学校代码：10284
分 类 号：TP181
密　　级：公开
U D C：004.8
学　　号：MG21370049

南京大学

硕 士 学 位 论 文

论 文 题 目　基于对比学习的八段锦动作质量评估

作 者 姓 名　张耕

专 业 名 称　计算机科学与技术

研 究 方 向　计算机视觉

导 师 姓 名　申富饶 教授

2024 年 5 月 21 日

答辩委员会主席 戴新宇 教授

评　　阅　　人 张剡 高工

徐明华 教授

论文答辩日期　2024 年 5 月 16 日

研究生签名:

导师签名:

Baduanjin Action Quality Assessment based on Contrastive Learning

by

Zhang Geng

Supervised by

Professor Shen Fu-Rao

A dissertation submitted to

the graduate school of Nanjing University

in partial fulfilment of the requirements for the degree of

MASTER

in

Computer Science and Technology

School of Artificial Intelligence

Nanjing University

May 21, 2024

南京大学研究生毕业论文中文摘要首页用纸

毕业论文题目：基于对比学习的八段锦动作质量评估

计算机科学与技术 专业 2021 级硕士生姓名：张耕

指导教师（姓名、职称）：申富饶 教授

摘　要

人们处于快节奏的生活方式之中，忽视了对自身健康的管理，近年来全球肆虐的新冠疫情给人们敲响了警钟，重新唤醒了人们对身体健康的重视。除了在健身房中健身外，中国的太极拳、八段锦等传统运动也可以起到锻炼身体的作用，且老少咸宜。但仅靠自己很难发现自身动作的不足，如果可以足不出户就能得到专业的指导，对自身的锻炼会有极大的促进作用。因此，本文聚焦于八段锦动作质量评估这一任务，分别从数据集、算法和系统三方面进行展开。

本文建立了第一个同时对每个动作起止时间及动作质量进行标注的八段锦骨骼动作质量评估数据集。该数据集内共有 178 个样本视频，总时长约 36 小时，总帧数为 3746750 帧。为了便于八段锦动作起止时间及动作质量的标注，本文开发了一个辅助标注系统，该系统同样适用于辅助其他需要标注动作起止时间及动作质量的数据集。

本文借鉴人类认识事物的原理，将样本的粗粒度标签信息加以利用，提出了利用粗粒度标签的有监督对比学习损失，将骨干网络的损失函数替换为该损失后，同时在分类和回归形式的八段锦动作质量评估任务上取得了优异的表现。该损失函数在其他公开的分类和回归数据集上同样获得了显著的效果提升，再次验证了本文提出的利用粗粒度标签的有监督对比学习损失的有效性。

本文利用上述建立的八段锦骨骼数据集及八段锦动作质量评估模型，开发了八段锦动作质量评估系统。用户通过拍摄自己的练习视频即可随时随地地获取动作质量打分，了解自身动作的不足，以便于达到更好的锻炼效果。

关键词：八段锦；动作质量评估；对比学习；粗粒度标签；图卷积神经网络

南京大学研究生毕业论文英文摘要首页用纸

THESIS: Baduanjin Action Quality Assessment based on Contrastive Learning

SPECIALIZATION: Computer Science and Technology

POSTGRADUATE: Zhang Geng

MENTOR: Professor Shen Fu-Rao

ABSTRACT

Living in a fast-paced lifestyle, people often neglect their health management. The global COVID-19 pandemic in recent years has sounded an alarm bell, reawakening people's attention to physical health. In addition to gym workouts, traditional Chinese exercises such as Tai Chi and Baduanjin can also help people of all ages to exercise their bodies. However, it is difficult to identify one's action flaws during self-practice. If professional guidance can be obtained without leaving home, it will greatly promote the exercise effect. Therefore, this paper focuses on the task of Baduanjin action quality assessment and conducts research from three aspects: dataset, algorithm, and system.

