大连理工大学本科毕业论文（设计）

基于 Wi-Fi 和视觉的多模态行为识别方法研究

Research on Multi-modal Human Activity Recognition Method Based on Wi-Fi and Vision

学 院（系）：软件学院

专　　业：网络工程

学 生 姓 名：张亦弛

学　　号：201992222

指 导 教 师：徐秀娟

评 阅 教 师：赵小薇

完 成 日 期：2023/5/28

大连理工大学

Dalian University of Technology

原创性声明

本人郑重声明：本人所呈交的毕业论文（设计），是在指导老师的指导下独立进行研究所取得的成果。毕业论文（设计）中凡引用他人已经发表或未发表的成果、数据、观点等，均已明确注明出处。除文中已经注明引用的内容外，不包含任何其他个人或集体已经发表或撰写过的科研成果。对本文的研究成果做出重要贡献的个人和集体，均已在文中以明确方式标明。

本声明的法律责任由本人承担。

作者签名：　　日　期：

关于使用授权的声明

本人在指导老师指导下所完成的毕业论文（设计）及相关的资料（包括图纸、试验记录、原始数据、实物照片、图片、录音带、设计手稿等），知识产权归属大连理工大学。本人完全了解大连理工大学有关保存、使用毕业论文（设计）的规定，本人授权大连理工大学可以将本毕业论文（设计）的全部或部分内容编入有关数据库进行检索，可以采用任何复制手段保存和汇编本毕业论文（设计）。如果发表相关成果，一定征得指导教师同意，且第一署名单位为大连理工大学。本人离校后使用毕业毕业论文（设计）或与该论文直接相关的学术论文或成果时，第一署名单位仍然为大连理工大学。

论文作者签名：　　日　期：

指导老师签名：　　日　期：

摘　要

人体行为感知在对老年人摔倒行为检测等领域有着举足轻重的作用。传统的非接触式人体行为感知模型只能在一些理想条件下表现良好，但是在特殊场景例如光线变暗或物体遮挡时，其性能通常会显著下降，进一步影响了其泛用性。

本文提出了 ViFi 多模态感知模型，通过视觉与 Wi-Fi 协同处理来解决特殊场景下的人体动作识别问题。ViFi 感知模型使用基于 YOLO 的视觉目标检测预处理与由卷积神经网络和循环神经网络组成的视觉感知模块提取视频摄像头采集的样本特征，通过基于多尺度的时空域卷积网络对 Wi-Fi 的信道状态信息（Channel State Information，CSI）进行特征提取，最后将两种模态的特征采取直接相联的方式融合。

与先进的 GaitFi CRNN 模型相比，ViFi 感知模型在视觉模态上提升了 1.56% 至 7.81%，Wi-Fi 的感知能力提高幅度在 3.65% 至 11.46% 之间，并在极端场景下的动作识别效果有超过 20% 的提升。此外，实验将多种场景下的数据合并混杂训练，得到了适用于更复杂场景的大型动作识别跨域模型预训练权重，在不同光照度和物体遮挡的场景中取得了较好效果，并且在未经学习过的新场景下也实现了 86.98% 的识别准确率。实验证明感知模型具有了更先进的识别效果和更高的鲁棒性和泛用性，未来可以被应用在智能家居、安防监控、数字娱乐等领域。

关键词：多模态；人体行为感知；深度学习

Research on Multi-modal Human Activity Recognition Method Based on Wi-Fi and Vision

Abstract

Human behavior perception plays a pivotal role in fields such as the detection of elderly fall behavior. Traditional non-contact human behavior perception models perform well only under ideal conditions, but their performance often declines significantly in special scenarios, such as when light dims or objects obstruct the view, which further impacts their versatility. This paper presents the ViFi multimodal perception system, which solves the problem of human motion recognition in special scenarios through collaborative processing of vision and Wi-Fi.

The ViFi system uses a vision target detection pre-processing based on YOLO and a vision perception module composed of convolutional neural networks and recurrent neural networks to extract sample features collected by video cameras. It also uses a multi-scale spatio-temporal convolutional network to extract features from Wi-Fi Channel State Information (CSI). Finally, the features of both modalities are fused directly.

Compared with the advanced GaitFi CRNN model, the ViFi perception model has improved the visual modality by 1.56% to 7.81%, and the Wi-Fi perception ability has improved by 3.65% to 11.46%. Furthermore, it has achieved more than a 20% improvement in action recognition under extreme scenarios. In addition, the experiment merged data from multiple scenarios for mixed training, obtaining pre-training weights for a large-scale action recognition cross-domain model suitable for more complex scenarios. It achieved good results in scenarios with different light intensities and object obstructions, and also achieved an 86.98% recognition accuracy rate in new scenarios that have not been learned before.

Experiments have proven that the system has more advanced recognition effects and higher robustness and versatility. In the future, it can be applied in areas such as smart homes, security monitoring, and digital entertainment.

Key Words：Multi-modal; Human Activity Recognition; Deep Learning

目　　录

1 文献综述

1.1 研究背景及意义

近年来，人口老龄化问题日益凸显，成为社会关注的焦点之一。随着老年人口的增加，老年人的健康和安全问题日益凸显，其中老年人的摔倒风险成为一个重要的议题。老年人的摔倒不仅可能导致身体受伤，还可能引发其他严重的后果，如骨折、低落心情、社交隔离等。因此，及早检测老年人的摔倒行为，并及时采取救援措施，具有重要的临床和社会意义。

在此背景下，人体行为识别（Human Activity Recognition，HAR）作为一种关键技术，引起了广泛的关注。HAR 旨在通过识别和分析人体的动作和行为，实现对人的行为状态的理解和判断。如图 1.1 所示，通过 HAR 技术，可以实时监测老年人的行为，包括日常活动、步行、坐卧姿势以及摔倒等，从而及时发现摔倒事件并采取紧急救援措施。

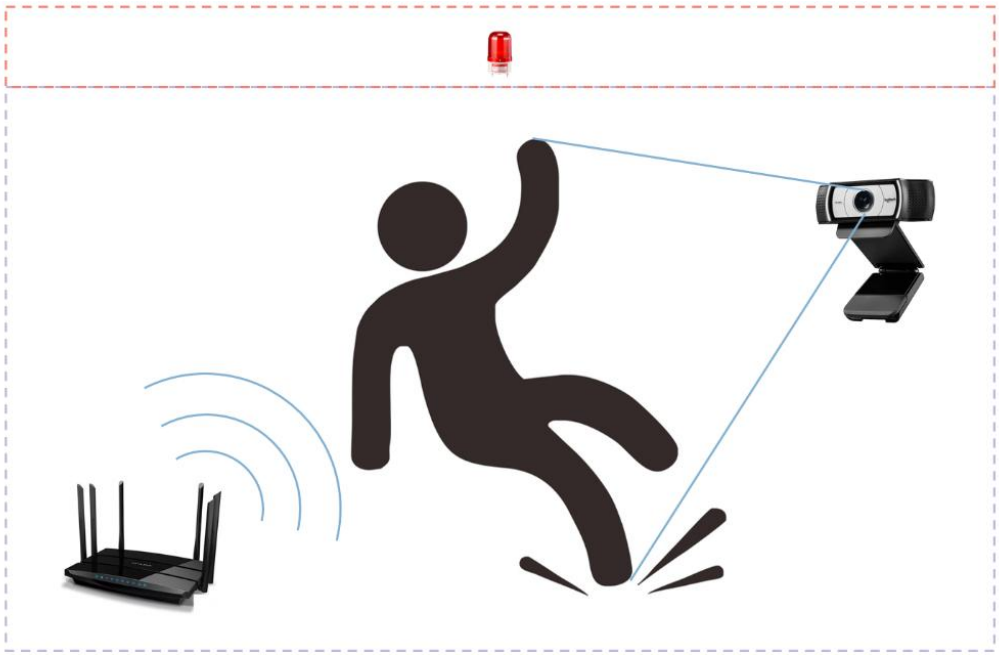


图 1.1 HAR 技术检测老年人行为并预警示意图

随着科技的不断进步和应用的广泛推广，人们对于智能感知和交互计算的需求不断增加。国家自然科学基金将面向真实世界的智能感知与交互计算列为信息科学部门的重点发展领域，这进一步证明了人体行为研究的重要性。此外，智能监测系统在公共场所

的部署也日益普及，旨在通过提取人体行为和身份识别相关的生物特征信息，加强对人群活动的监测和管理。

目前，人类识别技术如指纹识别、人脸识别和可穿戴设备等在人体行为识别领域得到了广泛应用。然而，这些技术在实际应用中存在一定的局限性。例如，指纹和虹膜识别的感应范围有限，人脸识别受到口罩遮挡的影响，可穿戴设备需要特殊设备的佩戴和频繁充电，存在使用不便和易损坏的问题。

与传统的人体行为感知技术相比，非接触式感知技术具有独特的优势。非接触式感知技术不需要人们佩戴任何感知设备，可以实现对人体行为的自动感知和识别。然而，这些非接触式感知技术在实际应用中也存在一些问题，如对光线和视线路径等工作条件的依赖，导致高成本和低效率，限制了其在日常生活场景中的应用。

因此，本研究旨在解决传统感知技术和非接触式感知技术在人体行为感知方面的局限性。本文将利用视频和 Wi-Fi 技术的多模态融合，提高人体行为感知的准确性和鲁棒性。通过结合视频和 Wi-Fi 所获取到的信息，可以克服一些特殊条件下（如暗光、遮挡）的人体动作感知度较低的问题。

该研究的成果具有广泛的应用前景。在家庭场景下，人体行为感知可以对屋主的健身情况和动作进行辅助指导，当有人摔倒时可以主动报警并及时送医。在监狱等场景下，人体行为感知系统可以用于检测是否有人跨越了限制区域以及是否发生暴力行为。在虚拟现实场景下，通过人体行为感知可以更好地捕捉游戏玩家的动作，从而提供更加真实和沉浸的游戏体验。

综上所述，本研究的研究背景和意义在于解决以老年人摔倒检测为典型的社会问题，推动人体行为识别技术的发展与应用，并为智能感知和交互计算领域的研究提供理论基础和实践应用。同时，通过多模态融合的方法，本研究试图克服传统感知技术和非接触式感知技术的局限性，提高人体行为感知的准确性和适用性，为相关领域的科学研究和工程应用做出贡献。

1.2 国内外研究现状

对于人体动作的感知和理解，通常有识别、定位和预测三个方向。通常来说，识别是感知系统的基础，通常通过各种检测系统实现对物体的定位和预测时，都需要先以实现识别为根基，这也是人体动作感知和理解最重要的组成部分。人体行为感知，在这里也即为人体动作识别，通常是将通过 Wi-Fi 天线、摄像头、可穿戴设备的陀螺仪等数据采集设备在人体进行活动时采集到的数据或持续采集的数据发生的变化，通过一些方法具体的呈现和放大，直到能通过数学或计算机的方式将不同的动作收集到的数据或数据

变化进行分类和识别，并对应到不同的动作上。如果动作实例跨越输入的整个长度，则称为修剪动作识别；如果动作实例没有跨越输入的整个长度，则称为未修剪动作识别。

在人体行为感知的研究初期，人们只能通过肉眼来观察和识别画面，这无法显著的提高效率，也收到了人工的限制。随着计算机技术和人工智能的发展，对于人体行为的感知和理解已经可以逐渐交给计算机来完成。越来越多基于机器学习和深度学习的多种数据采集载体的识别方式开始研究和应用，它们不仅提高了识别和感知的效率，也通过修改模型、增加模态等方式提高了对人体行为感知的准确率。但是由于人体动作的复杂性，以及对于一些特殊条件如较低的光线照明度、人物肢体的遮挡等条件、摄像头低像素或模糊等情况，对人体行为的感知研究仍然无法达到高可用的程度。

1.2.1 基于视频的人体动作识别

在处理人体行为感知方面，基于视觉进行深度学习的解决方案是最常用、最直接的方法[1]。人类活动识别技术使用外观和运动特征来学习视频的特征。Moez 等人[2]提出了一个完全的深度模型，该模型无需使用任何先前的知识来学习空间时间特征，并将卷积神经网络扩展到了 3 维。递归神经网络有一个长短期记忆单元的隐藏层，被训练来对总共 10 层的学习模型进行分类。Graham 等人[3]提出了使用 CNN 进行特征学习的概念。通过使用门控限制波尔兹曼机从图像序列中学习静态和动态特征，并以光流形式表示。Shuiwang 等人[4]提出了 3D CNN 架构，监督深度架构从相邻的输入帧生成多个信息通道，并在每个通道中进行卷积和子采样。最终的特征表示是通过结合所有通道的信息获得的。周升儒等人[5]使用基于 ResNet-50 姿态估计模型对网球运动视频进行人体目标检测并提取骨骼关键点并进行 PoseC3D 模型训练，相较于基于图卷积网络的方法具有更强的泛化能力。

在后来的研究中人们不再把目光局限于简单的 CNN 架构。Joe Yue-Hei Ng[6]提出了一种基于 LSTM 在 GoogleNet 和 AlexNet 上学习长期视频的空间时间特征的方法。原始帧和光流都被用作输入模式，而特征池被应用于动作识别的类分融合。Limin 等人[7]提出了轨迹池化的深度卷积描述器（TDD），具有手工制作和深度学习特征的优点。训练后的 CNN 被 TDD 使用，随后 Fisher 向量被用来编码它们，SVM 被用作分类器。CNN 在图像方面取得里程碑式的成就后，被扩展到 100 万个视频[8]，有 487 个类别，被命名为体育 1M 数据集。此外，两个空间分辨率的低分辨率上下文流和高分辨率眼窝流被用来实现更好的结果并减少训练时间。为了验证其他具有挑战性的数据集的结果，转移学习的概念被用于数据集 UCF 101 和缓慢融合。张执着[9]提出了一种姿态驱动的超分辨率(Pose-Driven Super-Resolution，PDSR)重建方法。该方法将姿态估计网络作为判别器，

以驱动 SR 网络隐式地学习对姿态估计具有高判别性的图像特征，并且利用 BASR 保持像素相似性和增强人体区域的信息表达。Fuqiang Gu 等人[10]提出了一种使用多个智能手机传感器如陀螺仪、加速度计、磁力计的数据来识别运动活动的深度学习模型，它的叠加去噪自动编码器使用了消除专家知识的训练。Su 等人起草了一个基于 RNN 的方法[11]，用于使用深度相机的健康和社会护理服务。身体上随时间变化的多个关节以时空矩阵的形式表示，随后对 RNN 进行了训练，并在病房中用于测试目的。与以前使用的方法相比，这种方法取得了最先进的成果。

但是以上用于从视频或图像序列中提取特征并进行识别都只能运行在正常条件，在只使用视频单模态时，在光线变暗或有一定遮挡的情况下无法进行稳定的识别。

1.2.2 基于 Wi-Fi 的人体动作识别

随着对于隐私性以及特殊场景的需求越来越高，许多研究开始将目标转成了 Wi-Fi。Yousefi 等人[12]首先利用 LSTM 来提取 Wi-Fi 信号的特征。然而，这种传统的 LSTM 只能处理前向的 Wi-Fi CSI 读数，一些特征可能会丢失。Chen 等人提出了基于注意力的双向 LSTM（ABLSTM），通过处理两个方向的 CSI 读数来学习特征[13]。Cao 等人采用双向 RNN 从睡眠时身体运动的 Wi-Fi 信号中学习背景信息，引入了带有 Balloon 机制的深度神经网络算法来处理大内存消耗的问题，通过扩大跨层充气通道的宽度来降低训练的复杂性[14]。由于 Wi-Fi 信号带宽较窄且时间分辨率低，导致基于 Wi-Fi 的人体感知技术在大尺度动作感知方面由于信号特征非独立于背景环境导致大尺度动作感知模型跨域能力差，在小尺度动作感知方面目前基于移动终端的人体呼吸感知模型尚不完善[15]。Lin 等人提出了一个名为 WiWrite 的自步调密集卷积网络，用于一应俱全的细粒度手指手写识别[16]。Widar3.0 提出了身体坐标速度曲线（BVP）的概念。它结合 CNN 和门控递归单元来提取 BVP 特征，用于识别跨域手势[17]。张东恒[18]提出了一种基于 Wi-Fi 的人体呼吸追踪技术，基于稀疏重构的弱信号提取方法，能够有效挖掘和聚合 Wi-Fi 信号中存在的信息，抑制多径和噪声等干扰，实现对人体呼吸状态的细粒度分辨。此外，对抗网络、转移学习等新技术也被引入到无线传感领域[19,20]。

