权重进行训练，所以学习率初始设置相对较小。我们将姿态估计的损失和最终动作估计的损失进行加权操作，加权权重分别为 1 和 0.01，因为 Focal Loss 损失的梯度比姿态估计的损失要更大。

5.4.3 姿态识别评估

本次使用的姿态识别评估的数据集是 MPII。从表 5-2 的测试结果可以看出，姿态估计模块的精度能够接近当前先进方法的性能，但由于姿态估计算法采用的是轻量级的网络 LimitNet-A，在推理速度上会有一定的优势。同时对比了是否使用 S-Soft-Argmax 的准确率对比，可以发现使用 S-Soft-Argmax 后能够提升姿态估计的 0.8%精度。

表 5-2 姿态估计准确率对比

方法	PCKh@0.2	AUC@0.2	PCKh@0.2	AUC@0.5
Iter. Error Feedback	46.8	20.6	81.3	49.1
Heatmap regression	61.8	28.5	89.7	59.6
DeeperCut	64	31.7	88.5	60.8
Pose Machines	64.8	33	88.5	61.4
Stacked Hourglass	66.5	33.4	90.9	62.9
Mulyi-Context Att	67.8	34.1	91.5	63.8
Self Adversarial	68	34	91.8	63.9
Ours	66.3	33.1	90.1	62.5
S-Soft-Argmax	67.1	33.5	90.4	63.0

5.4.4 行为识别评估

本小结对行为识别算法的测试结果进行对比分析。分别对基于姿态特征、外观特征以及两者结合的方法分析测试结果，详细细节如下所述。

5.4.4.1 基于姿态的行为识别评估

首先只使用姿态信息对行为分类，在数据集 Penn action 和教室学生数据集的性能结果如表 5-3 所示：

表 5-3 准确率

数据集	Acc%(pre anno)	Acc%(true anno)
Penn action	82.7	85.4
静态图像	85.72	-

可以发现只使用预测的关节点坐标进行人体行为识别，在数据集 Penn action 的准确率只有 82.7%，静态图像数据集的准确率只有 85.72%。即使用真实的关节点坐标作为动作识别网络的输入，在 Penn action 数据集上的准确率也只有 85.4%。

对算法分类详细测试结果进行统计分析，如表 5-4 所示：

表 5-4 教室数据集下模型识别结果统计

真实类别	被分为站立（%）	被分为坐下（%）	被分为趴桌子（%）
站立	98.56	1.44	0
坐下	3.37	86.56	10.07
趴桌子	0.3	27.12	72.68

从以上结果分析，只基于姿态信息进行人体行为识别，能够保证站立目标的分类准确率，但难以区分趴桌子和坐下两类行为，这是因为趴桌子的行为与桌子的特征有很大关系，而姿态信息只有人体的关节信息，姿态估计模块是几乎不能获取道桌子的任何特征。因此只用姿态信息在教室数据集下是很难做到高准确率，Penn action 数据集准确率低的原因也一样，这里就不过多分析。

5.4.4.2 基于外观的动作识别评估

此外，本章也对只使用外观信息对人体行为进行分类，和其他普通分类任务一样。在各个数据集上的识别准确率如表 5-5 所示：

表 5-5 识别准确率

数据集	Acc%（pre anno）	Acc% (true anno)
Penn action	83.6	85.4
教室数据集	87.7	-

结果显示，在 Penn action 的准确率同样只有 83.6%，而在教室学生数据集上只有 87.7%，即使用真实关节点坐标对 Penn action 中视频片段行为识别的准确率也只有 85.4%，远远低于现有的先进的方法。同样对各类的识别结果进行统计分析，如表 5-6 所示：

表 5-6 教室数据集下模型识别结果统计

真实类别	被分为站立%	被分为坐下%	被分为趴桌子%
站立	98.61	1.39	0
坐下	3.15	88.49	8.36
趴桌子	0.4	23.42	76.18