This paper constructs the first Baduanjin skeleton action quality assessment dataset that simultaneously annotates the start and the end time of each action and the action quality. The dataset contains 178 sample videos with a total duration of about 36 hours and a total of 3 746 750 frames. To facilitate the annotation of Baduanjin action start and end time and action quality, this paper develops an auxiliary annotation system, which is also applicable to other datasets that need to annotate action start and end time and action quality.

Inspired by the way humans perceive things, this paper utilizes the coarse-grained label information of the samples and proposes a supervised contrastive learning loss with coarse-grained labels. After replacing the loss function of the backbone network with this loss, the model achieves excellent performance on both classification and regression tasks for Baduanjin action quality assessment. This loss

function also demonstrates significant performance improvements on other publicly available classification and regression datasets, further validating the effectiveness of the proposed supervised contrastive learning loss with coarse-grained labels.

This paper develops a Baduanjin action quality assessment system based on the Baduanjin skeleton dataset and the Baduanjin action quality assessment model mentioned above. Users can obtain action quality scores and understand their action flaws anytime and anywhere by shooting practice videos, achieving better exercise effects.

KEYWORDS: Baduanjin; Action Quality Assessment; Contrastive Learning; Coarse-grained Labels; Graph Convolutional Neural Network

目　　录

插图目录

表格目录

第一章　绪论

1.1　八段锦动作质量评估的研究背景及研究意义

现代社会的高速发展给人们带来极大便利的同时，也带来了极大的身心压力。快节奏的生活方式使人们忽视了对自身健康的管理，久坐久站也使得人们的颈椎腰椎处于亚健康状态。然而，根据马斯洛的需求层次理论，健康属于人类需求层次中最基本的生理需求，是实现其他高层次需求的基础。近年来，全球肆虐的新冠疫情给人们敲响了警钟，重新唤醒了人们对身体健康的重视。

一提到健身，很多人首先想到的是在健身房中挥汗如雨，而不是太极拳和八段锦等“慢悠悠”的中国传统运动，甚至有不少人觉得太极拳和八段锦只适合公园里晨练的老年人。其实，太极拳和八段锦等中国传统运动不仅适合老年人，也适合各个年龄段的人群。太极拳和八段锦动作舒缓，对关节和肌肉的冲击较小，不同年龄段的人可以根据自身情况调整练习的强度，获得适合自己的锻炼效果。年轻人可以通过太极拳和八段锦提高身体素质，改善睡眠和肩颈健康[1-2]；而对于老年人来说，适当的锻炼可以帮助保持身体的灵活性、稳定性和协调性，延缓肌肉和关节的衰老，提升生活质量[3-4]。

与此同时，2021 年国务院印发了《全民健身计划（2021 - 2025 年）》，其中提到要提升科学健身指导服务水平、推动体育产业高质量发展、营造全民健身社会氛围和提供全民健身智慧化服务。这一计划的出台意味着政府对全民健身的重视，并将全民健身纳入国家发展的重要议程。在 2024 年春晚的舞台上，著名歌手周深演唱了一首名为《健康到到令》的歌曲（如图1-1所示），并伴随着改编自八段锦的舞蹈动作。这一表演不仅彰显了八段锦作为中国传统运动的魅力，也对全国观众发出了健康生活的呼吁。

八段锦是一种源自中国古代的健身功法，起源可以追溯到北宋时期，至

图 1-1 健康到到令

今已有八百多年的历史。它由八个不同的动作组成（不计入“预备式”和“收式”），因此得名为“八段锦”。目前，人们学习八段锦的方式主要是观看视频教程、阅读文字书籍以及接受他人的指导。尽管通过模仿视频中的标准动作，人们可以在较短的时间内初步学习这些动作，但难免会存在一些不规范的地方，而这些问题可能很难自己察觉。而依赖于他人的指导，个人的动作水平也会受到指导者的影响，且并非所有人都能够获得专业人士的指导。因此，本文旨在利用深度学习的方法开发一套八段锦动作质量评估系统，使人们能够在家中获得较为准确的动作质量评估，从而纠正自身动作的不足。