Wi-Fi 单模态识别一直在致力于提高识别效果，但是在多数场景下都无法稳定的达到接近视频单模态的识别率。

1.2.3 多模态融合的人体动作识别

多模态学习是一种建模方法，旨在通过融合、共同学习和其他方法来处理和联系多种传感模式的信息，从而提取其新颖性。多模态机器学习的动机来自于这样一个事实，即单模态模型有其自身的缺点，使得它们的表现不尽如人意，例如，当图像在光照、相

机角度或背景杂波方面有问题时，基于视觉的模型就不能很好地工作。在这种情况下，来自不同模式的数据集合被证明是有益的。此外，人类对世界的体验是多模态的，有视觉、听觉、语言和嗅觉感受器等，这鼓励了多模态学习的引入。

有很多研究已经在尝试通过视觉和 Wi-Fi 的多模态融合来提高对人体行为感知的识别效果。H Zou 等人[21]提取了细粒度的 Wi-Fi 通道信息将它们转换为 Wi-Fi 帧并设计了一个特别的卷积神经网络模型用于在 Wi-Fi 帧中提取高级代表性特征，以便提供人类活动的估计。Alamgir 等人[22]设计的多模态情感识别框架 InceptionV3DenseNet，从视频中提取诸如镜头长度、照明关键、运动和颜色等特征，从音频中提取零交叉率、梅尔频谱系数（MFCC）、能量和音调，从文本中提取单词、大词和 TF-IDF，使用多组综合典型相关分析（MICCA）进行融合，模型有很大的提升。Huang Z 等人[23]提出了一种新的动态内外模态注意力（DIIA）模型，有效地融合了音频和文本两种模态信息。其中的多模态知识蒸馏（MKD）模块，对多模态 MC 模型能够仅基于文本或音频准确预测答案有很大帮助。P Gao 等人[24]通过内部和跨模态信息流动态融合多模态特征，这种方法可以强大地捕获语言和视觉领域之间的高级交互，从而显著提高视觉问题回答的性能。但是以上多模态融合的讨论都局限在正常的环境和条件下，L Deng 等人[25]开始讨论多模态感知模型在低光照场景中的步态感知识别，但是没有对遮挡场景进行测试。

多模态机器学习涉及两个主要的技术，即融合和协同学习，这在本质上有助于提高它的新颖性。融合包括在特征或决策层面结合来自多种传感模式的信息，以进行预测。基于特征的融合在不同模式的特征被提取后立即进行整合，而基于决策的融合则在每个模式做出决策（如分类或回归）后进行整合。基于特征的融合学会利用不同模态的低级特征之间的相关性，并且由于它涉及到训练一个单一的模型，所以训练起来很简单。另一方面，基于决策的融合使用一些机制，如平均法、投票方案或学习的模型来融合单模态决策。回过头来看，基于决策的模型有几个重要的优点。首先，它们允许不同的模型用于不同的模态，从而具有更大的灵活性。其次，它们对来自单一或多个模式的数据损失更加稳健。协同学习通过利用其他资源丰富的模态的知识，使资源贫乏的模态（缺乏注释的数据，嘈杂的输入）的模型化。它通过利用最先进的迁移学习和领域适应方法实现这一能力[26,27]。目前的机器学习算法在很大程度上依赖注释数据进行训练，而这些数据的获得需要大量的人力投入。因此，关于改进无监督学习方法的讨论已经很多了。在这种情况下，共同学习证明了其存在的价值。

1.3 本文研究内容与结构

本文旨在探索一种基于 Wi-Fi 和视觉技术的多模态感知方法，使其能够适应各种特殊环境，无论是在不同光照条件下，还是在环境中存在不同程度的遮挡情况下，都能够达到高精度和较高实时性的感知和识别人体行为的效果。为了实现这一目标，本文提出了一种新颖的多模态融合模型。该模型在现有的 GaitFi 模型的基础上，进行了诸多创新和改进，在许多特殊场景下都显示出优异的性能。

（1） 本文提出了 ViFi 模型，对视频摄像头采集到的样本增加目标检测作为数据预处理的一部分，在单视频模态的训练上都有了不同程度的提升，测试结果准确率提高幅度从 1.56%到 7.81%不等，使得多模态模型更加稳定，并在极端场景下有超过 20%的准确率提升。

（2） 本文使用对时空域分别扩散卷积的时空域卷积网络[28]替换了 CRNN 模型中对于 Wi-Fi 数据的处理，在单 Wi-Fi 模态的训练上提升幅度在 3.65%到 11.46%不等，对多模态模型的提升在不同遮挡的场景下提升约为 2%，不同光照的场景提升约为 1%。

（3） 本文尝试了多种模态融合方式（直接相联、相加、加权）并在不同场景下进行消融实验，并最终选定出最为稳定的融合方式。根据单模态的效果进行调整的权重加权融合通常会取得更为优秀的效果，直接相联相比最优加权的准确率降低幅度通常不超过 1.5%。

（4） 本文最终将多种场景下的数据进行合并混杂训练，最终得到可以应对多种复杂场景的大型识别模型预训练权重，在特殊光线场景下准确率超过 93.75%，在两种遮挡场景的识别准确率均为 95.833%。

本文的剩余部分组织如下：第 2 章从 Wi-Fi 的 CSI 数据、视频的采样和编码、深度学习和多模态融合三个方面简要介绍论文中涉及的相关技术；第 3 章对本文提出的 ViFi 模型进行详细的介绍；第 4 章主要介绍以 LRM 预训练跨域模型；第 5 章将对比 ViFi 模型的视频和 Wi-Fi 两个模态以及在视频预处理、Wi-Fi Backbone 模型、模态融合方式前后的对比实验过程和结果，并对未来进行展望；最后一章对本文进行总结陈述。

2 相关技术概论

本研究主要通过收集 Wi-Fi 的 CSI 数据和摄像头的视频数据，并通过深度学习进行多模态的融合，从而生成对人体行为感知的训练模型，最终达到对新的人体动作数据的预测。

2.1 Wi-Fi 信道状态信息

在无线通信领域，基于 Wi-Fi 的 CSI（Channel State Information，信道状态信息），指的是已知通信链路的信道属性，它描述了信号从发射器到接收器传播的方式，以及散射、衰减和距离影响等效应的综合影响。通过信道估计的方法获取 CSI，可以使传输适应当前的信道条件，这对于实现多天线系统中高数据速率的可靠通信非常重要。通常情况下，CSI 需要在接收器端进行估计，进行量化并反馈给发射器。因此，发射器和接收器可能会有不同的 CSI，发射器的 CSI 和接收器的 CSI 分别被称为 CSIT 和 CSIR。

通常无线通信领域将 CSI 分为两个层次：瞬时 CSI 和统计 CSI。瞬时 CSI（或短期 CSI）指的是已知当前的信道条件，相当于已知数字滤波器的脉冲响应。这使得传输信号可以适应当前的脉冲响应，从而优化接收信号的空间复用或实现低误码率。统计 CSI（或长期 CSI）指的是已知信道的统计特征。这些特征可以包括衰减分布类型、平均信道增益、视线成分和空间相关性。与瞬时 CSI 类似，统计 CSI 也可以用于传输优化。但是，CSI 的获取受到信道条件变化速度的限制。在快速衰减系统中，信道条件在单个信息符号传输下迅速变化，只有统计 CSI 是可行的。另一方面，在慢速衰减系统中，瞬时 CSI 可以以合理的精度进行估计，并在过期前用于传输适应。在实际系统中，可用的 CSI 通常处于这两个层次之间，即具有一定的估计/量化误差的瞬时 CSI 与统计信息相结合。

基于 802.11n 协议的 Wi-Fi 技术采用 MIMO-OFDM 系统。在一个具有多个发射和接收天线（Multiple Input Multiple Output，MIMO）的窄带平消信道中，系统被建模为

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{n} \tag{2.1}$$

其中，**y** 代表接收端收到的信号向量，**x** 代表发射端发出的信号向量，**H** 和 **n** 分别代表信道矩阵和噪声。噪声通常被建模为圆形对称复数正态，被表示为

$$\boldsymbol{n} \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{S}) \tag{2.2}$$

其中均值为零，噪声协方差矩阵为 **S** 是已知的。对于瞬时 CSI 而言，理想情况下，通道矩阵 **H** 是完全已知的。由于信道估计的误差，信道信息可以表示为

$$vec\,(\boldsymbol{H}_{estimate}) \sim \mathcal{CN}\big(vec\,(\boldsymbol{H}), \boldsymbol{R}_{error}\big) \tag{2.3}$$

其中，$\mathbf{H}_{\text{estimate}}$ 是信道估计，而$\mathbf{R}_{\text{error}}$ 是估计误差协方差矩阵。矢量化 vec() 被用来实现列叠加的 **H**，因为多变量随机变量通常被定义为矢量。而对于统计 CSI，**H** 的统计数据是已知的。在一个雷利消逝信道中，对于一些已知的通道协方差矩阵 **R**，这相当于知道

$$vec(\boldsymbol{H}) \sim \mathcal{CN}(0, \boldsymbol{R}) \tag{2.4}$$

通过收集路由器通过一条或多条 Wi-Fi 天线发射的 CSI 数据，使用 CSI tools 对数据包进行解码和分析。与 MAC 层的信号强度（Received Signal Strength Indicator, RSSI）相比，CSI 描绘了子载波层面上的无线通信链路的信道衰减程度，包括信号散射、环境衰减和距离衰减，它可以同时测量多个子载波的频率响应，从而更全面地刻画了多径传播。因此，CSI 在幅度和相位上可以提供无线链路更细粒度的时间和频谱结构，比单一的 RSSI 信息更加丰富。对于多输入多输出正交频分复用的无线网络系统，信道状态信息是一个三维复值矩阵，可以表示多径无线网络信道的幅度衰减和相位。这使得 CSI 在多天线系统中发挥着重要的作用，能够更好地适应复杂的无线信道环境，提高系统的数据传输速率和可靠性。

具有 **M** 个发射天线、**N** 个接收天线和 ϕ 个子载波的 CSI 是一个矩阵 $\mathbf{H} \in \mathbf{R}^{\mathbf{M}\times\mathbf{N}\times\phi}$，表示多路径信道的振幅衰减和相移。CSI 矩阵的时间序列从不同的领域，即空间（$\delta = \mathbf{M}\times\mathbf{N}$ 天线对/信道）、频率（ϕ 子载波）和时间 t（收集数据包的数量）来描述 MIMO 信道的变化。因此，第 i 个通道的第 j 个子载波在时间 t 的 CSI 可以定义为

$$\boldsymbol{H}_t^{i,j} = \boldsymbol{A}_t^{i,j} e^{j\phi_t^{i,j}} \tag{2.5}$$

其中，$\mathbf{A}_t^{i,j}$ 和 $\boldsymbol{\phi}_t^{i,j}$ 分别代表振幅和相位。

2.2 视频的采集及编码

视频采样是指将连续的视频信号转化为离散的数字信号的过程。视频编码是将采样得到的数字信号进行压缩编码，以减小数据量并提高传输效率。本文研究的实验通过摄像头采集了不同条件下的视频数据，并经过数据预处理后输入给模型。

从扫描方式上看，视频采样可以分为逐行扫描和隔行扫描两种方式。逐行扫描是按照行的顺序逐个采样每个像素，优点是采样的像素数量多，图像清晰度高。隔行扫描是按照偶数和奇数行交替采样，优点是采样速度快，数据量小。目前，逐行扫描已成为主流采样方式。从视频编码上看，早期的视频编码主要采用无损编码方式，这种编码方法可以完全还原原始视频信号，但数据量较大，传输效率低。20 世纪 80 年代中期，出现了基于变换的编码方法，例如离散余弦变换（DCT），它可以把视频信号从时域变换到

频域，再进行压缩编码，数据量明显减小，传输效率得到提高。到了 20 世纪 90 年代，运动估计编码方法开始流行，它通过对相邻帧之间的运动进行估计，来减少编码时需要传输的信息量。常见的运动估计编码方法有 MPEG-1、MPEG-2 等。随着网络技术和存储技术的不断发展，视频编码标准也得到了不断的更新迭代。当前最常用的视频编码标准是 H.264/AVC 和 H.265/HEVC，它们采用了更加先进的编码技术，例如运动补偿、帧内预测和变长编码等，可以在保证视频质量的同时，进一步提高传输效率和压缩比。

在对视频的处理问题上，主流思路分为了帧采样和光流法两种。帧采样是将视频以时间为维度，将视频的每一帧分离。从这个角度上分析，视频效果等同于直接将照片连续播放快速变换，可以用公式表示为

$$\mathcal{M} = R_{fps} \times t \tag{2.6}$$

其中，R_{fps} 和 t 分别代表视频的帧率（frame per second）和视频片段的时间长度，$\mathcal{M}$ 代表等效的照片张数，一段长度为 3 秒、每秒 30 帧的视频等同于将其转换为 90 张连续的照片。另一种思路是光流法。光流（Optical Flow）是一种通过计算视频帧沿水平、竖直和时间方向的梯度（Gradient），推断视频帧中像素移动方向和速度的方法。光流一般用箭头来表示，箭头的方向指示了像素移动的方向，箭头的长度表示像素移动的速度。可以预想在处理连续运动的图像时，通过光流法就可以将其视为照片上的物体或像素在进行坐标的移动，这就为连续图片转换成为向量创造了条件。通过将视频以光流法的角度分析，将其中移动的物体转换为持续变换的向量，就可以更好的反映出物体的运动状态等信息。但是这也造成一些例如遮挡、光照变化等的复杂场景无法很好的转换为向量，对模型的数据造成影响。

时间建模问题是视频动作识别的关键，主要涉及短距离运动信息建模和长距离运动信息建模。短距离运动信息（Short-range motions）指相邻帧之间的信息，而长距离运动信息（Long-range aggregations）则涉及到长期特征融合。时序信息的表达和获取是当前视频动作识别的难点和研究热点。

2.3 深度学习和多模态融合

深度学习是人工智能的一个重要子领域，它以神经网络为基础，特别是以多层感知器（Multilayer Perceptron，MLP）为基础，进一步发展出一系列复杂、深度的网络结构。这些网络能够自动从大量数据中学习并提取有用的特征，这种自我学习的能力使得深度学习在许多复杂任务中表现出色，如图像识别、语音识别、自然语言处理等。它的核心思想是使用多层非线性处理单元进行特征提取和变换，每一层都使用前一层的输出作为输入。这种分层的结构使得深度学习模型能够处理非常复杂的数据，并且可以捕捉到数

据中的高级抽象特征。这一点与依赖于人工设计的特征提取器的传统的机器学习方法有明显的区别。深度学习模型的训练依赖于反向传播（Backpropagation）和梯度下降（Gradient Descent）等优化方法。反向传播是一种高效的方法，用于计算神经网络中每个参数的梯度，而梯度下降则用于更新这些参数以最小化损失函数。尽管深度学习在许多领域都取得了显著的成功，但它依然面临着一些挑战。首先，深度学习模型通常需要大量的标记数据进行训练。对于一些领域来说，获取这些数据可能非常困难。此外，深度学习模型的训练过程往往需要大量的计算资源，并且训练过程可能需要很长的时间。这些因素都限制了深度学习的应用。此外，深度学习模型通常被认为是“黑箱”模型，因为它们的内部工作机制往往难以理解。这种缺乏可解释性可能在一些需要高度透明性和可解释性的领域（如医疗诊断和金融）中导致问题。因此，如何提高深度学习模型的可解释性，是当前研究的一个重要方向。最后，虽然深度学习在处理固定任务时效果显著，但它的泛化能力仍然有待提高，每当面对一个新的场景时，就需要重新收集数据并训练，这使得它仍然没有足够高的普适性。