从以上结果发现，只有站立和趴桌子之间容易区分，其他类别之间的区分能力有限；这是因为视觉信息的局限性，人体行为中姿态信息在某些类别比外观信息更重要。基于外观的行为识别方法会丢失人体姿态对行为动作的影响，这是不能接受的。但我们发现在教师数据集上只基于外观信息的方法要比只基于姿态的方法的准确率要低，而在 Penn action 数据集的准确率比只基于姿态信息的准确率要高，这是因为 Penn action 数据集中包含了大量背景信息，所以外观信息的影响要更大。

5.4.4.3 基于姿态和外观特征的行为识别评估

本节对基于姿态特征和外观的特征融合后的行为识别模型做测试结果对比分析，在 Penn action 数据集和教室数据集上的识别准确率如表 5-7、表 5-8 所示：

表 5-7 方法准确率对比

方法	Annota pose	Estimated pose	准确率
Nie et al.[55]	-	√	85.5
Ipbal et al.[56]	-	√	92.9
Cao et al.[57]	√	-	98.1
		√	95.3
Luvizon et al.[58]	√	-	98.6
		√	97.4
Ours	√		97.4
		√	96.2

表 5-8 在教室数据集上的准确率

数据集	Acc%（pre anno）	Acc% (true anno)
Penn action	97.4	96.2
教室数据集	91.4	-

在 Penn action 数据集的测试结果中，和现有的先进的行为识别算法进行比较，

结果表明本文的算法可以达到较高的准确率，只比 Luvizon et.al[58]低 1.2%，但其采用 Xception 作为主干网络，推理时延为 70ms，而本文的算法只有 60ms。同时利用真实关键点坐标和使用姿态估计预测的关节点坐标的准确率只相差 1.2%，这证明了姿态识别模块的有效性。对比表 5-3 和表 5-5 中的测试结果，可以发现融合两者的特征后对人体行为识别的准确率有显著提升，这验证了姿态信息与外观信息在行为识别任务中的重要性。

5.5 本章小结

本章设计实现单阶段多任务的人体行为识别算法，可以进行人体关节点检测和行为识别任务。在姿态识别模块，使用基于 heatmap 的姿态估计方法，提出 S-Soft-Argmax 方法从 heatmap 中直接生成关节点坐标，让算法可以端到端的训练，提高人体关节点检测的准确率。在行为识别模块，基于姿态识别模块的结果和浅层的人体外观特征对人体行为分类，其中设计了一种自适应加权的方法将外观特征和姿态特征进行融合，在准确率上能够显著提升。本章做了大量的对比实验验证各模块的有效性，最后在公开数据集 Penn action 下能够达到 96%的精度，高于现有的大多数算法，且推理时延只有 60ms。

第六章 全文总结

6.1 论文总结

人体检测与行为识别在视频监控、国家安防等领域的有着重要的研究价值和意义。本文详细介绍了当前人体检测与行为识别技术研究现状，包括数据集、技术难点、当前主要研究方法以及相关技术理论等。然后针对人体检测与行为识别的难点问题，本文主要以下几个方面做了深入的研究：

1、为了提升人体检测与行为识别算法推理速度，本文基于 ResNets 的残差结构设计了一种轻量级的特征提取网络 LimitNet，并且使用网络 op 融合的方法在推理阶段对模型加速，最后 LimitNet 网络在输入分辨率为112×96时，单帧推理时延只有 4.1ms。同时为了验证网络的性能，构建了的几种复杂的测试集，最后该网络在几种测试集下的整体性能表现比当前的流行的轻量级网络更突出。并且将其应用到人体检测与行为识别算法模型中同样能表现出很高的性能，进一步验证了该网络的高鲁棒性与通用性。

2、在人体检测算法的深入研究中，针对当前检测多尺寸目标的网络推理延迟大、计算量大的问题，对 CenterNet 的检测算法进行改进，提出 Refine_CenterNet 检测算法。其中使用轻量级网络 LimitNet 作为头部特征网络，然后通过设计特征融合模块增强算法在小目标实例的检测效果。同时在预测模块设计了一种多阶段分支预测模块改进算法整体性能。最后与 CenterNet 检测方法对比，该算法不仅能够显著提升检测准确率，而且在推理时延上有一定程度的减小。