通过动作质量评估，锻炼者可以了解自身的姿势和动作幅度是否正确，以确保身体负荷和刺激适当，从而提高训练效果。此外，不正确的动作可能会增加运动损伤的风险，纠正错误动作还可以减少运动损伤，这对于长期练习八段锦的人群来说尤为重要，可以保障他们免于“好心办错事”。

研究八段锦动作质量评估的意义并不局限于八段锦本身，动作质量评估可以应用于各个领域，例如专业运动员的训练[5-6]、课堂教学[7-8]、体育比赛评分（如图1-2所示的东京奥运会体操 AI 辅助评分系统）和愈后康复训练[9]等。

图 1-2 东京奥运会体操 AI 辅助评分系统

1.2 八段锦动作质量评估的研究现状及难点分析

1.2.1 研究现状

动作质量评估研究现状

动作质量评估是评价一个动作执行得有多好，并对其打分的任务。动作质量评估方法大致可以分为两种：基于人工特征的模型和基于深度学习的模型，其中基于人工特征的模型在 2014 年之前比较常见。

Gordan 等人[10]首先提出了动作质量评估任务，但限于当时计算机硬件性能及算法发展水平，文章更多地是提出一种对自动动作质量评估的设想，并以体操中的跳马动作为例分析了动作质量评估的可行性。Ilg 等人[11]提出了一种模型，用分层算法建立学习的原型示例序列与新轨迹之间的时空对应关系，以便于进行样本之间的动作质量比较。C¸eliktutan 等人[12]提出了一种基于图的方法来对齐两个动态骨骼序列，同时进行动作识别任务和动作质量评估任务。Pirsiavash 等人[13]更为清晰地形式化了动作质量评估任务，并使用基于姿势的特征、时空兴趣点和分层卷积特征完成分数的计算。

在 2014 年之后，动作质量评估也进入了深度学习时代。研究人员进行了一系列工作，探索将 RGB 视频或骨骼序列作为输入，使用深度学习模型评估动作质量。

一些研究工作使用 RGB 视频作为输入，利用 CNN[14-15]、LSTM[16]等模型进行评估。Li 等人[17]开发了一个端到端的动作质量评估框架，提取 C3D 特征[18]来回归分数。Gao 等人[19]对进行交互动作的个体间的非对称关系进行建模。他们将个体分为主要个体和次要个体，使模型在评估执行动作的主体时表现更好。Dong 等人[20]提出了一种多阶段回归模型（MSRM），用于从视频的不同隐藏子阶段学习和融合特征以进行动作质量评估。Wang 等人[21]提出了一种用于动作质量评估的管道自注意网络。他们引入了一个单一对象跟踪器，使模型能够连续区分和分析前景与背景信息以进行动作质量评估。

另一些方法则使用骨骼序列作为输入，骨骼序列可以通过深度相机（如 Kinect[22]）捕捉到，或者可以使用姿态估计算法从 RGB 视频中提取。Li 等人[23]的工作提出了时空姿势提取模块和动作间时序关系提取模块，对输入的

骨骼时间序列进行动作质量评估。Pan 等人[24]使用关节关系图来进行质量评估，通过关节共性模块和关节差异模块分别学习特定身体部位的一般运动规律及运动差异规律。Pan 等人[25]在之后的工作中又提出了一种自适应动作评估系统，根据关节间的交互自动构建适用于不同动作类型的不同评估架构。在 Yan 等人[26]提出时空图卷积神经网络（ST-GCN）用于骨骼动作识别任务后，Bruce 等人[27]也将其迁移到了骨骼动作质量评估任务中。此后出现了更多利用图卷积神经网络的工作，例如 Bruce 等人[28]之后又提出了双任务图卷积神经网络（2T-GCN），同时进行异常检测和质量评估任务。Liu 等人[29]提出了一种集成学习的图卷积神经网络（EGCN），同时使用骨骼位置流和骨骼方向流作为输入，在数据、特征、抉择和模型层级进行集成学习。而 Lei 等人[30]的工作则提出了一种多骨骼结构的图卷积神经网络，同时对关节的自连接、身体某部分内的连接和身体各部分之间的连接进行建模。