基于多模态的深度学习是一种建模方法，其目的是通过融合、共同学习和其他方法来处理和联系多种传感模式的信息，从而提取它们的新颖性。多模态机器学习的动机来自于这样一个事实，即单模态模型有其自身的缺点，使得它们的表现不尽如人意，例如，当图像在光照、相机角度或背景杂波方面有问题时，基于视觉的模型就不能很好地工作。在这种情况下，来自不同模式的数据集合被证明是有益的。此外，从仿生学的角度来看，人类对世界的体验是多模态的，有视觉、听觉、语言和嗅觉感受器等，这使得多模态融合的深度学习广受关注。

深度学习在多模态环境中的应用主要依赖于两种核心技术：融合与协同学习，这两者都有助于提升模型的创新性。融合过程涉及整合多种传感模态的信息，这些信息可能位于特征或决策层面，用以进行预测。特征层面的融合在从不同模态提取特征后立即进行，而决策层面的融合则在每种模态完成其决策（例如分类或回归）后实施。特征层面的融合利用了不同模态间的低级特征相关性，并且因为只需训练一个模型，所以实现相对简单。另一方面，决策层面的融合采用了诸如平均法、投票策略或学习模型等机制来整合各模态的决策。然而，决策层面的模型具有几个显著的优点。首先，它们允许为不同的模态使用不同的模型，从而提供更大的灵活性。其次，这些模型对于来自单一或多种模态的数据丢失具有更高的稳健性。协同学习则通过引入其他模态的知识，以帮助那些资源匮乏（如缺乏标记数据或有噪声输入）的模态进行建模，这是通过应用最新的迁移学习和领域适应方法来实现的。

3 ViFi 感知模型的设计

本章主要介绍 ViFi 感知模型的总体设计，包括了基于 YOLO 的视觉目标检测预处理模块、视觉感知模块、基于多尺度的 Wi-Fi 无线感知模和特征融合模块的介绍，最后介绍本文提出的大型跨域感知模型预训练。

3.1 ViFi 感知模型总体架构

本文提出的感知模型由视频和 Wi-Fi 共同协作完成，二者在不同场景下相辅相成，所以将其命名为 ViFi（Video and Wi-Fi）。ViFi 感知模型的总体架构如图 3.1 所示。预处理模块将 Wi-Fi CSI 数据进行解码和解析，提取天线接收到的振幅信息并处理成 $3 \times 30 \times 1000$ 的高维矩阵；将摄像头捕获的视频流以帧为单位使用 YOLO 目标检测模型进行预处理，检测和裁剪出人体正方形图片。ViFi 将处理好的两种模态数据分别输入到基于多尺度的 Wi-Fi 无线感知模块和视觉感知模块获取到 Wi-Fi 特征和视觉特征，最终在特征融合模块进行特征融合，并通过全连接层进行分类和映射，输出的结果为模型识别出的动作类型（在本文实验中共有 8 种）。

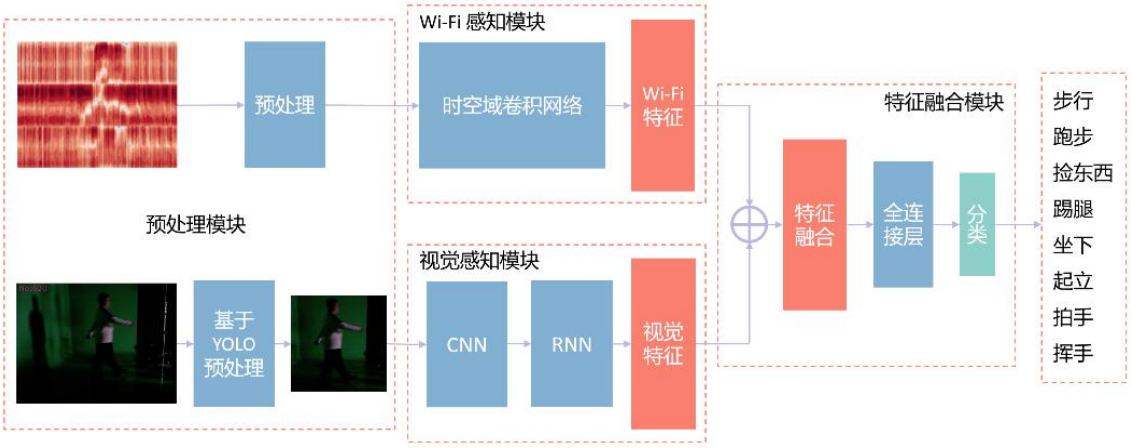


图 3.1 ViFi 感知模型总体架构示意图

3.2 基于 YOLO 的视觉目标检测预处理

为了能让视频模型更好的聚焦到正在进行动作的人体上，消除或减弱周围环境的变化对模型识别的影响，同时对采集的数据集数据增强，如进行清洗、筛选和检测，本文的 ViFi 模型使用了 YOLOv5 视觉目标检测模型对摄像头采集到的视频数据进行预处理，预处理流程如图 3.2 所示。

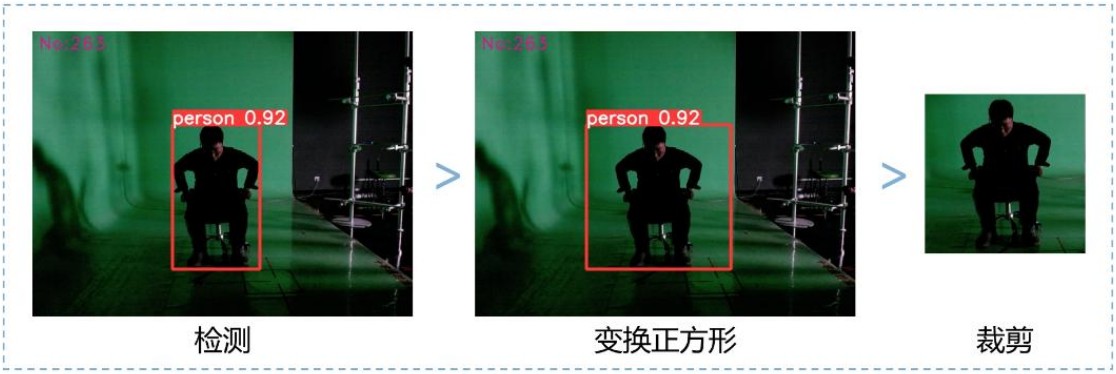


图 3.2　数据预处理部分中的 YOLO 目标检测

YOLO（You Only Look Once）是一种实时目标检测算法，由 Joseph Redmon 等人于 2016 年提出[29]。与传统的目标检测方法不同，YOLO 将目标检测视为单一的回归问题，直接从图像像素预测出目标的边界框和类别概率。这使得 YOLO 在目标检测时具有较高的速度，尤其适用于实时应用。YOLO 模型使用一个单一的卷积神经网络（CNN）对输入图像进行处理。这个网络将图像分割成 $S \times S$ 个网格，如果某个对象的中心落入某个网格中，那么这个网格就负责预测这个对象。每个网格预测 B 个边界框和这些边界框的置信度，以及 C 个类别的条件概率，其中 S 和 B 被设定为固定值，在原始的 YOLO 中，$S = 7$，$B = 2$，C 取决于数据集的类别数量。

3.3　视觉感知模块

在基于视频的人体动作识别深度学习网络模型中，最常使用的是卷积神经网络(Convolutional neural network, CNN)[25,30-32]和 C3D (3DCNN) [21,33-35]型，也有一些通过注意力机制[36]、生成对抗网络[37]等方式进行特征学习的例子。CNN 是一类最常用于分析视觉图像的人工神经网络，它也被称为移位不变或空间不变人工神经网络（SIANN），基于卷积核或滤波器的共享权重结构，沿着输入特征滑动，通过滑动窗口与卷积核进行矩阵的卷积运算来生成结果，如图 3.3 所示。

卷积神经网络由输入层、隐藏层和输出层组成。输入层用于接收原始数据，如图像矩阵。在隐藏层中，卷积层可以产生一组平行的特征图（Feature Map），它通过在输入图像上滑动不同步长 Z_S 的卷积核并将两个矩阵进行卷积运算组成，即卷积核与输入图像之间会执行一个元素对应乘积并求和的运算以将感受野内的信息投影到特征图中的一个元素。一张特征图中的所有元素都是通过一个卷积核计算得出的，也即一张特征图共享了相同的权重和偏置项。线性整流层（Rectified Linear Units layer，或激活层）可以

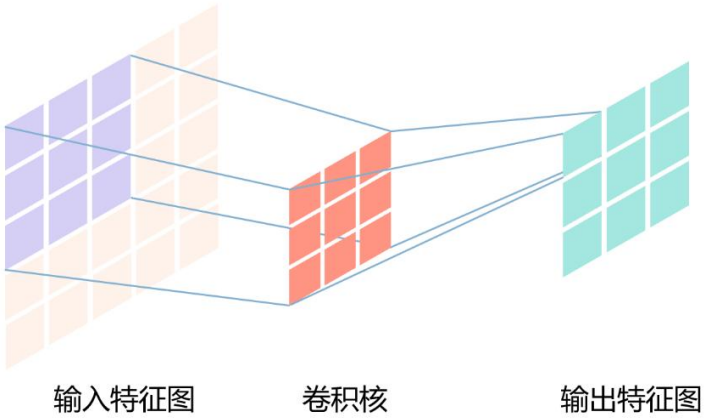


图 3.3　卷积核到特征图映射过程示例

对卷积层的输出应用非线性激活函数，如 ReLU、tanh 和 Sigmoid，它可以增强判定函数和整个神经网络的非线性特性，而本身并不会改变卷积层。这使得神经网络可以学习复杂的非线性函数。池化层是一种非线性形式的降采样，通过降低数据的空间维度，以减少计算复杂性和过拟合。常见的池化方法有最大池化（Max Pooling）和平均池化（Average Pooling）。在经过几个卷积和最大池化层之后，神经网络中的高级推理通过完全连接层来完成，它将前一层的输出压缩成一个一维向量，并通过全连接的神经元计算高级特征。它被用于进行分类或回归任务。使用卷积神经网络 CNN，卷积层可以学习局部特征，这意味着网络可以识别任意位置的特征；卷积核在不同位置共享参数，降低了模型的参数数量，减少了过拟合的风险；通过使用不同大小的卷积核和池化操作，CNN 可以在多个尺度上处理信息。

在对视频模态的学习模型中，为了能在兼顾单张视频帧和连续的时间序列，在引入 CNN 模型的同时使用了循环神经网络（Recurrent Neural Network，RNN），如图 3.4。在每个时间步，RNN 循环层接收来自当前时间步的输入和前一个时间步的隐藏状态，然后更新隐藏状态，并输出给定时间步的结果。这种结构使得 RNN 可以捕获到序列数据中的时间依赖性。但是单纯的 RNN 无法处理随着递归、权重指数级爆炸或梯度消失问题，难以捕捉长期时间关联，所以本文结合了长短期记忆(LSTM)网络。LSTM 是 RNN 的一种变体，通过引入门结构（输入门、遗忘门、输出门）和细胞状态，它能够更好地处理长序列中的长期依赖问题。相比单纯的 CNN 模型，RNN 具有处理序列数据的能力，这使得 RNN 在诸如自然语言处理、时间序列分析以及这里的视频数据处理等任务中表现出色；RNN 可以处理动态长度的输入和输出，对于不同长度的视频采集数据，都可以通过 RNN 进行训练；虽然基本的 RNN 在实践中难以捕获长期依赖，但其变体（如

LSTM 和 GRU）通过特殊的结构成功地解决了这个问题，使得模型能够学习序列中的长期模式。

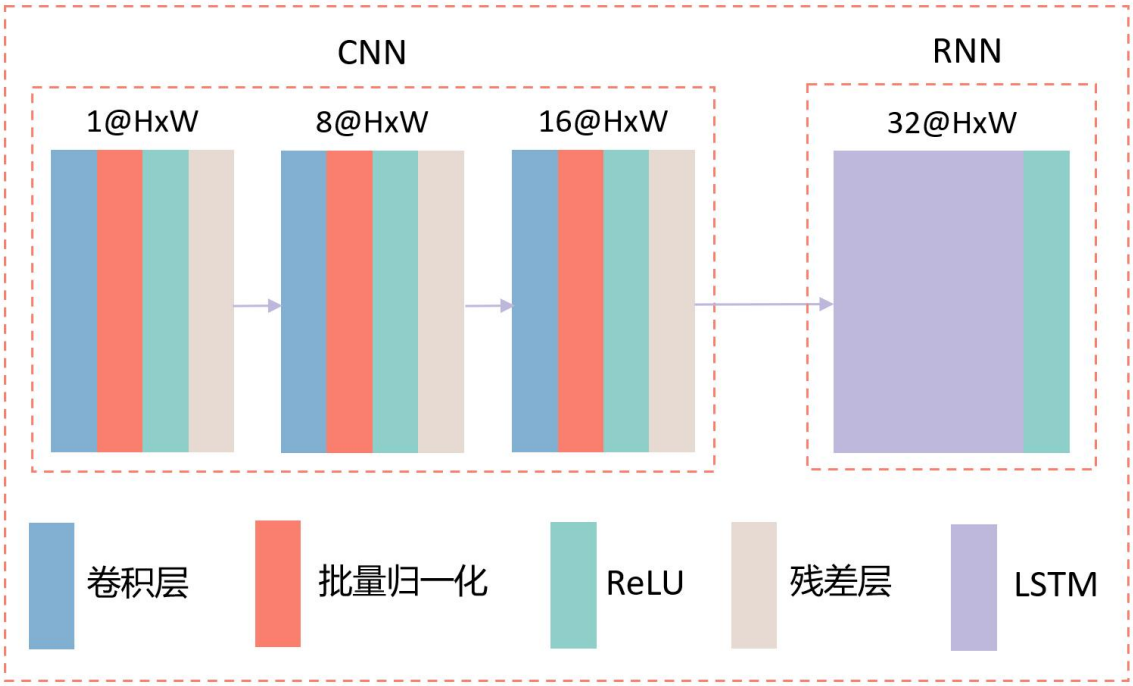


图 3.4　视觉感知模块模型框架图

考虑到视频采集样本的场景以及时间连续性，本文在对人体动作的视频数据处理时选择了 CNN 与 RNN 协同作用的方式，将压缩的视频帧样本先后经过多个 CNN 卷积层、标准化处理、激活层、RNN LSTM 结构，最终通过全连接层进行分类和输出。

3.4　基于多尺度的 Wi-Fi 无线感知模块

在对于 Wi-Fi 数据的训练模型时，起初选用的 CNN 模型并没有达到很好的效果。这里假设 Wi-Fi 的发射端与接收端之间有 γ 个通信信道，也即有同样数量的天线对，每条信道里有 ϕ 数量的子载波，Wi-Fi 的 CSI 数据经过 $\gamma \times \phi$ 个信道子载波（在本文实验采集时使用了 3 条信道，每条信道有 30 个子载波）收集后，在时间 τ 时的振幅强度矩阵可以表示为

$$\boldsymbol{A} = \begin{bmatrix} A_1^{1,1}, & \cdots, & A_1^{1,\phi} \\ \vdots, & \ddots, & \vdots \\ A_\tau^{\gamma,1}, & \cdots, & A_\tau^{\gamma,\phi} \end{bmatrix} \tag{3.1}$$

其中，A_t^{ij} 表示在 t 时刻，第 i 条信道的第 j 个子载波的 CSI 振幅数值。如果对于采集到的 $\gamma \times \phi \times \tau$ 的 CSI 数据（本文实验采集到的数据为 $3 \times 30 \times 1000$）直接采用 CNN 模型进行大核的卷积操作，就等于放弃了 CSI 数据时序性和空间连续性的特点，必须对两者进行取舍。时空域卷积网络[28]模型通过在多个尺度上分别进行不同的卷积操作，从多个维度进行特征提取，从而很好的解决了使用单一 CNN 无法同时兼顾时间和空间两个维度的连续性的缺点。