3、在人体行为识别的算法研究中，基于人体的多模态特点，设计了一种多任务网络模型。其中在姿态估计中，提出 S-Soft-Argmax 方法直接从 heatmap 中回归对应的人体坐标，使得模型可以端到端的训练，提升姿态估计的准确率。在行为识别模块，基于姿态与外观特征，设计一种自适应特征加权的方法对两者进行融合，显著提升了模型的准确率。该模型在 Penn action 数据集中能够达到了 96%的识别精度，且推理时延只有 60ms。该模型不仅保持着高准确率、低时延，而且能够获取到人体的姿态信息，在实验还分析中验证了姿态特征与外观特征对行为识别的重要意义。

最终整体的人体检测与行为识别算法在 1080TI GPU 上不仅能够达到高准确率，而且总体时延只有 67ms，整体性能在目前的算法中很有竞争力。

6.2 后续工作展望

本文设计了一种轻量级网络 LimitNet，并将其应用到人体检测与行为识别算法中，并在人体检测和人体行为识别的深入研究中也取得了阶段性的成果。但由于时间和设备的限制，还有很多技术问题亟待解决和优化，主要包括以下几个方面：

1、在第三章中，提出了一种轻量级网络 LimitNet，该网络在各方面的性能与当前的轻量级网络中对比中表现更突出。但在整体的网络结构的构建中，网络结构的深度、宽度等主要基于当前经典网络（如 ResNets 等）的经验进行设置，而这种方式对于网络的结构设置考虑不全面，必定会存在推理速度、准确率更高的网络结构。未来可以使用模型自动搜索的方法对网络模型各层的深度与宽度以及每阶段的层数做搜索，然后进一步提升该网络的性能，但这种方法对设备的要求比较高。

2、在人体检测算法的研究中，主要基于两种数据集对算法的性能进行验证，但数据集的场景较单一，样本多样性不足，导致泛化性能差。由于当前存在人体检测数据集较少，且规范不统一的问题，未来可以对现有人体检测数据集进行整合优化，构建一种多样性的数据集，这将大幅提升算法模型泛化性能。同时本文中教室学生数据集是非专业人士人工标注，且每个人标准不统一，所以可能存在一些误标等问题，造成对模型训练的影响，后面可以对其进行优化。

3、在第五章的人体行为识别的研究中，设计一种多分支网络对人体动作进行分类，在基于姿态特征的分支中使用图像序列中各关节点坐标矩阵 Concat 后的特征作为行为识别的输入，该方式的缺点是坐标信息矩阵对于人体关节点的联系还有待验证。未来可以通过图卷积或者人工设计一种通用性关节点特征表示等方式对人体关节点信息生成更具代表性的行为特征。

4、在行为识别算法的研究中，由于当前数据集不足和硬件设备落后等原因，导致不能将整个视频直接输入网络模型中提取特征，而是对视频进行间隔采样的得到的图像序列输入到网络模型中提取特征，这对于某些动作的识别会有较大的干扰。同时由于行为视频很容易获得，但对其进行准确的标注会耗费巨大人力。基于未标注的行为数据样本，可以使用无监督方法生成一种对抗网络模型，然后再用已标注的数据进行微调，这对于提高行为识别的鲁棒性有帮助。

5、在第四章的人体检测算法的研究中，设计的检测算法是基于 heatmap 进行中心点的回归，由于下采样 4 倍后 heatmap 的分辨率固定为 128×128，模型的输入只能固定分辨率 512×512，导致模型的对于数据输入的泛化性不好，如对于高分辨率的样本几乎不适用。这是当前算法还需要改进的地方。

6、在本文的算法测试实验中，只对比了在 1080TI 上的测试结果，且设计的网络模型只针对 GPU 平台。由于目前手机、摄像头等移动设备的普遍应用，对该算

法在嵌入式设备的研究意义同样重要。后续可以对该网络模型设计一种高效的压缩方法，如量化、剪枝等，在不丢失太多精度的前提下实现嵌入式设备下的算法时延的降低。

致　谢

参考文献

[1] X. Wang, T. X. Han, and S. Yan. An HOG-LBP human detector with partial occlusion handling. In ICCV, 2009.

[2] P. Viola, M. J. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. IJCV, 63(2):153– 161, 2005.

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In In CVPR, pages 886–893, 2005.