大多数工作均将动作质量评估看作回归任务，但除了对网络结构的改进外，也有一些工作探索了更适合动作质量评估任务的损失函数。Li 等人[31]和 Xu 等人[32]的工作均引入了对比学习的思想，将网络学习的任务改为了学习一个样本相对于另一个样本的相对分数。而 Wang 等人[33]则将回归损失改为了 triplet 损失[34]，同时考虑正负样本。

越来越复杂的网络结构导致了计算复杂度的上升，不利于模型在实际使用时达到实时推理的速度。因此相比于对网络结构进行改进，本文更倾向于对损失函数进行改进，在几乎不增加计算量的前提下提升模型的表现。同时，对损失函数的改进是与模型无关的，这意味着几乎所有的动作质量评估模型都可以使用本文提出的损失函数进行训练，获得更好的性能。

八段锦动作质量评估研究现状

目前，虽然已经有关于太极拳的动作质量评估研究[35-37]，但八段锦动作质量评估的研究尚处于初级阶段，尽管该功法在中国已有悠久的历史和广泛的实践应用，但专门针对八段锦动作质量评估的研究相对较少。截至目前为止，没有以八段锦为内容的公开数据集，陈静[38]在其论文中开发了一个八段锦动作质量评估系统，但其建立的数据集中练习者数量较少，且质量标注仅为较差、一般和良好三级，文中质量评估实际上是多分类问题。

在应用市场中可以搜索到“跟我练八段锦”（如图1-3所示）、“健身八段锦”和“智能八段锦”等八段锦 APP，但是其中前两者只提供了视频跟练或者文字说明，而后者虽然提供了关节坐标识别功能，但并没有显式地给出质量评分。