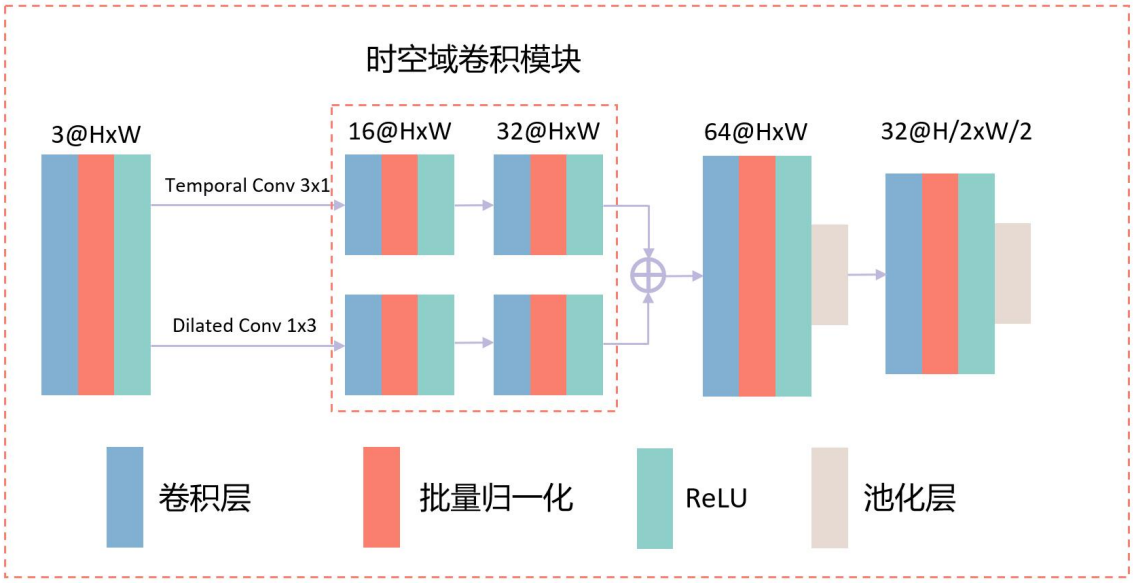


图 3.5　基于多尺度的 Wi-Fi 无线感知模块模型框架图

在基于多尺度的 Wi-Fi 无线感知模块中的时空域卷积网络模型（如图 3.5）中，对于两个分别在时间和空间维度上进行特征提取的小卷积核，使用了扩张卷积增加卷积核的感受野。扩张卷积（如图 3.6，Dilated Convolution），也称为空洞卷积（Atrous Convolution），是一种在深度学习中用于增加感受野（Receptive Field）的卷积方法，其最早在波形数据处理中被提出和应用，后来发现计算机视觉领域也有重要的作用。这种方法在不增加卷积核尺寸和参数数量的情况下，有效地扩大了网络的感受野，有助于捕捉更大范围的上下文信息。扩张卷积的基本思想是在卷积核的每个元素之间引入一个固定的间隔（称为空洞率，Dilation Rate）。空洞率为 1 时，扩张卷积等同于常规的卷积。当空洞率大于 1 时，扩张卷积的感受野变得更大，可以捕捉到更广泛的上下文信息。时空域卷积网络模型通过引入扩张卷积，可以在不增加卷积核尺寸和参数数量的情况下，扩大网络的感受野，有助于捕捉更大范围的上下文信息，同时扩张卷积的引入并不会提高计算复杂度。

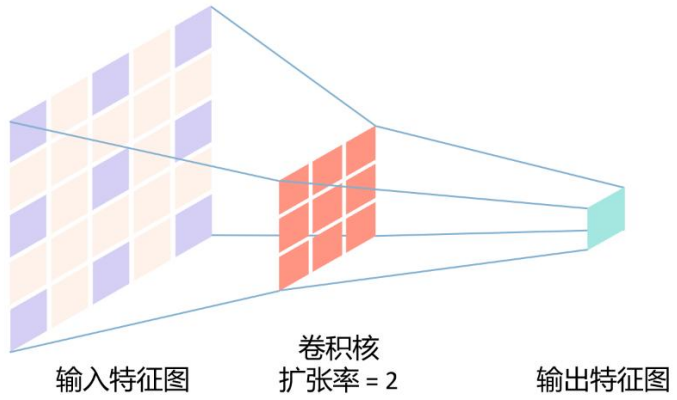


图 3.6　扩张卷积过程示例

根据 Wi-Fi CSI 数据的结构，时空域卷积网络模型提出了并行时空神经网络，在时间域和空间域上使用更小的卷积核并行卷积计算的方式，使得时空卷积模块能显著的提高对于 Wi-Fi CSI 数据在相同信道的不同子载波和连续时间段上的关联和变化的感知，引入两个 1×3 和 3×1 的小卷积核。通过这种方式，所有信道的所有子载波都能通过不同的方式被遍历，最终在连接层中合并这些数据流的计算特征，以获得对 Wi-Fi CSI 读数的隐藏层特征图。

3.5　多模态融合的 ViFi 模型

参考了同样由视觉和 Wi-Fi 两种模态进行深度学习训练的 GaitFi 网络，本文同样使用 TripletLoss 作为模型训练过程中的损失值作为反向传播的重要指标。三元组损失函数（TripletLoss）是一种用于学习深度神经网络嵌入的损失函数，尤其在图像识别和面部识别中广泛使用。它通过比较一个“锚点”样本与一个正样本和一个负样本的相对距离进行计算，目标是使得锚点样本与正样本之间的距离小于与负样本之间的距离。锚点样本与正样本来自同一类，而锚点样本与负样本来自不同的类。通常还会加入一个边距参数（margin）来确保锚点样本与正样本之间的距离相比比与负样本的距离小得多。三元组损失函数可以用欧几里得距离表示为公式 3.2。

$$\mathcal{L}(\boldsymbol{A}, \boldsymbol{P}, \boldsymbol{N}) = \boldsymbol{max}(\| \boldsymbol{f}(\boldsymbol{A}) - \boldsymbol{f}(\boldsymbol{P}) \|^2 - \| \boldsymbol{f}(\boldsymbol{A}) - \boldsymbol{f}(\boldsymbol{N}) \|^2 + \boldsymbol{\alpha}, \boldsymbol{0}) \qquad (3.2)$$

这里的 $\boldsymbol{A}$ 是一个锚点样本，$\boldsymbol{f}(\boldsymbol{x})$ 表示神经网络对样本 $\boldsymbol{x}$ 的嵌入，$\| \boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{y}) \|^2$ 表示样本 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 经过神经网络后的欧氏距离的平方，$\boldsymbol{\alpha}$ 是正负样本对之间的松弛边距。通过引入三元组损失函数 TripletLoss，可以学习到良好的样本嵌入，从而反映出样本之间的相似性。在使用边距参数后，它可以进一步提升模型的性能，使得同类样本之间的距离比异类样本之间的距离更小。

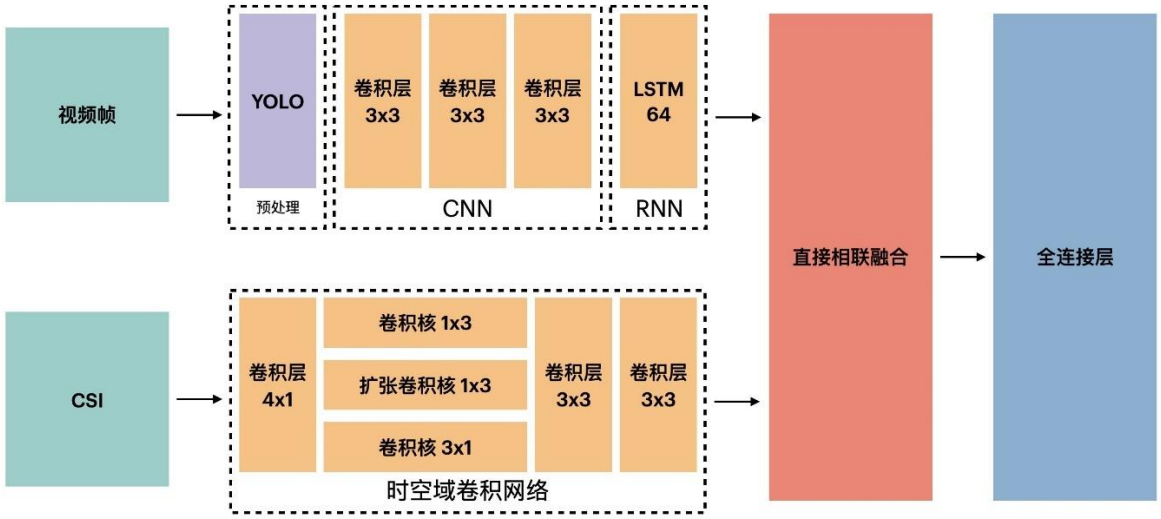


图 3.7　ViFi 模型细节设计图

在视频的帧数据经过基于 YOLO 的视觉目标检测模型进行数据增强后得到检测出人体的正方形聚焦裁剪图，先后经过视觉感知模块中的 CNN 和 RNN 模型，同预训练的权重进行多层的卷积计算得到视觉特征，与 Wi-Fi CSI 数据经过 Wi-Fi 感知模块中的基于多尺度的时空域卷积网络的 Wi-Fi 特征进行模态的特征融合操作，如图 3.7 中所示。两个模态经过多层卷积运算得到的矩阵均为大小 64×32 的张量，在进行合并时，ViFi 提供了三种融合方式，分别为直接相联、参数相加（即平均权重分配）和依据权重分配进行加权平均的方式。直接相联保留了两种模态最多的信息，同时也意味着最后的全连接层的输入扩大一倍，经过合并后的张量大小变为 128×32。张量直接相加也是一种平等的加权运算，使视觉和 Wi-Fi 两种模态在最终的融合结果中占据相同的分量，输出张量大小为 64×32，但也可能因为量纲不一致导致权重出现较大的偏差。通过参考视觉和 Wi-Fi 两种单模态在本文实验的各种场景（不同光照度、不同遮挡情况）下的测试结果，可以对两者在融合过程中的权重占比进行调整，从而使模型相比参数相加平均权重分配效果更好，但是这也通常意味着需要先分别得到两种单模态的预测准确率，并根据两个模态的测试结果进行权重更细微的调整。相比之下，直接相联的方式虽然得到的模型效果并不一定是最佳，但是比较稳定，与最佳权重的效果相差不大。经过特征融合后的多模态数据最终通过最后一层全连接层进行标签分类，将其对应到在每个场景下的不同动作。

3.6 基于 LRM 的跨域感知模型预训练

智力是一个多方面的、难以捉摸的概念，目前还没有一个普遍认同的定义，但有一个方面是被广泛接受的是，智力并不局限于某个特定的领域或任务，而是包含了一系列广泛的认知技能和能力。建立一个能表现出这种广泛行为的人工系统是人工智能研究的一个长期的人工智能研究的一个长期和雄心勃勃的目标。今年以来，大模型成为了人工智能相关领域的热门话题。大模型通常具有更多的参数和更强大的学习能力，可以生成更准确、更连贯的文本、图像或音频等内容；可以更好地理解上下文、语义和语法规则，从而生成更自然、更富有创造力的结果；可以处理更复杂、更具挑战性的任务，为各种应用场景提供更好的解决方案。

本文试图训练一个大型跨域感知模型（Large Recognition Model，LRM），让它可以不局限在某一个特定场景的预训练权重，而是预载一个基于所有场景的混杂数据集进行预训练得到的权重并将其运用在不同的情况，使其能应对更为复杂、多变的场景。在本文的实验中只使用了（100, 10, dark）三种光照度降低的场景和白板、桌子遮挡的场景，通过典型的环境训练模型并让它能在更多不同的光照强度和不同的遮挡环境下对人体行为进行感知和识别，使模型有更好的鲁棒性和泛用性。

LRM 预训练权重的跨域指模型能使用一个预训练结果的最优权重使用不同场景的测试集进行预测，并达到较为优秀的识别准确率。而当面临一个新的场景时，可以在没有进行数据标记和训练的情况下直接对其进行人体动作识别。为了达到这个目标，就需要对数据集的种类进行筛选、合并和混杂，让预训练模型能学习到在不同场景下的视觉和 Wi-Fi 特征。

4 ViFi 感知模型的实现

4.1 基于 YOLO 的视觉目标检测预处理流程

在 ViFi 模型对视觉部分数据的预处理中，使用了 YOLO 官方提供的 yolov5m 预训练权重对图片帧中的人体进行检测、裁剪和处理。

表 4.1 基于 YOLO 进行视觉数据检测和裁剪预处理

算法 1
Input: $\mathcal{V}$ 是视觉帧数据样例; $\mathcal{F}$ 是 YOLO 目标检测模型; acc 物体的识别可信度; $\mathcal{C}$ 图片裁剪工具
Output: 被检测者 $\mathcal{P}$
1: **for** each $\tilde{v} \in \mathcal{V}$ **do**
2: $Z_i \leftarrow \mathcal{F}(\tilde{v})$
3: **if** Z_i is not person and $Z_i^{acc} < acc$ **then**
4: discard
5: **end if**
6: expand the identified target Z_i into a square Z_i^{square}
7: $\mathcal{P} \leftarrow \mathcal{C}\left(Z_i^{\text{square}}\right)$
8: **end for**

在预处理的过程中（如算法 1 所示），在从视频数据集中按照时间流顺序抽取视频帧后，将其输入到基于 YOLO 的视觉目标检测模型进行识别，并检测识别结果是否为人体且是否达到可信阈值，将无效数据丢弃，并将人体区域裁剪出来，抛弃图像中其他的物体和背景中的环境信息，对每一种实验场景生成新的 crop 数据集。为了保持模型输入的一致性和目标检测的唯一性，将超过可信阈值的被检测物体图像扩张为正方形。而对于极端条件下无法检测出物体的场景，则直接对图像数据进行重构处理。

4.2 三元组损失的计算

在 ViFi 模型训练时的，使用三元组损失（Triplet Loss）对当前迭代的模型训练权重的预测结果与真实标签的偏差值进行表示。

算法 2 描述了 TripletLoss 类在每个数据批次中如何计算三元组损失。第 2-5 行将从数据批次中提取核解析到的全局特征和标签在需要时进行归一化处理，然后通过函数 $\mathcal{E}$ 计算所有样本之间成对的欧氏距离，并挖掘出与每个样本最接近的正样本和最远的负样本，计算与这两个样本的距离。最后根据是否指定了边界值，使用 MarginRankingLoss 或 SoftMarginLoss 计算损失并返回，用来在后续的训练迭代中进行反向传播和优化。

表 4.2 ViFi 模型训练过程中的三元组损失计算函数

算法 2

Input: M 是三元组损失值计算方式选择; $\mathcal{E}$是欧式距离计算函数; $\mathfrak{F}$ 是全局特征; $\mathfrak{L}$ 是全局标签

Output: 三元组损失值 $\mathcal{L}$

for each batch of data **do**
  $\mathfrak{F}, \mathfrak{L} \leftarrow$ extract global features and labels
  if required normalize the global features **then**
    $\mathfrak{F} \leftarrow \text{normalize}(\mathfrak{F})$
  end if
  dist_mat $\leftarrow \mathcal{E}(\mathfrak{F})$
  dist_ap, dist_an $\leftarrow$ hard_example_mining(dist_mat, , $\mathfrak{L}$, mask)
  if M is not None **then**
    $\mathcal{L} \leftarrow$ MarginRankingLoss(dist_an,_dist_ap)
  else
    $\mathcal{L} \leftarrow$ SoftMarginLoss(dist_an − dist_ap)
  end if
end for

4.3 ViFi 模型数据训练和预测标签整体流程

表 4.3 ViFi 感知模型预测人体动作流程

算法 3

Input: $\mathcal{D}$ 是视觉和 Wi-Fi CSI 数据和标签; $\mathcal{F}$ 是 ViFi 感知模型; R 是动作类别; y_r 是第r类的 one-hot 真实标签