[4] D. G. Lowe. Distinctive image Features from scale-invariant keypoints. Int. J. Comput. Vision, 60(2):91–110, Nov. 2004.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In NIPS, 2012.

[6] Girshick R, Donahue J, Darrell T, et al. Rich Feature hierarchies for accurate object detection and semantic segmentation. In CVPR, 2014.

[7] He K, Zhang X, Ren S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition. In ECCV, 2014: 346-361.

[8] Girshick R. Fast R-CNN. In ICCV, 2015: 1440-1448.

[9] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 2015: 91-99.

[10] J. Redmon, S. Divvala, R. Girshick, A. Farhadi: You only look once: Unified, real-time object detection. In CVPR, 2016.

[11] Liu W, Anguelov D, Erhan D, et al. SSD: Single Shot MultiBox Detector. In ECCV, 2016.

[12] J.Redmon, A.Farhadi. YOLO9000: Better, faster, stronger. In CVPR, 2017.

[13] J.Redmon, A.Farhadi.YOLOv3: An incremental improvement. 2018.

[14] Zhou X , Wang D , Krhenbühl, Philipp. Objects as Points[J]. 2019.

[15] JI S W, XU W, YANG M, et al. 3D Convolutional Neural Networks for Human Action Recognition[C]// Proceedings of the International Conference on Machine Learning.2010:495-502.

[16] JI S W, XU W, YANG M, et al. 3D Convolutional Neural Networks for Human Action Recognition[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2013, 35(1):221-231.

[17] TRAN D, BOURDEV L, FERGUS R, et al.Learning Spatiotemporal Features with 3D Convolutional Networks[C]// Proceedings of IEEE International Conference on Computer Vision. 2015: 4489-4497.

[18] WANG L, GE L, LI R, et al. Three-Stream CNNs for Action Recognition[J]. Pattern Recognition Letters, 2017, 92(C):33-40.

[19] GIRDHAR R, RAMANAN D, GUPTA A, et al. ActionVLAD: Learning Spatio-Temporal Aggregation for Action Classification[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. 2017.

[20] DONAHUE J, HENDRICKS L A, GUADARRAMA S, et al. Long-Term Recurrent Convolutional Networks for Visual Recognition and Description [C]// The IEEE Conference on Computer Vision and Pattern Recognition. 2015 :2625-2634.

[21] WU Z X, WANG X, JIANG Y G, et al. Modeling Spatial-Temporal Clues in a Hybrid Deep Learning Framework for Video Classification[C]// Proceedings of the ACM international Conference on Multimedia (ACM MM). 2015:461-470.

[22] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers et al. Selective Search for Object Recognition. International Journal of Computer Vision, 2013, 104(2): 154-171.

[23] Girshick R. Fast R-CNN. In ICCV, 2015: 1440-1448.

[24] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 2015: 91-99.

[25] J. Redmon, S. Divvala, R. Girshick, A. Farhadi: You only look once: Unified, real-time object detection. In CVPR, 2016.

[26] J.Redmon, A.Farhadi. YOLO9000: Better, faster, stronger. In CVPR, 2017.

[27] J.Redmon, A.Farhadi.YOLOv3: An incremental improvement. 2018.

[28] Liu W, Anguelov D, Erhan D, et al. SSD: Single Shot MultiBox Detector. In ECCV, 2016.

[29] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Doll´ar. Focal Loss for dense object detection. arXiv preprint arXiv:1708.02002v2, 2018.

[30] BOBICK A F, DAVIS J W. The Recognition of Human Movement using Temporal Templates[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(1):142-158.

[31] FUJIYOSHI H, LIPTON A J, KANADE T. Real-time Human Motion Analysis by Image Skeletonization[J]. IEICE Transacions on Information and Systems, 2004, E87-D(1):113-120.

[32] YANG X D,TIAN Y L.Effective 3D Action Recognition using EigenJoints[J]. Journal of Visual Communication and Image Representation, 2014, 25(1):2-11.

[33] WANG H, ULLAH M M, KLASER A, et al. Evaluation of Local Spatio-Temporal Features for Action Recognition[C]// Proceedings of the 2009 British Machine Vision Conference. 2009:124-135.