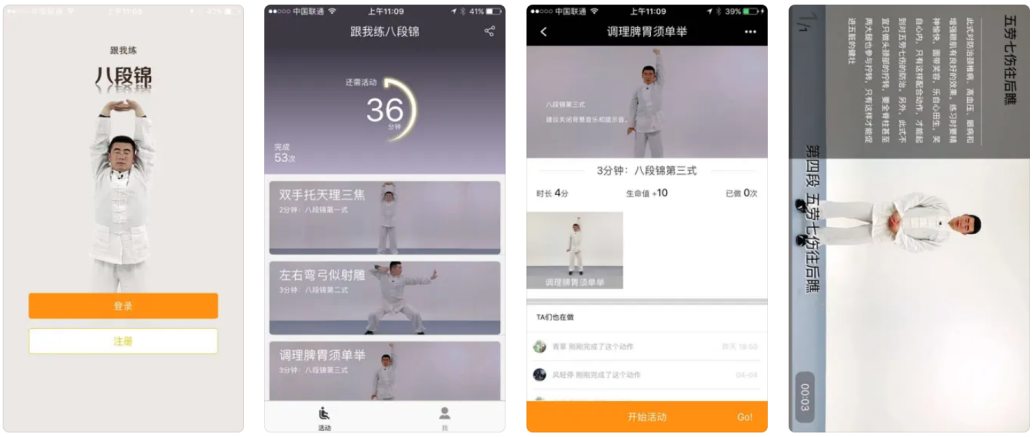


图 1-3 “跟我练八段锦”APP

1.2.2 难点

八段锦动作质量评估面临以下几个难点：

- 数据集制作：在目前的研究中，尚未拥有专门用于八段锦动作质量评估的数据集。因此，为了构建这样一个数据集，第一步是通过网络搜索和收集与八段锦相关的视频。这些视频可以来自各种在线视频平台、体育比赛录像和健身教程等。在收集过程中，需要确保视频的质量以及内容与八段锦动作密切相关。筛选完成后，需要对视频进行片段标注和质量标注。片段标注是指将视频划分为具有明确开始和结束时间的八段锦动作片段；质量标注即是评估每个片段的动作质量，并为其打分。
- 动作质量评估标准的制定：目前，八段锦动作质量评估尚缺乏一套统一的评估标准。不同研究和实践中可能采用不同的评估标准，如动作准确性、协调性、稳定性和幅度是否到位等。为了推动八段锦动作质量评估的研究，需要制定一套科学可行的评估标准，以便进行客观的评估和比较分析，最终完成对数据集中动作片段的质量分数标注。

- 长时间段动作质量评估：不同于跳水、体操等短时间动作，一整套八段锦动作整体时长约有十二分钟，每段动作时长约一两分钟，均显著长于大部分其他动作质量评估数据集中的动作。长时间段动作可能会导致数据的稀疏性和冗余性。因此，评估模型需要具备强大的时空建模能力，能够从多个时间尺度和空间层次上理解和分析动作；或者需要对评估模型进行一定的优化以降低计算复杂度。

1.3 本文贡献

首先，针对数据集缺失的问题，本文建立了第一个同时对每个动作起止时间及动作质量进行标注的八段锦骨骼动作质量评估数据集。本文参照资料制定了详尽的八段锦动作规范，为质量评估提供了理论依据，同时可以指导现实生活中的八段锦练习。该数据集内共有 178 个样本视频，总时长约 36 小时，总帧数为 3746750 帧。为了便于八段锦动作起止时间及动作质量的标注，本文开发了一个辅助标注系统，该系统同样适用于辅助其他需要标注动作起止时间及动作质量的数据集。

其次，本文借鉴人类认识事物的原理，将样本的粗粒度标签信息加以利用，提出了利用粗粒度标签的有监督对比学习损失，将骨干网络的损失函数替换为该损失函数后，同时在分类和回归形式的八段锦动作质量评估任务上取得了优异的表现。该损失函数在其他公开的分类和回归数据集上同样获得了显著的效果提升，再次验证了本文提出的利用粗粒度标签的有监督对比学习损失的有效性。针对长时间段的质量评估任务，本文使用了重采样的方式以降低计算复杂度。

最后，本文利用上述建立的八段锦骨骼数据集及八段锦动作质量评估模型，开发了八段锦动作质量评估系统。用户通过拍摄自己的练习视频即可随时随地地获取动作质量打分，了解自身动作的不足，以便于达到更好的锻炼效果。

1.4 论文结构

如图1-4所示，本文围绕着八段锦动作质量评估展开，共分为六章。第一章为绪论，简要介绍了研究八段锦动作质量评估的研究背景及研究意义；第二章介绍了对比学习、图神经网络和动作质量评估的相关基础知识；第三章则详细介绍了八段锦动作质量评估数据集的建立过程；第四章介绍了利用粗粒度标签的有监督对比学习算法，及其在八段锦动作质量评估数据集和若干公开数据集上的表现；第五章则将八段锦动作质量评估数据集及算法应用于演示系统中；第六章为总结和展望，对全文进行总结并列举本文工作的未来研究方向。