Output: 被检测者 $\mathcal{P}$

$(\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{test}}) \leftarrow \text{split}(\mathcal{D})$
for each $\boldsymbol{D} \in (\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{test}})$ **do**
  crop video frames using Algorithm 1
end for
for each $\widetilde{\boldsymbol{D}} \in \mathcal{D}_{\text{train}}$ **do**
  $Z \leftarrow \mathcal{F}(\widetilde{\boldsymbol{D}})$
  for each $Z_i \in Z$ **do**
    predicted probability $p(r|\widetilde{\boldsymbol{D}};\theta) \leftarrow \dfrac{\exp(Z_i)}{\sum_j \exp(Z_j)}$
  end for
  minimize $\mathcal{L}(\theta) \leftarrow$ Triplet Loss using Algorithm 2 by Adam optimizer
end for
$\mathcal{F}_{\text{best}} \leftarrow$ bestModel (θ)
$\nabla \leftarrow \mathcal{F}_{\text{best}}(\mathcal{D}_{\text{test}})$

算法 3 详细的阐述了 ViFi 感知模型的数据预处理、数据训练、模型筛选和结果标签预测的整体流程。

ViFi 感知模型首先将分割后的数据集中的视觉数据使用基于 YOLOv5 的视觉目标检测模型（如算法 1 中流程）进行检测、裁剪和处理，并生成 crop 数据集。将训练集经过预处理后的视频和 Wi-Fi 数据按照批依次输入到 ViFi 感知模型进行训练并通过计算三元组损失进行反向传播，从而使模型预测的结果更加准确，最终得到在实验过程中的所有模型预训练权重的结果。从中选取识别效果最好的结果对应的预训练权重重新载入，使其能够对训练数据集进行分类和动作预测。整个感知模型的时间开销主要集中在 2-4 行的视频数据帧预处理和 5-11 行的模型训练部分。

4.4 ViFi 模型测试结果准确率计算

ViFi 感知模型准确率计算算法 4 先得到每个测试集特征和训练集特征之间的平方欧氏距离，并存储在距离向量 dist 中，在其中找到每个测试集特征的最小距离的索引，并存储在索引向量 index 中。使用索引向量 index 将训练集标签进行关联，以获取预测标签 pred，并确保 pred 的种类与测试集标签相同。最后统计出所有预测标签与真实的测试集合标签相同的实验结果个数，与总测试用例个数相除就能得到测试结果的准确率 $\boldsymbol{\alpha}$。

表 4.4 ViFi 感知模型准确率计算

算法 4
Input: $\mathfrak{F}_G$ 是训练集特征; $\mathfrak{L}_G$ 是训练集标签; $\mathfrak{F}_P$ 是测试集特征; $\mathfrak{L}_P$ 是测试集标签; $\mathcal{E}$ 是欧式距离计算函数
Output: 准确率 α
1: dist ← $\mathcal{E}(\mathfrak{F}_P, \mathfrak{F}_G)$
2: index ← minimum value of every line in dist
3: pred ← $\mathfrak{L}_G$[index]
4: **assert** same shape (pred, $\mathfrak{L}_P$)
5: α ← summation number of (pred == $\mathfrak{L}_P$) ÷ total number

5 实验设计与结果分析

本文通过大量的对比实验和消融实验介绍 ViFi 多模态人体行为感知模型的详细实现和识别性能。本章将介绍 ViFi 感知模型的承载实验的软件和硬件环境、模型训练和测试使用的数据集、模型对比实验和消融实验的结果、模型参数量和计算复杂度、对模型的实时性和鲁棒性进行分析，以及在最后一部分介绍基于 LRM 的预训练权重的评估结果。

5.1 实验设置

5.1.1 实现环境

为了能采集人体动作感知数据，本研究使用了一台商用的工作在 5GHz 频段（关闭 2.4GHz 频段的信号发射）的 TP-LINK AC1750 无线路由器作为 Wi-Fi 信号的发射端，通过简易的接收天线按照时间序列每秒接收和存储 500 个信号包，并传输到一台搭载了商用的 Intel 5300 NIC 网卡和装载 CSI Tools [38]驱动来对抓取到的 CSI 信号数据包进行解析和处理，提取出其中的振幅数据，并在旁边使用 Logitech Webcam C930 摄像头同步对志愿者进行视频的采集，视频采集的分辨率为 640 × 480 像素，实验数据集采集环境如图 5.1 所示。发送端和接收端天线相距 4.5 米，离地高度 0.9 米。最终的 ViFi 模型在 3 台搭载了 Nvidia 3070Ti（8GB RAM）、1 台 Nvidia 3070（8GB RAM）的电脑主机和 1 台双路 Nvidia Titan Xp（12GB RAM）的服务器同步训练和测试，使用 2.0.1 稳定版本 PyTorch 搭配 11.8 版本 CUDA 作为实验训练环境的标准。

5.1.2 数据集

为了评估 ViFi 模型在两种不同程度遮挡场景和三种不同光照亮度情况的效果，研究采集了 8 名不同年龄（年龄范围在 20 岁至 29 岁）、不同身高（身高范围在 158 厘米到 189 厘米）、不同体型志愿者（其中 5 名男性，3 名女性）在正常光照无遮挡条件、光照度分别为 3.7Lux, 1.5Lux, 0Lux（使用德力西 DLY-1802 照度计在实验感知区域的中轴线不同位置测量取平均值，分别由 100, 10, dark 指代）、使用桌子、白板或书架进行遮挡共 7 种场景下的视频和 Wi-Fi CSI 数据，志愿者的信息去除个人隐私后如图 5.2 所示。

实验对每个人的每个场景都采集了 8 种动作，其中 4 种粗粒度活动（步行、跑步、捡东西、踢腿）和 4 种细粒度活动（坐下、起立、拍手、挥手），动作如图 5.3 中所示，每位志愿者每项动作需要重复 40 次，每次持续时间约 2 秒。对于采集结果不理想的数

据（如采集时间过短、Wi-Fi 天线无法接收数据、环境干扰较多等），在采集的过程中直接抛弃并重新采集该条数据。

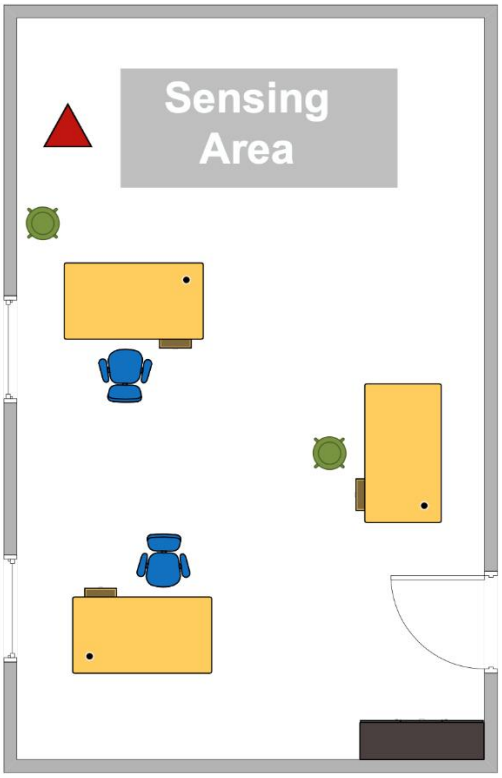


图 5.1　实验数据采集环境示意图

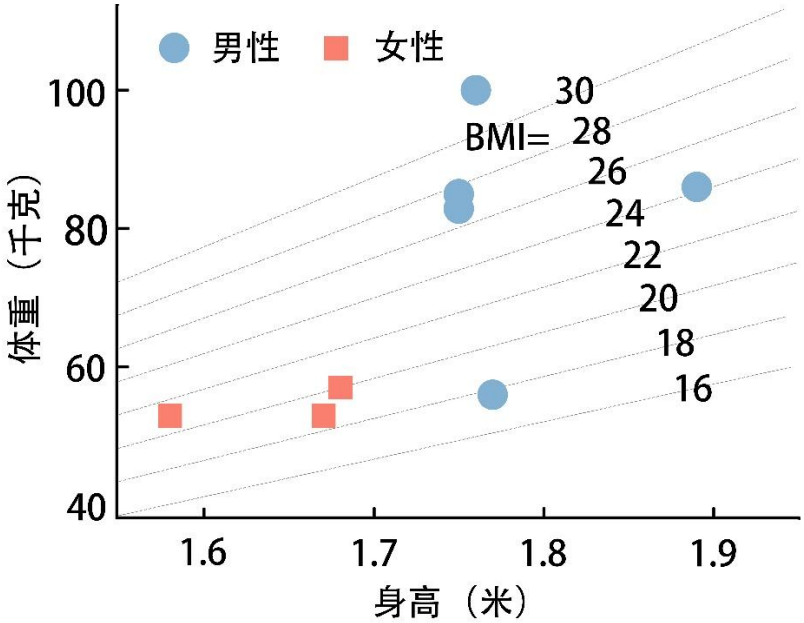


图 5.2　实验数据采集志愿者信息

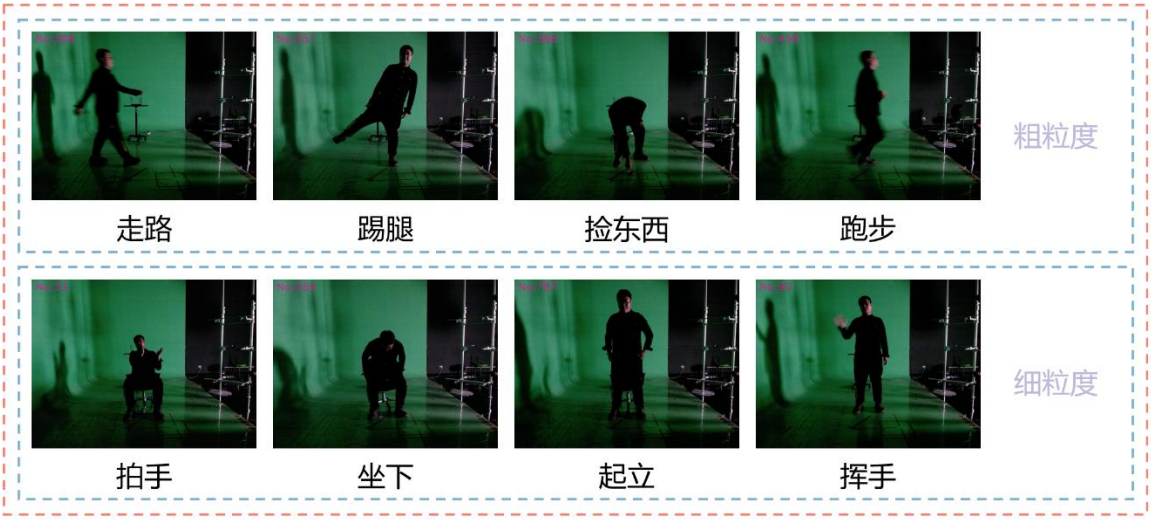


图 5.3　实验采集粗细粒度动作示意图

最终采集到的数据集总共包括 35840 个动作样本（7 种场景 × 40 次实例 × 8 种动作 × 8 名志愿者 × 2 种模态），总计大小为 49.5GB（包括经过视频预处理的 crop 数据集和未经解码处理的原始 CSI 数据集）。在采集的过程中没有环境因素变化的影响，无动态干扰因素。最终原始 Wi-Fi CSI 的 CSV 数据集和视觉的 Video 数据集被分别预处理成 Mat 和 crop 数据集并按照 8:2 分割成训练集和测试集，如算法 5 所示。

表 5.1　数据预处理算法

算法 5
Input: $\mathcal{D}$ 是视觉和 Wi-Fi CSI 数据和标签; $\mathcal{S}_F$ 是随机函数; $\mathcal{S}_P$ 是数据流分割函数; $\mathcal{M}$ 将文件从原位置移动到分割路径
1: **for** each data in $\mathcal{D}$ **do**
2:　shuffled_data $\leftarrow \mathcal{S}_F$(data)
3:　$\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{test}} \leftarrow \mathcal{S}_P$(shuffled_data)
4:　preprocess $\mathcal{D}_{\text{Video}}$ into $\mathcal{D}_{\text{crop}}$ by Algorithm 1
5:　resize $\mathcal{D}_{\text{crop}}$
6:　$\mathcal{M}(\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}})$
7: **end for**

在进行数据集分割的过程中，先将所有的数据打乱顺序，并按照比例划分和截取作为训练集和测试集，并将其中的视频数据通过算法 1 裁剪进行数据增强操作，最终经过数据帧压缩存储到合适的位置。

5.2 模型评估

5.2.1 对比实验和消融实验结果

消融实验（Ablation Experiment）是一种在科学研究和医学领域中常用的实验方法，旨在通过破坏或去除特定组织或结构来研究其功能、相互关系和生理过程。在计算机领域，消融实验通常指的是一种评估模型的特征重要性的方法。它用于确定模型中哪些特征对模型的性能起到重要作用，以及在去除这些特征后模型的性能如何受到影响。

在人工智能和数据挖掘任务中，特征选择是一个重要的问题，因为选择合适的特征可以提高模型的效果和解释性。消融实验通过反复训练模型并去除一个或多个特征，然后比较去除前后的性能差异，从而确定特征的重要性。它可以使用不同的方法来进行。其中一种常见的方法是逐个消除特征。首先，使用所有特征进行模型训练和评估，得到基准性能。然后，逐个去除每个特征，重新训练和评估模型，并记录每个特征的去除对性能的影响。通过比较不同特征去除后的性能变化，可以确定它们对模型性能的贡献。另一种方法是随机消融实验，其中随机选择一组特征并将其从模型中去除，然后评估模型的性能。重复此过程多次，并记录每次消融的特征组合和性能结果。通过统计分析这些结果，可以得出特征的重要性排名。消融实验在模型解释、特征工程和模型优化等方面具有重要的应用。通过了解特征的重要性，可以简化模型、提高模型的解释性、减少计算成本，并发现不必要的特征或噪声对模型性能的影响。

在本文的研究中，对 ViFi 模型与原 GaitFi 的 CRNN 模型之间、添加数据预处理与否、单个模态与多模态融合对比、不同的模态融合方式等方面进行了对比实验和消融实验，共对模型在不同的实验环境场景下测试 200 余次，总训练和测试用时超过 100 小时，最终生成的单模态与多模态对比数据如图 5.4 所示。

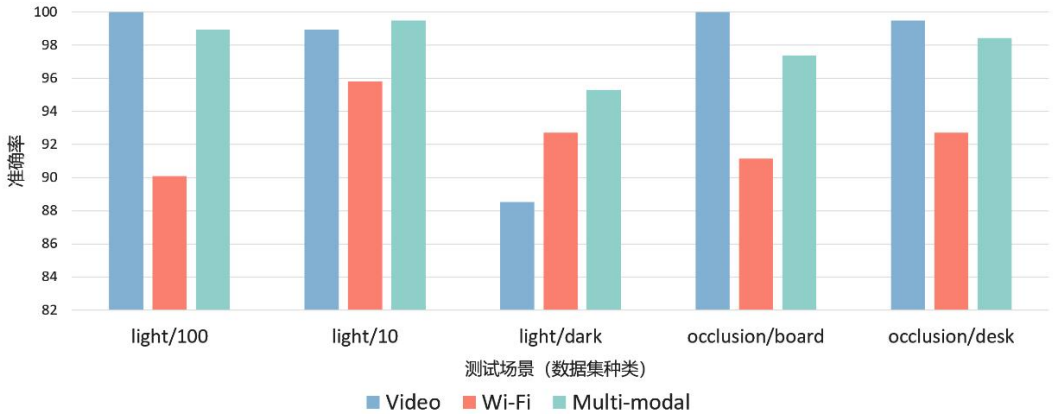


图 5.4 对 ViFi 多模态融合模型的单模态与多模态对比实验结果

对于采取基于 YOLO 的视觉目标检测作为视觉数据的预处理部分前后，从对 ViFi 多模态融合模型的消融实验结果对比图 5.5 中可以看到，在光照度分别为 100, 10, dark 的三种条件下，增加基于 YOLO 的视觉目标检测作为视频的数据预处理的一部分都对视频单模态的效果有一定的提升作用，其中在光照度最高的 100 标值场景下提升了 4.18%，在对视频采集数据影响最大的接近全黑环境下测试也有 2.64%的准确率提高，达到 88.542%。在两种遮挡场景下，通过添加基于 YOLO 的视觉目标检测预处理将模型的注意力聚焦在人体上，抛弃周围环境的影响也同时降低了遮挡物体在视觉中的占比和对模型训练的影响，最终白板遮挡场景提高了约 4.69%，而书桌遮挡场景也从没有使用基于 YOLO 的视觉目标检测预处理时的 92.188%提升到在常规数据集下都能够实现精准的预测。