[34] Sun K, Xiao B, Liu D, et al. Deep high-resolution representation learning for human pose estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 5693-5703.

[35] Zhang Z, Tang J, Wu G. Simple and lightweight human pose estimation[J]. arXiv preprint arXiv:1911.10346, 2019.

[36] Li J, Wang C, Zhu H, et al. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 10863-10872.

[37] Cao Z, Hidalgo G, Simon T, et al. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields[J]. IEEE transactions on pattern analysis and machine intelligence, 2019, 43(1): 172-186.

[38] SIMONYAN K, ZISSERMAN A. Two-Stream Convolutional Networks for Action Recognition in Videos[M]// Advances in Neural Information Processing Systems.Berlin:Springer,2014:568-576.

[39] JI S W, XU W, YANG M, et al. 3D Convolutional Neural Networks for Human Action Recognition[C]// Proceedings of the International Conference on Machine Learning.2010:495-502.

[40] Wang, J.; Liu, Z.; Wu, Y.; and Yuan, J.2012. Mining actionlet ensemble for action recognition with depth cameras. In CVPR. IEEE.

[41] Shahroudy, A.; Liu, J.; Ng, T.-T.;andWang, G. 2016. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In CVPR, 1010–1019.

[42] Yan S , Xiong Y , Lin D . Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition[J]. 2018.

[43] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 6848–6856, 2018. 1, 6, 9.

[44] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In Proceedings of the European conference on computer vision (ECCV), pages 116–131, 2018. 1, 3, 4, 9.

[45] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. MobilleNets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861, 2017. 1, 9.

[46] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobilleNetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE conference on computer vision and pattern recognition,pages 4510–4520, 2018. 1, 9.

[47] Chollet F . Xception: Deep Learning with Depthwise Separable Convolutions[J]. arXiv e-prints, 2016.

[48] Franc¸ois Chollet. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1251–1258, 2017. 1

[49] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016. 1, 2, 5, 6

[50] Andreas Veit, Michael J Wilber, and Serge Belongie. Residual networks behave like ensembles of relatively shallow networks. In Advances in neural information processing systems, pages 550–558, 2016. 2, 4, 8

[51] Ding X , Zhang X , Ma N , et al. RepVGG: Making VGG-style ConvNets Great Again[J]. 2021.

[52] Yu C , Wang J , Peng C , et al. BiSeNet: Bilateral Segmentation Network for Real-time Semantic Segmentation[C]// Springer, Cham. Springer, Cham, 2018.

[53] Jie H , Li S , Gang S , et al. Squeeze-and-Excitation Networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, PP(99).

[54] N.Bodla, B.Singh, R.Chellappa, and L.S.Davis. Soft-nms improving object detection with one line of code[C]//ICCV.Venice, USA:IEEE Press, 2017:5562-5570.

[55] B. Xiaohan Nie, C. Xiong, and S.-C. Zhu. Joint action recognition and pose estimation from video. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR),June 2015.

[56] U. Iqbal, M. Garbade, and J. Gall. Pose for action – action for pose. FG-2017, 2017.

[57] C. Cao, Y. Zhang, C. Zhang, and H. Lu. Body joint guided 3d deep convolutional descriptors for action recognition. CoRR,abs/1704.07160, 2017.

[58] D. C. Luvizon, D. Picard and H. Tabia, 2D/3D Pose Estimation and Action Recognition Using Multitask Deep Learning.

[59] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei, ‘Integral human pose regression’, in Proceedings of the European Conference on Computer Vision (ECCV), pp. 529–545, (2018).

[60] Cheng B , Xiao B , Wang J , et al. HigherHRNet: Scale-Aware Representation Learning for Bottom-Up Human Pose Estimation[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020.

[61] B. Xiaohan Nie, C. Xiong, and S.-C. Zhu. Joint action recognition and pose estimation from video. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR),June 2015.

攻读硕士学位期间取得的成果

[1] 项目：******技术研究研制（涉密），项目编号：******，2018 年

[2] xxx(7).A lighten CNN-LSTM model for speaker verification on embedded devices[J]. Future Generation Computer Systems, 2019, 100: 751-758.