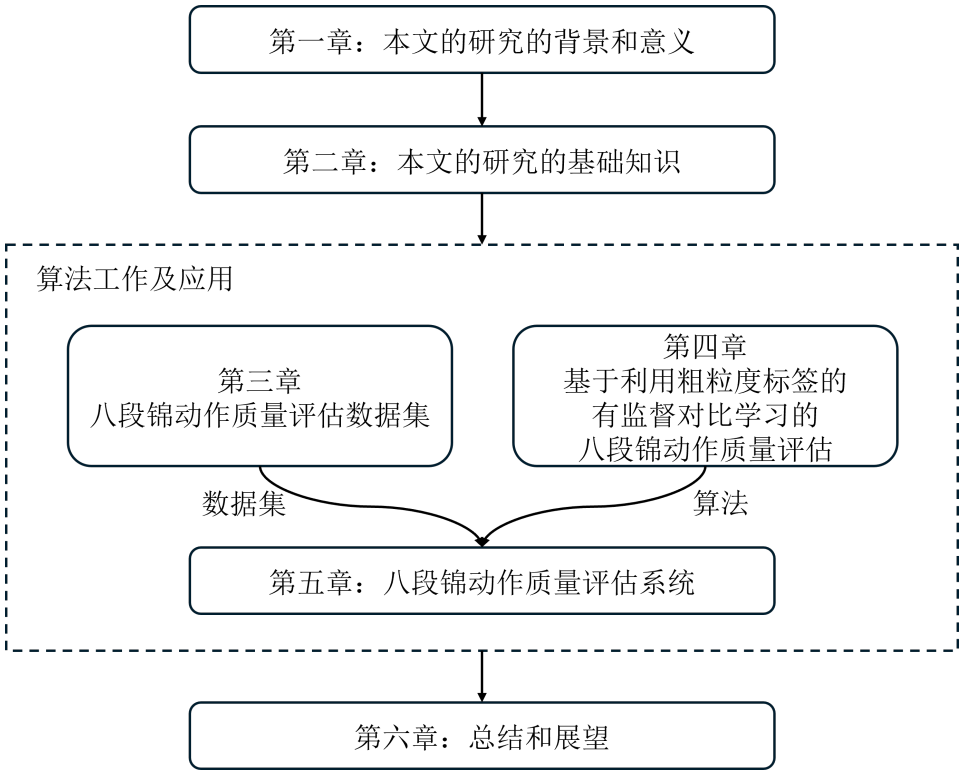


图 1-4 论文结构图

第二章　相关工作

本文提出的利用粗粒度标签的有监督对比学习方法属于对比学习的一种，因此本章将首先简介对比学习的基础知识，随后对八段锦动作质量评估用到的图卷积神经网络以及动作质量评估范式进行介绍。

2.1　对比学习

2.1.1　自监督对比学习

对比学习的思想非常简单，即学习一个从样本到嵌入特征（embedding）的映射网络，使得相似的样本在嵌入空间中的表示较近，而不相似的样本在嵌入空间中的表示较远。因此，需要一种相似度度量的方式来计算两个样本之间的相似度，最常用的一种计算方式就是余弦相似度：

$$\mathrm{sim}_{\cos}(a,b)=\frac{a\cdot b}{\|a\|\cdot\|b\|}. \tag{2-1}$$

谈到对比学习，就需要先了解 NCE（Noise Contrastive Estimation）损失[39]。NCE 损失是一种用于训练概率模型的损失函数，以估计离散的概率分布。它是一种非常有效的无监督学习方法，通常用于语言模型和词嵌入等任务中。例如在语言模型中，模型通常会在最后一层输出一个预测的词，这就需要使用 softmax 函数计算模型在词表所有词上的概率，当词表非常大的时候，由于 softmax 的分母会包含所有词的项，这会使得计算的开销变得很大。而 NCE 损失的思想在于通过采样，将在非常大的词表上做多分类的问题转换成了区分目标词和噪声词的二分类问题。NCE 损失的基本形式为：

$$\mathcal{L}^{\mathrm{NCE}}=-\log\frac{\exp(\mathrm{sim}(q,k_{+}))}{\exp(\mathrm{sim}(q,k_{+}))+\exp(\mathrm{sim}(q,k_{-}))}, \tag{2-2}$$

其中 $\text{sim}(\cdot,\cdot)$ 是相似度计算函数，如式2-1所述的余弦相似度；q 为待查询样本的嵌入特征，k_+ 为正样本嵌入特征，k_- 为负样本嵌入特征。在训练过程中，模型会逐渐调整自己的参数，以使正样本的预测概率更高，噪声样本的预测概率更低。