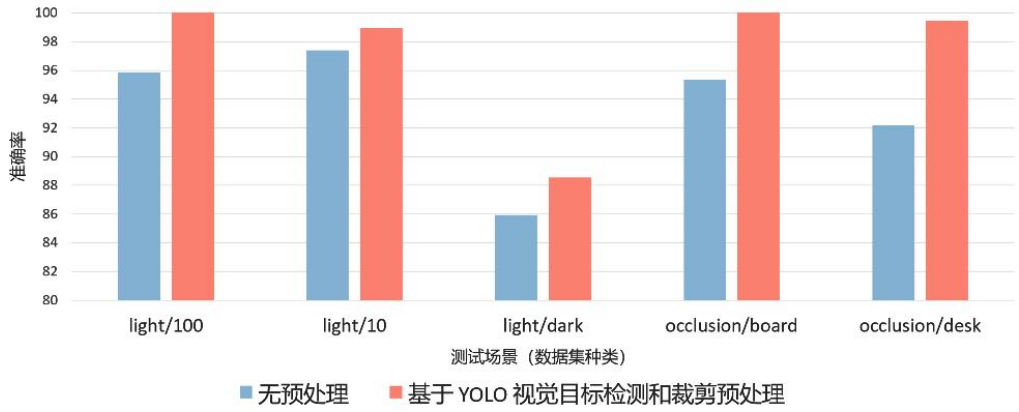


图 5.5　添加基于 YOLO 的视觉目标检测作为预处理对模型实验结果的影响

在使用了时空域卷积网络替换了原多模态模型 Wi-Fi 部分的 CNN 之后，对于 Wi-Fi CSI 数据在同一条信道相同子载波的时间序列上的特征和相同信道内不同子载波之间的变化被多个并行小卷积核所捕获，并通过扩张卷积增大了对原始数据的感受野，使卷积核可以以较小的模型参数和运算时间为代价从更长的时间段和更大子载波信号振幅中抽象出特征图。

从图 5.6 中可以看出时空域卷积网络模型达到了更好的检测和识别效果。在光照度最大的 100 标值实验中，替换的时空域卷积网络模型就对 Wi-Fi 有 3.65%的提升；而在更暗的 10 光照度和接近全黑的 dark 数据集上，提升幅度分别能达到 7.29%和 6.25%。由于白板遮挡的面积更大，对于信道子载波的影响也更为明显，即便在通过空间上增加卷积核进行特征提取也仅有 4.17%的提升；而书桌遮挡面积较小，且无线信号可以更为轻易的通过反射等方式“穿过”遮挡物，提升幅度也相对大得多，达到 11.46%。

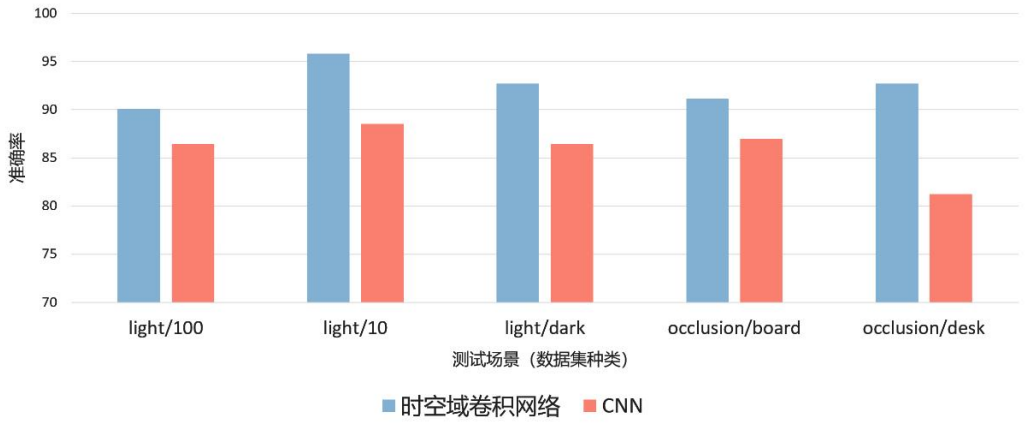


图 5.6　时空域卷积网络模型与原 CNN 模型在 Wi-Fi 单模态上的效果对比

在视觉和 Wi-Fi 的实验数据分别经过各自的单模态模型后，得到的结果需要进行合并之后再输入全连接层。在本文的实验中，着重以桌子遮挡的场景下，模型在使用不同的融合方式以及不同的权重值时进行对比实验，结果如图 5.7 所示，图中的权重表示为（视觉权重，Wi-Fi 权重）。

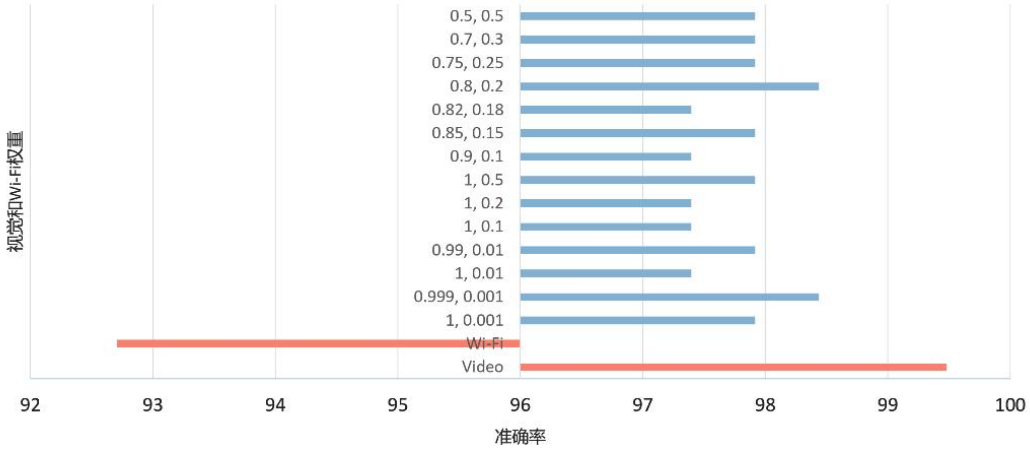


图 5.7　桌子遮挡场景下不同权重设置对实验结果的影响

可以看出，在对应的视频和 Wi-Fi 单模态的测试结果分别为 99.479%和 92.708%，视觉识别的效果比单纯使用 Wi-Fi 更准确时，多模态融合的效果并非与测试结果更好的单模态在融合时的权重占比呈线性相关，但总体上仍然出现当视觉权重和 Wi-Fi 权重分别为 0.999、0.001 和 0.8、0.2 时多模态模型效果达到最优。在多个模态的融合过程中，不同模态之间的特征相关性可能不同，在这里即虽然视觉的准确率较高，但如果其特征与任务目标的相关性较低，将其权重设置得过高可能并不会带来最好的效果；相反，从

Wi-Fi 的 CSI 数据中提取到的一些特征可能与任务目标更相关。除此之外，也需要考虑模型的鲁棒性和泛化能力，即便视觉在训练集上得到的损失函数较小，表现出更好的准确率，但如果其在未见过的数据上的泛化能力较差，那么将其权重设置得过高可能导致过拟合和性能下降。同时，在数据的采集等过程中也会产生一定量的噪声，它们或许是因为采集的方式不能够达到预期，或是由于设备的不一致等。在权重变化的过程中，不仅是提取到的特征值被放大，噪声也随着倍率变得更为明显。如果此时恰好在权重的作用下，其中一个模态的噪声掩盖掉了另一个模态原本可以利用的特征，就会对模型的学习和预测产生一些影响。这也是为什么在 ViFi 模型中，单纯的提高单模态时效果更好的一方并不一定得到更好的多模态测试结果。

5.2.2 参数量和模型计算复杂度

参数量和 GFlops 是对模型性能和复杂程度的重要评估方式。模型的参数量指的是模型中需要学习的可调整参数的数量。这些参数是通过训练模型时自动调整的，以使模型能够适应给定的数据集。参数量通常是衡量模型大小的指标之一。较大的参数量意味着模型具有更多的自由度来学习复杂的模式和关系。参数量越多，模型通常越能够拟合训练数据，但也容易过拟合。GFlops（GigaFlops）是衡量计算机或计算设备每秒能够执行的浮点运算次数的度量单位。Flops 是每秒能进行的浮点运算次数的缩写。在人工智能领域，GFlops 常常被用来衡量模型的计算复杂度和运算效率,它可以用于评估模型在训练和推理过程中所需的计算资源。

表 5.2 模型的参数量和复杂度对比

模型	模态	参数量	GFlops
CRNN	视觉	37392876	0.032438272
	Wi-Fi	73660	0.004306224
	多模态	37466524	0.76392424
ViFi	视觉	37392876	0.032438272
	Wi-Fi	456810	0.731486224
	多模态	37849418	0.76392424

从表 5.2 中可以看到,原模型 CRNN 的视觉部分模型的参数量是 Wi-Fi 部分的约 500 倍，相应的复杂程度（由 GFlops 反映）也达到约 7.5 倍，最终完整的多模态模型参数量超过 3746 万，而其 GFlops 约为 0.76。在将 Wi-Fi 模型替换为时空域卷积网络的 ViFi 模型中，单 Wi-Fi 部分模型参数量提高到原模型的 6 倍，多模态模型也要比 CRNN 参数

量多 38 万，但是由于添加的均为较小的用于在时间和空间上分别提取特征的卷积核，并行的通过扩张卷积计算，在参数量可控的变化下扩张 Wi-Fi 模型的感受野，并融合多维度特征的结果综合学习，使得 ViFi（直接相联融合模态）在 GFlops 也即模型计算复杂度上与原模型持平，达到了惊人的效果。

除了模型的参数量和 GFlops 之外，时间也是衡量模型计算复杂度的一个重要性能指标。在深度学习的相关任务中，模型的计算复杂度直接影响了模型训练和预测得出结果的时间开销。对于大规模的数据集或需要满足实时性能的应用，计算复杂度就成为了一个至关重要的考虑因素。为了评估模型的计算复杂度，本文考虑了模型在训练迭代过程中和结果预测过程中所需要消耗的时间。训练时间是指完成一次训练迭代所需要的时间，而推理时间是指对一个输入样本进行预测得出结果所需要的时间。这些时间开销可以由模型的计算复杂度决定，因为更复杂的模型通常需要更多更复杂的计算操作，例如浮点计算，这必然将导致需要更长的计算时间。对于计算资源受限的情况，如嵌入式设备或移动设备，时间成为了一个重要的约束条件。在这种情况下，就需要选择计算复杂度较低的模型，以保证模型能够在有限的时间内完成训练和推理任务。一些模型压缩和加速技术，如剪枝、量化和模型优化算法，可以帮助降低模型的计算复杂度，从而提高计算效率。总之，时间作为衡量模型计算复杂度的直接指标之一，对于选择合适的模型、优化计算性能以及满足特定应用需求都具有重要意义。综合考虑参数量、架构和时间等因素，可以更好地评估和比较不同模型的计算复杂度，并选择最适合特定应用场景的模型。

表 5.3　模型的参数量和复杂度对比

模型	模态	任务		
		epoch 平均时长	输入输出处理用时	模型训练时长
CRNN	视觉	37.3	33.0	4.3
	Wi-Fi	31.7	30.4	1.3
	多模态	37.9	33.1	4.8
ViFi	视觉	37.3	33.0	4.3
	Wi-Fi	32.5	30.5	2.0
	多模态	38.8	33.5	5.3

在本文的实验中统计了 ViFi 模型和 CRNN 模型两者在每一次训练迭代中从数据集中读取数据到内存并转存到显存的过程所花费的时间，实际每一个批处理（batch）所耗费的时间和迭代的总时间，并由每次迭代的时间开销减去输入输出处理所用时间得出模

型在每次迭代时训练前向传播（Forward Propagation）或者正向计算（Forward Computation）以及反向传播（Backpropagation）所需要花费的时间，最后整理并汇总的结果如表 5.3 所示。

通过对比图 5.8 可以更直观的看到，在每一次循环迭代的过程中，输入输出处理所占用的时间最长，而不同的数据集所导致的每一次的输入时间也并不相同。为了能更好的对比模型训练所花费的时间，这里在对比时忽略了模型前的数据流读入和处理的时间，但是在实际应用时，运行在 CPU 和内存上的输入输出可能需要花费大量的时间，对使用模型进行预测造成很大的困扰。从模型的训练时长可以看到，时空域卷积网络相比 CNN-WiFi 需要多花费 0.7 秒，映射到多模态模型时每次需要多 1 秒，考虑到视频占比仍然是多模态模型中时间开销中较高的一个，ViFi 在训练时间，也即模型计算复杂度上与 CRNN 比较接近，处在可以接受的范围内。

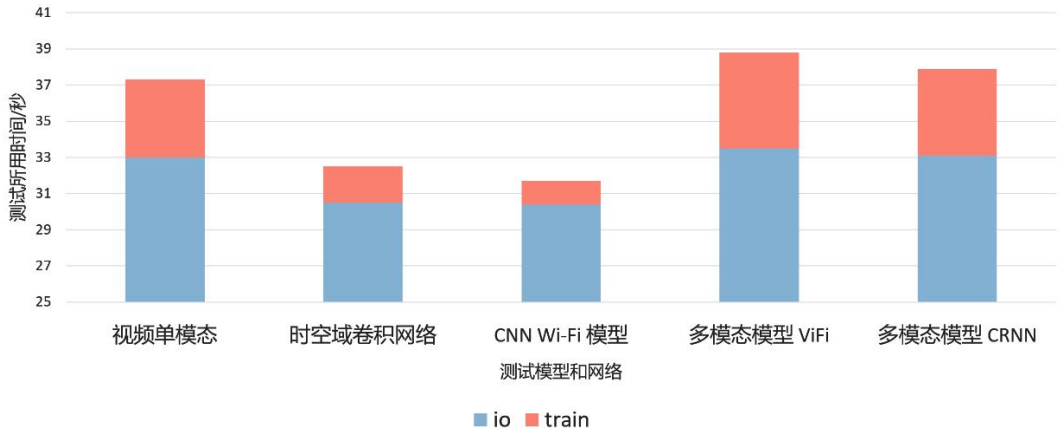
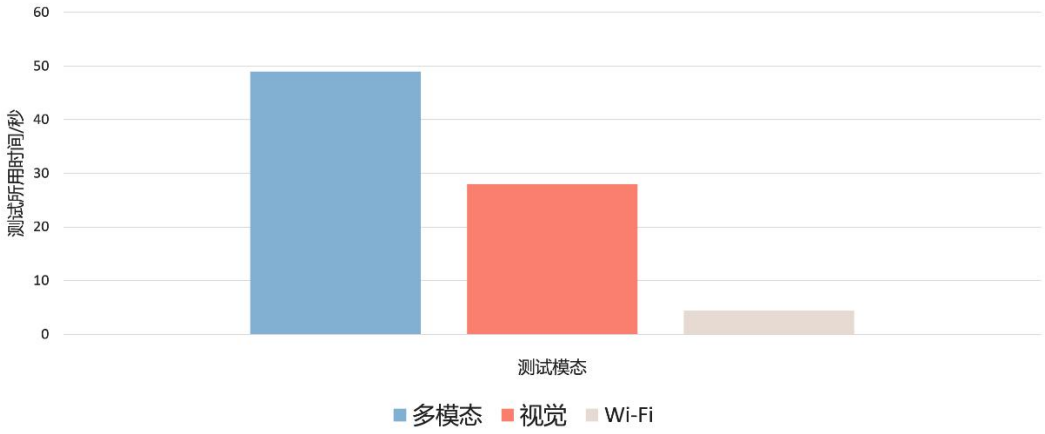


图 5.8　ViFi 和 CRNN 在单模态和多模态训练时的 IO 和训练时间

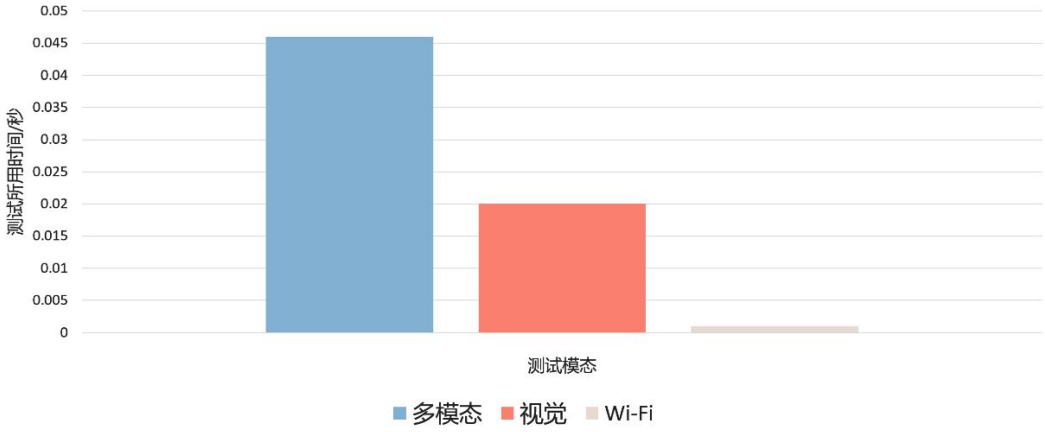
5.2.3　实时性分析

通过图 5.9 可以看到，ViFi 模型在实时性方面仍然有较大的潜力。人体动作检测深度学习模型的实时性意味着该模型能够在实时应用场景中快速、准确地检测和识别人体动作。它可以用于实时反馈和控制系统、运动分析和辅助系统、健康监测和照护系统、安全和监控系统等，在日常生活中有非常重要的意义。

在图 5.9(b)中可以看到，ViFi 不管是视觉和 Wi-Fi 的单模态还是融合后的多模态模型，对每一个批（batch，在这里为 32 次帧采样）的预测所耗费的时间都非常短，使用时空域卷积网络模型对 Wi-Fi 数据可以实现每 0.001 秒进行一次批处理，而多模态模型需要消耗 0.046 秒。仅参考模型的预测速度时，ViFi 可以满足实时性检测的要求。但是



(a)　ViFi 模型测试总用时



(b)　ViFi 模型单个批测试用时

图 5.9　ViFi 模型时间开销测试

在图 5.9(a)中考虑了对数据采集后的读取和内存显存之间的传输时间后，模型处理 6 批次的数据共用时 49 秒，且没有考虑对视频帧的 YOLO 预处理和对 Wi-Fi CSI 数据的解码和解析，所以在实时性的道路上仍然任重道远。

5.2.5　鲁棒性分析

鲁棒性是指系统或算法对于输入数据的变化、异常情况或不完美条件的适应能力。在实际应用中，数据往往会包含噪声、异常值或不完整信息，因此鲁棒性成为评估一个系统或算法优劣的重要指标。在进行实验数据采集的过程中，由于实际环境的复杂性和不确定性，有许多相互干扰的无线信号在空气中存在和传播，这与 ViFi 感知模型在实际应用中的场景类似，但是会对模型的训练有一些干扰，使得模型的输入中会掺杂一定量的噪声信息。在视频采集部分由于为了能更好的模拟实际应用的场景，使用了商用廉价的摄像头，在无法提高感光度的情况下使得视觉捕捉的效果并非最佳状态，同样会对

模型的感知造成影响。除此之外，不确定性和扰动的来源多样，可能会有仪器的测量误差、模型参数初始化状态不同等。ViFi 感知模型由两种模态相互协作，互为补充。实验最终采用蒙特卡洛方法，将模型的初始权重设置了不同的随机值，并使模型在不同的场景下进行测试，最终得到图 5.10。可以看到，原 CRNN 模型在设置不同的随机值并生成随机的初始权重的情况下，表现不稳定，在 Wi-Fi 单模态的测试结果中准确率上下相差 11.46%，在多模态模型的测试表现也有 2.65%的差距；而在 ViFi 感知模型在使用小批量进行测试后表现出较好的稳定性。最终实验的所有测试都运行在随机种子为 3407 的初始权重下，这也是经过测试表现最为稳定的随机种子。这些结果表明，ViFi 感知模型对于不确定性和扰动具有强大的抵抗能力，表明其在实际应用中的潜力，有较高的鲁棒性。

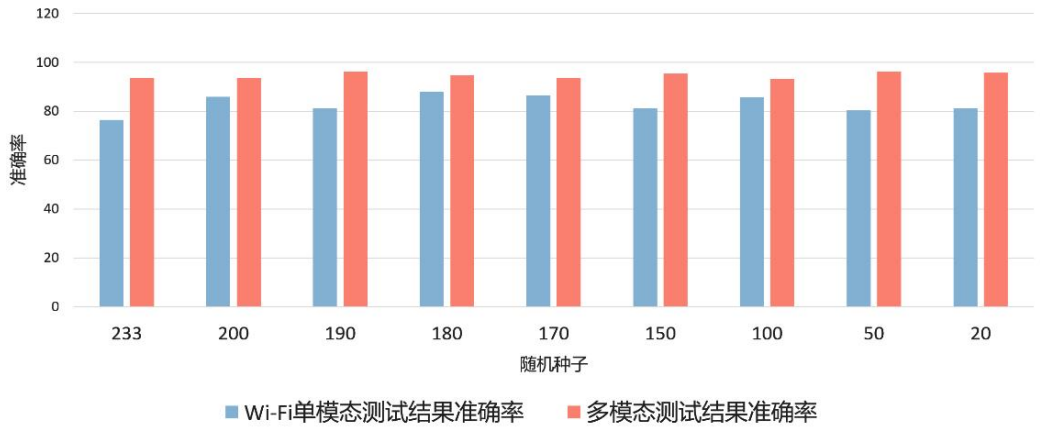


图 5.10　模型在 Wi-Fi 单模态和多模态模型不同随机种子时的测试结果

5.3　基于 LRM 的预训练评估

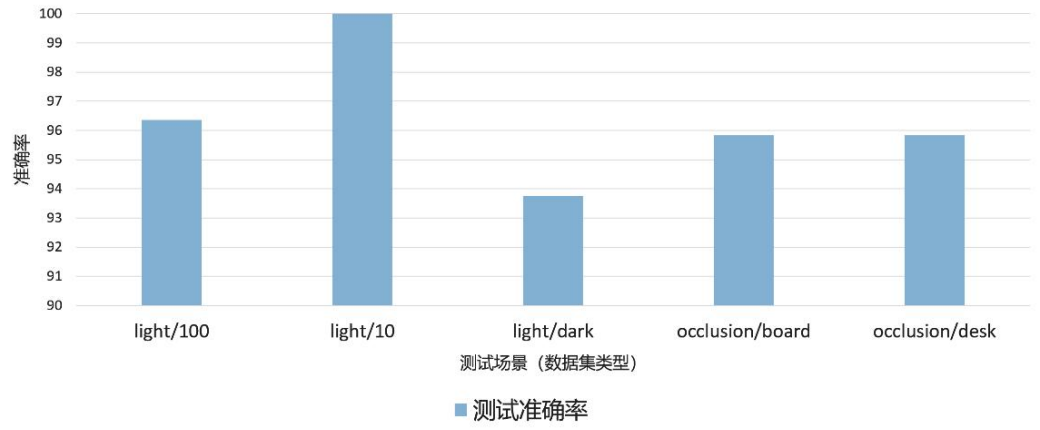


图 5.11　ViFi LRM Stream 训练在不同场景数据集上测试结果

在实现 ViFi 的 LRM 模型预训练的过程中，本文对比了两种训练方式（Stream 和 Collect）。如图 5.11 是 Stream 方式将场景数据集依次输入模型进行训练得到的结果，由于不同场景下数据的差异性较大，可以看到模型在接收新的数据集反向传播的过程中，会削弱原本对于上一个数据集适用的模型参数的作用，使最终的 LRM 模型对先训练的数据集没有很好的识别效果，仅能与 CRNN 在 Wi-Fi 环境下的识别率相当（63.021%）。所以本文调整了 LRM 的训练方式，Collect 将所有场景的数据集融合并重新打乱顺序，最终训练得到的 LRM 在不同环境的测试集上结果如图 5.12 所示。

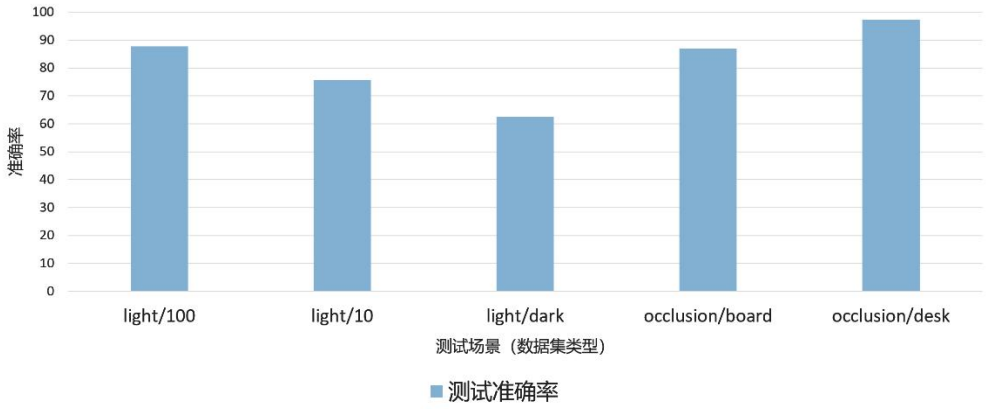


图 5.12　ViFi LRM Collect 训练在不同场景数据集上测试结果

可以看到 Collect 方式训练的 ViFi LRM 预训练模型在非极端环境（dark 光照度数据集）上都可以实现超过 95%的识别准确率，即便在几乎完全黑暗的情况下，ViFi LRM 也能正确的识别出 93.75%的动作。本文实验中也尝试将使用其他场景预训练得到的权重进行跨域测试，输入经过算法 1 预处理过的书架遮挡数据，测试结果准确率达到 86.98%。

5.4　对 ViFi 感知模型的深入分析与优化

本文在对 ViFi 感知模型进行实验和测试的过程中发现过一些问题，它们或许是因为时间原因无法再进行更深层次的对比实验，或是对实验结果仍然不够满意。

尽管对视频数据帧增加 YOLO 目标检测预处理后可以让模型更聚焦在被识别的人体上，从而舍弃周围环境对模型训练造成的影响，从结果上来看这一预处理也对视觉单模态和最终的多模态模型识别准确率都有一定的提升，但是通过检测和裁剪的方式也放弃了人体在视觉图像帧中的位置信息，在一些粗粒度动作（如走路、跑步等）的识别中，人体在摄像头中从一边到另一边的速度也可以作为区分不同动作的依据之一，但由于深

度学习是黑箱模型，并不能知道在特征学习的过程中它是否会学习到特征主体在图像中的位置变化信息，所以这一点对模型是否有影响还有待考证。

添加 YOLO 目标检测预处理的时间开销在前文中一直没有被统计，因为这步预处理通常是在整理数据集时就被执行，而非模型训练过程中，这也导致对模型的实时性计算并不够准确。整个模型运算时间的主要开销在视觉部分，如前文中图 5.9 所示，ViFi 模型对批进行处理和预测时仅需要花费 0.046 秒，而在输入输出处理时需要近千倍的时间，所以如果将处理流程进行优化，将数据仅读入一次并连续的经过预处理、解码、模型计算，可能会在实时性方面有更好的效果。

同样在 YOLO 目标检测的过程中，对于一些极端的实验环境的数据并没有很好的识别效果，例如在接近全黑暗的 dark 数据集中，一组 30 帧的视觉动作数据中只有约 5 张图片中可以检测到人体且可信度超过 50%。在实验中对这种情况时是将整个场景都无法识别到人的视觉帧不经过预处理直接输入给模型，当其中有部分识别到人体时视为其他数据质量较低直接从数据集中丢弃。在多模态融合的权重设置时，本文曾考虑过将 YOLO 目标检测的结果作为权重设置的直接相关因素，根据视频数据帧中每一次采集到的 30 张图片中能检测到人体的图片所占的比重，自动调整在多模态融合时视觉与 Wi-Fi 两部分的权重值。根据在本文实验中的结果来看，视觉仍然是识别准确率较高的一个，所占比重在大多数场景下也更高。而目标检测预处理的结果通常也会直接反映实验环境是否对视觉采集有较大的影响。当环境较极端时，目标检测时识别到的人体照片占所有采集的视觉帧也会更低，在系统捕获到这一信息后，降低模型中视觉部分结果在多模态融合中的权重，最终实现自动取得一个更适合当前环境的 ViFi 模型。同时，在本文实验中一些场景会出现其中某一模态的效果超前，可以直接近似作为最终的模型使用的情况，此时就可以将调整模态融合的权重改为直接进行模态的选择，生成根据环境情况动态升降的多模态模型。

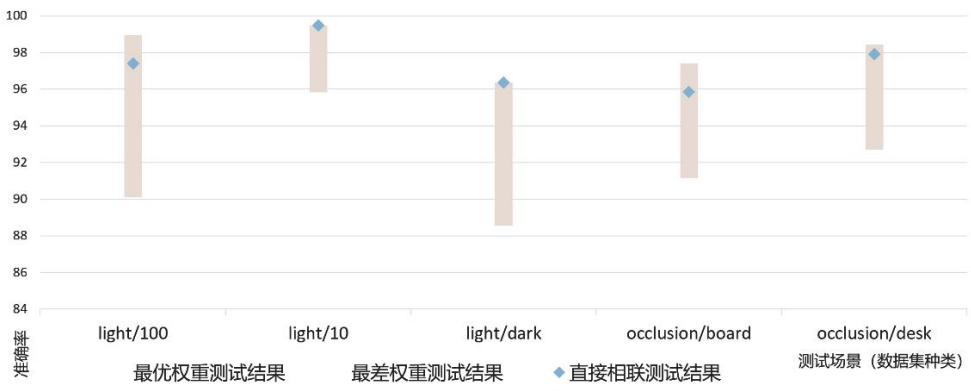


图 5.13　ViFi 模型直接相联与权重的最优最差测试结果对比

在本文对多模态融合的消融实验中（如图 5.13），虽然最终选用的直接相联方式较为稳定，准确率也在最优权重测试结果和最差权重测试结果中较优的位置，但是仍然无法达到实验最优结果。而为每个场景单独调整权重虽然一般都可以获得一个更优于直接相联方式的测试结果，但是由于对每个场景都需要单独微调，在需要花费大量时间的同时，不利于 ViFi 模型在 LRM 方面进行统一，对不同场景的鲁棒性和泛用性低。对于最终多模态模型在模态融合方式的选择需要因地制宜的进行取舍。

5.5 未来展望

随着计算机软件和硬件的迭代发展，算力在过去几年有了飞跃式的提升，这使得一些大模型能更便捷的训练和使用。在人体行为感知方向上，本文期望可以有一个对多种动作在多种不同的场景下通用的预训练模型参数出现，使其可以应对更为复杂多变的环境。实验中模型训练的动作只有 4 种粗粒度动作和 4 种细粒度动作，但在实际生活中人会有许多不同或相似的动作，都会对模型的实际应用造成影响。想要实现自学习的大型人体动作识别模型，就需要在采集时不给数据添加标签类型，让模型通过无监督的方式通过与环境的交互学习，结合强化学习算法来进一步提升性能，使模型可以通过自主探索和试错来提高自身的动作理解能力，并将学习到的结果泛化到更多的人体动作识别类型上。