当引入更多的负样本时，即可以得到基于 NCE 损失改进的 InfoNCE 损失[40]：

$$\mathcal{L}^{\text{InfoNCE}} = -\log \frac{\exp(\text{sim}(q,k_+))}{\exp(\text{sim}(q,k_+)) + \sum_{i=1}^{K}\exp(\text{sim}(q,k_i))}, \tag{2-3}$$

其中 k_i 为负样本。

到此为止，InfoNCE 就已经非常接近对比学习广泛使用的损失函数形式，而 Chen 等人提出的 SimCLR[41]（A **Sim**ple Framework for **C**ontrastive **L**earning of Visual **R**epresentations）则是第一个将其成功运用到对比学习中的方法。

SimCLR 使用数据增强来获得一个样本对应的正样本，由于其不需要样本标签，所以属于自监督方法的一种。具体来说，在训练阶段数据加载器每次都会从数据集中随机采样得到一个大小为 N 的 batch，其中包含样本 $x_1 \cdots x_N$。而在 SimCLR 方法中，这 N 个样本中的每一个都会被独立地应用两次数据增强（如水平翻转、裁剪、旋转和颜色失真等），得到总共 $2N$ 个数据增强后的样本 $\tilde{x}_1 \cdots \tilde{x}_{2N}$，也称作数据增强后的双视图 batch。其中 $\tilde{x}_i$ 和 $\tilde{x}_{N+i}$ 由同一个样本数据增强得来，称为同源样本。如图2-1所示，这 $2N$ 个样本会通过神经网络模型（编码器）得到其对应的嵌入特征，通过投影头投影再进行归一化即可使得这 $2N$ 个嵌入特征分布在半径为 1 的超球面上。对于第 i 个样本（称作锚样本），可以计算其 SimCLR 损失：

$$\mathcal{L}_i^{\text{SimCLR}} = -\log \frac{\exp(z_i \cdot z_{j(i)}/\tau)}{\sum_{a\in A(i)}\exp(z_i \cdot z_a/\tau)}. \tag{2-4}$$

其中 $z_l = \text{proj}(\text{Enc}(\tilde{x}_l))$ 为数据增强后的样本经过编码器和投影头后归一化的嵌入特征，z_i 和 $z_{j(i)}$ 则是一对同源正样本；$A(i) = \{1 \cdots 2N\} \backslash \{i\}$ 为除了锚样本 i 之外的其它样本集合；$\tau \in \mathcal{R}^+$ 为温度系数，用于控制训练过程中对困难样本的关注程度，温度系数越小，模型越倾向于在优化过程中将某个样本

与其相似的困难样本区分开，得到更均匀的特征表示。然而困难样本中也有可能存在潜在的正样本，如果温度系数设置太小会将正样本也推远，不利于模型学习到好的表示。在温度系数趋于 0 时，SimCLR 将只关注最困难的负样本；而温度系数趋于无穷大时，SimCLR 对所有负样本一视同仁，不再关注困难样本。

图 2-1 自监督对比学习方法结构图

对所有样本的损失求和，即可得到最终的 SimCLR 损失：

$$\mathcal{L}^{\text{SimCLR}} = \sum_{i=1}^{2N} \mathcal{L}_i^{\text{SimCLR}} = -\sum_{i=1}^{2N} \log \frac{\exp(z_i \cdot z_{j(i)}/\tau)}{\sum_{a \in A(i)} \exp(z_i \cdot z_a/\tau)}. \tag{2-5}$$