在现实应用中，往往面临着样本稀缺的情况。未来的动作识别模型可能会更好地适应小样本学习的挑战，通过少量样本就能快速学习和泛化到新的动作类别或场景。在本文的实验中，也考虑过视觉模型是以图片流的三维数据进行输入，而 Wi-Fi CSI 数据只采用了一个天线接收到的振幅信号。可以将三条接收天线接收到的信息作为第三个输入层的维度，并且将无线信号的相位等信息也作为模型训练输入的一部分，通过传统方法将振幅和相位综合考虑。并且在时空域卷积网络论文中提到，可以对 Wi-Fi 天线采集的 CSI 数据设置滑动窗口和步长，将一条数据拆分成多个用来训练模型，从而解决样本量不足的问题。

当涉及到动作识别模型时，跨领域迁移学习是一种有潜力的方法，可以提高模型的适用性并减少对大量标注数据的需求。在跨领域迁移学习中，模型可以通过在一个领域中学习到的知识和模型参数，迁移到另一个相关领域中进行动作识别。传统的动作识别模型通常需要大量标注数据才能进行训练，这对于每个特定领域都需要耗费大量的时间和人力成本。然而，通过跨领域迁移学习，可以利用一个已有的数据集（称为源领域）来训练一个动作识别模型。然后，这个模型可以迁移到目标领域中，该领域可能没有足够的标注数据可用。在迁移过程中，模型会保留从源领域学到的知识和模型参数，并将

其应用于目标领域。通过这种方式，模型可以利用源领域中的数据和知识来辅助目标领域中的动作识别任务。这种迁移可以帮助填补目标领域中的数据缺失，并提供有关动作的一般特征和模式。总而言之，跨领域迁移学习为动作识别模型提供了一种有效的方法，通过利用已有的数据和知识，减少对大量标注数据的需求，并在相关领域中提高模型的适用性。这种方法有助于加速模型的训练过程并提高动作识别的性能，解决模型在不同场景或不同动作的泛用性的问题，同时降低了实际应用中的成本和难度。

实时性和效率是未来动作识别模型关注的重要方面。随着应用场景的多样化和对实时性的需求增加，模型需要能够在高实时性要求下进行快速准确的动作识别，并具备高效的计算和存储能力。为了提高实时性，未来的动作识别模型可能会采用轻量级的网络结构和优化算法，以减少模型的计算负担和响应时间。这包括模型压缩和加速技术，例如模型剪枝、量化和分布式推理等方法，以提高模型的推理速度，使其能够运行在资源受限、更低功耗的边缘计算设备上。此外，模型的存储和传输效率也是实时性和效率的关键因素。依据前文中图 5.9(a)，数据传输的速度在 ViFi 模型的时间开销中占比较高，未来的动作识别模型可能会采用模型压缩和量化技术，以减少模型的存储空间和传输带宽需求。同时，模型的离线训练和在线推理策略可以更好地平衡计算和存储资源的使用。综上所述，未来的动作识别模型将致力于提高实时性和效率，以满足不同应用场景的需求。通过采用轻量级结构、优化算法、硬件加速和模型压缩等技术，模型可以实现快速准确的动作识别，并在计算和存储资源受限的环境中高效运行。这将推动动作识别技术在实际应用中的广泛应用和发展。

目前的动作识别模型在真实世界的应用方面有很深的潜力，随着技术的进步和研究的深入，模型将能够在非常多领域实现更智能、更个性化的应用，为现实世界带来更多的便利和创新。

结　论

常规的人体行为感知模型通常只能在正常光照无遮挡的场景下识别，但是在实际的应用场景中会遇到光线变暗、物体遮挡等不可预知的情况，这就导致它无法做到更高的泛用性。为了解决在特殊场景下的人体动作识别，本文提出了 ViFi 多模态感知模型，通过视觉与 Wi-Fi 协同工作，在特殊场景使用 Wi-Fi 进行辅助识别处理。本文的主要贡献在于：

（1） ViFi 模型对视频摄像头采集到的样本使用基于 YOLO 的视觉目标检测作为数据预处理的一部分，使模型聚焦在人体并减少周围环境的干扰,相比目前最先进的 GaitFi CRNN 模型在视觉模态上提升了 1.56%到 7.81%。

（2） 在 Wi-Fi 特征提取部分对时空域分别添加扩张卷积，在时间连续性和空间分布上进行感受野更大的特征提取，最终得到了非常好的效果，提升幅度在 3.65%到 11.46%之间，并且在对极端场景下的动作识别效果有超过 20%的提升。

（3） 在对比实验部分将多种模态融合方式在不同场景下进行消融实验，最终选定效果最为稳定的直接相联融合方式。

（4） 本文提出将多种场景下数据合并混杂训练，得到一种可以应对多种复杂场景的大型动作识别模型预训练权重，在不同光照度和不同物体遮挡的场景都取得了较好的效果，在非极端环境下可以识别超过 95%的动作，即便在接近完全黑暗的场景中也有 93.75%的准确率。通过这种方式可以使用同一个预训练权重应对在不同光照度、不同物体遮挡场景下的人体动作数据而不再需要对其进行数据标记和训练，有更高的鲁棒性和泛用性。

ViFi 感知模型在智能家居、安防监控、数字娱乐等领域均展示出了巨大的潜力。然而，由于其实时性和泛用性仍有待提高，这导致了其在满足一些应用需求时仍有局限性。在提升 ViFi 模型的性能方面，系统优化、权重调整以及动态选择升降模式等方法被认为是具有有效性的策略。这些策略能够在各个层面上提升模型的效果，从而进一步增强其实时性和泛用性。同时，小样本学习和跨领域迁移学习等方法也被认为是提高 ViFi 模型在人体行为感知适应性方面的有效途径。这些方法能够帮助模型更加精确地理解和识别人体行为，从而扩大模型在实际应用中的适用范围。总结而言，尽管 ViFi 感知模型在某些应用方面仍有不足，但是通过持续的优化和学习，有望进一步挖掘其潜力，从而使其在更多的实际应用领域中得到广泛应用。

参 考 文 献

[1] KHURANA R, KUSHWAHA A K S. Deep learning approaches for human activity recognition in video surveillance-a survey[C]//2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC). IEEE, 2018: 542-544.

[2] BACCOUCHE M, MAMALET F, WOLF C, et al. Sequential deep learning for human action recognition[C]//Human Behavior Understanding: Second International Workshop, HBU 2011, Amsterdam, The Netherlands, November 16, 2011. Proceedings 2. Springer Berlin Heidelberg, 2011: 29-39.

[3] TAYLOR G W, FERGUS R, LECUN Y, et al. Convolutional learning of spatio-temporal features[C]//Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part VI 11. Springer Berlin Heidelberg, 2010: 140-153.

[4] JI S, XU W, YANG M, et al. 3D convolutional neural networks for human action recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2012, 35(1): 221-231.

[5] 周升儒, 陈志刚, 邓伊琴. 基于 PoseC3D 的网球动作识别及评价方法[J]. 计算机工程与科学, 2023, 45(1): 95.

[6] YUE-HEI NG J, HAUSKNECHT M, VIJAYANARASIMHAN S, et al. Beyond short snippets: Deep networks for video classification[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 4694-4702.

[7] WANG L, QIAO Y, TANG X. Action recognition with trajectory-pooled deep-convolutional descriptors[C]//Proceedings of the IEEE conference on computer vision and pattern RECOGNITION. 2015: 4305-4314.

[8] KARPATHY A, TODERICI G, SHETTY S, et al. Large-scale video classification with convolutional neural networks[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2014: 1725-1732.

[9] 张执着. 低分辨率图像下的二维人体姿态估计算法研究[D]. 北京: 北京交通大学, 2021.

[10] GU F, KHOSHELHAM K, VALAEE S. Locomotion activity recognition: A deep learning approach[C]//2017 IEEE 28th annual international symposium on personal, indoor, and mobile radio communications (PIMRC). IEEE, 2017: 1-5.

[11] PARK S U, PARK J H, AL-MASNI M A, et al. A depth camera-based human activity recognition via deep learning recurrent neural network for health and social care services[J]. Procedia Computer Science, 2016, 100: 78-84.

[12] YOUSEFI S, NARUI H, DAYAL S, et al. A survey on behavior recognition using WiFi channel state information[J]. IEEE Communications Magazine, 2017, 55(10): 98-104.

[13]CHEN Z, ZHANG L, JIANG C, et al. WiFi CSI based passive human activity recognition using attention based BLSTM[J]. IEEE Transactions on Mobile Computing, 2018, 18(11): 2714-2724.

[14]CAO Y, WANG F, LU X, et al. Contactless body movement recognition during sleep via WiFi signals[J]. IEEE Internet of Things Journal, 2019, 7(3): 2028-2037.

[15]杨旭. 基于 WiFi 的室内人员非接触式感知方法研究[D]. 徐州: 中国矿业大学, 2021.

[16]LIN C, XU T, XIONG J, et al. Wiwrite: An accurate device-free handwriting recognition system with COTS wifi[C]//2020 IEEE 40th International Conference on Distributed Computing Systems (ICDCS). IEEE, 2020: 700-709.

[17]ZHENG Y, ZHANG Y, QIAN K, et al. Zero-effort cross-domain gesture recognition with Wi-Fi[C]//Proceedings of the 17th annual international conference on mobile systems, applications, and services. 2019: 313-325.

[18]张东恒. 基于无线信号的室内人体感知技术研究[D]. 成都: 电子科技大学, 2021

[19]XUE H, JIANG W, MIAO C, et al. DeepMV: Multi-view deep learning for device-free human activity recognition[J]. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2020, 4(1): 1-26.

[20]ZHANG J, TANG Z, LI M, et al. CrossSense: Towards cross-site and large-scale WiFi sensing[C]//Proceedings of the 24th annual international conference on mobile computing and networking. 2018: 305-320.

[21]ZOU H, YANG J, PRASANNA DAS H, et al. WiFi and vision multimodal learning for accurate and robust device-free human activity recognition[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. 2019: 0-0.

[22]ALAMGIR F M, ALAM M S. Hybrid multi-modal emotion recognition framework based on InceptionV3DenseNet[J]. Multimedia Tools and Applications, 2023: 1-28.

[23]HUANG Z, LIU F, WU X, et al. Audio-oriented multimodal machine comprehension via dynamic inter-and intra-modality attention[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2021, 35(14): 13098-13106.

[24]GAO P, JIANG Z, YOU H, et al. Dynamic fusion with intra-and inter-modality attention flow for visual question answering[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 6639-6648.

[25]DENG L, YANG J, YUAN S, et al. Gaitfi: Robust device-free human identification via wifi and vision multimodal learning[J]. IEEE Internet of Things Journal, 2022, 10(1): 625-636.

[26]VOLPI R, MORERIO P, SAVARESE S, et al. Adversarial feature augmentation for unsupervised domain adaptation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 5495-5504.

[27]ZOU H, ZHOU Y, YANG J, et al. Consensus adversarial domain adaptation[C]//Proceedings of the AAAI conference on artificial intelligence. 2019, 33(01): 5997-6004.

[28]LU X, WANG L, TIAN Y, et al. Towards WiFi-based Real-time Sensing Model Deployed on Low-power Devices[C]//2022 IEEE 19th International Conference on Mobile Ad Hoc and Smart Systems (MASS). IEEE, 2022: 385-393.

[29]REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.

[30]OUYANG X, SHUAI X, ZHOU J, et al. Cosmo: contrastive fusion learning with small data for multimodal human activity recognition[C]//Proceedings of the 28th Annual International Conference on Mobile Computing And Networking. 2022: 324-337.

[31]ZHAO Y, TU P, CHANG M C. Occupancy sensing and activity recognition with cameras and wireless sensors[C]//Proceedings of the 2nd Workshop on Data Acquisition To Analysis. 2019: 1-6.

[32]WANG F, ZHOU S, PANEV S, et al. Person-in-WiFi: Fine-grained person perception using WiFi[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 5452-5461.

[33]SHENG B, GUI L, XIAO F. Ts-net: Device-free action recognition with cross-modal learning[C]//Wireless Algorithms, Systems, and Applications: 16th International Conference, WASA 2021, Nanjing, China, June 25–27, 2021, Proceedings, Part I 16. Springer International Publishing, 2021: 404-415.

[34]SHENG B, SUN C, XIAO F, et al. MuAt-Va: Multi-attention and Video-auxiliary Network for Device-free Action Recognition[J]. IEEE Internet of Things Journal, 2023: 10870-10880.

[35]GUO J, SHI M, ZHU X, et al. Improving human action recognition by jointly exploiting video and WiFi clues[J]. Neurocomputing, 2021, 458: 14-23.

[36]WU Z, ZHANG D, XIE C, et al. RFMask: A simple baseline for human silhouette segmentation with radio signals[J]. IEEE Transactions on Multimedia, 2022: 1-12.

[37]YU C, WU Z, ZHANG D, et al. Rfgan: Rf-based human synthesis[J]. IEEE Transactions on Multimedia, 2022: 1-1.

[38]HALPERIN D, HU W, SHETH A, et al. Tool release: Gathering 802.11 n traces with channel state information[J]. ACM SIGCOMM computer communication review, 2011, 41(1): 53-53.

修改记录

一、毕业设计（论文）题目修改

原题目：基于快速组网验证算法的优化算法

修稿后题目：基于 Wi-Fi 和视觉的多模态行为识别方法研究

二、毕业设计（论文）内容重要修改记录

第一次修改记录：

括号格式，修改前：中英文括号混用。

修改后：全部使用中文括号，在第一次出现的专有名词缩写后添加详细解释。

模型架构图，修改前：缺少输入输出示例。

修改后：模型图添加数据集样例，修改模型介绍部分结构和顺序。

模型的 Wi-Fi 模块命名方式，修改前：直接使用原论文中模型名称 LiST。

修改后：所有 Wi-Fi 模型更改为以模型特点为命名依据，使用基于多尺度的 Wi-Fi 感知模型或时空域卷积网络替换。

第二次修改记录：

模型介绍部分，修改前：模型介绍。

修改后：添加模型总体架构介绍，对每一个模块进行简单介绍。

论文格式，修改前：列表换行没有顶头，公式变量没有统一。

修改后：修改列表格式，统一字母大小，全文统一变量指代内容，图标添加标注。

专有名词语言，修改前：使用英文缩写进行表述。

修改后：论文中文字和图片中的英文专有名词改为中文对应的名称。

第三次修改记录：

摘要和结论，修改前：内容较为混乱。

修改后：重写了摘要和结论部分，强调重要性和模型设计。

相关工作部分，修改前：多模态部分较少。

修改后：单模态和多模态添加相关工作的结论，添加多模态部分的相关文献。

三、毕业设计（论文）正式检测重复比

去除本人文献复制比：4.5%

记录人（签字）：

指导教师（签字）：

致　　谢

转眼间，校园生活即将结束。此篇论文完稿之际，要感谢众多师长和亲友，谢谢你们的期望与鼓励。此时此刻，我无法找到合适的言语来表达我内心深处最真挚的谢意。

首先衷心感谢我的毕设导师徐秀娟老师、wilna 实验室王雷老师、研究生导师张鹏老师、博士生卢欣欣学姐、硕士生韩斌学长等人的教诲，一直给予我生活、学习上的帮助，他们严谨的治学态度，帮助我学习和理解文献、修改毕业论文时的认真细致，仍然深深刻印在我脑海中，挥之不去，难以忘怀。

感谢同窗的各位同学。在我的大学四年时间里，在团学、社团、心协、科研和竞赛等地方都留下了一些自己的印记，也借此机会结识了非常多朋友，人数太多我无法在这里一一列举。他们为我在学习和生活中提供了大量的无私帮助，我也非常幸运得以在本科期间过的如此充实。感谢 DV 工作室在我毕设的实验阶段提供用于训练和计算的电脑设备和环境，能让我在有限的时间内对模型进行足够多的测试。

我要感谢我的父母多年来默默的支持、理解、信任和期盼，这是我一直前行的动力。在我对科研道路迷茫时，是他们在电话里鼓励我，让我勇敢的迈出下一步。

最后，由于我的学术水平有限，所写论文难免有不足之处，恳请各位老师和同学提出批评和指正。