2.1.2 三元组损失

与 SimCLR 不同，三元组（triplet）损失顾名思义，一次只同时考虑 3 个样本之间的关系：

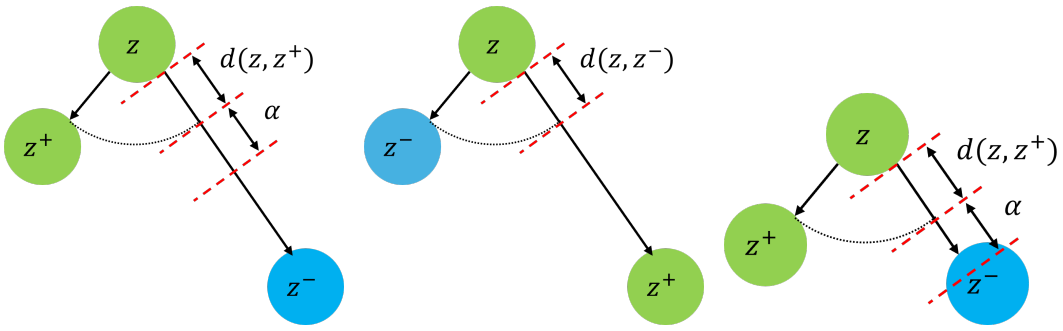
$$\mathcal{L}^{\text{triplet}} = \max(d(z, z_+) - d(z, z_-) + \alpha, 0), \tag{2-6}$$

其中 z 为锚样本，z_+ 为正样本，z_- 为负样本，$d(\cdot,\cdot)$ 为距离函数，α 为 margin。直观理解，锚样本与正样本之间的距离应该至少比其与负样本之间的距离小 α，α 控制了正负样本之间的差异。如果不设置 α，公式将变为

$$\mathcal{L}^{\text{triplet\_bad}} = \max(d(z, z_+) - d(z, z_-), 0), \tag{2-7}$$

在这种情况下，网络只需要使得 $d(z, z_+)$ 和 $d(z, z_-)$ 尽量接近，也就是锚样本与正样本之间的距离和其与负样本之间的距离差异很小就可以使得损失很低，此时达不到区分正负样本的目的。因此需要设置 α，迫使网络将正负样本区分开。

数据集中的样本根据损失的大小可以分为三类。第一类为如图2-2(a)所



(a) 简单三元组 (b) 困难三元组 (c) 半困难三元组

图 2-2 三种类型的三元组

示的简单三元组，满足 $d(z,z^+)+\alpha < d(z,z^-)$，即损失为 0，这种三元组已经满足要求，对训练没有帮助。第二类为如图2-2(b)所示的困难三元组，满足 $d(z,z^+) > d(z,z^-)$，即 $\mathcal{L}^{\text{triplet}} > \alpha$，这种情况下锚样本与正样本之间的距离大于其与负样本之间的距离，需要尽量优化。第三类为如图2-2(c)所示的半困难三元组，此时 $d(z,z^+) < d(z,z^-) < d(z,z^+)+\alpha$，即 $\mathcal{L}^{\text{triplet}} < \alpha$，锚样本与负样本之间的距离大于其与正样本之间的距离，但是差距未达到 α，$\mathcal{L}^{\text{triplet}}$ 依然大于 0，需要被优化。因此在训练过程中，应该尽量采样困难三元组和半困难三元组，帮助模型更好地学习到正负样本之间的差异，最后获得更好的特征表示。

2.2 图卷积神经网络

CNN 网络中的卷积运算过程如图2-3所示，可以看作卷积核矩阵（图中蓝色矩阵）在输入特征图（图中灰色矩阵）上移动，输入特征图中与卷积核重叠的部分（图中橙色部分）与卷积核计算哈达玛乘积再求和后得到输出特征图中的一个元素。

CNN 网络经常用在图像等具有网格结构的数据上，这类欧几里得数据中的每个节点的邻居数量是固定的（如图像中一个像素周围有 8 个像素），而且邻居的顺序也是固定的，因此可以用普通的卷积运算。而对于图这种非欧几里得数据来说，每个顶点的邻居数量是不固定的，且邻居之间也没有顺序可言，因此图卷积需要特殊的设计。

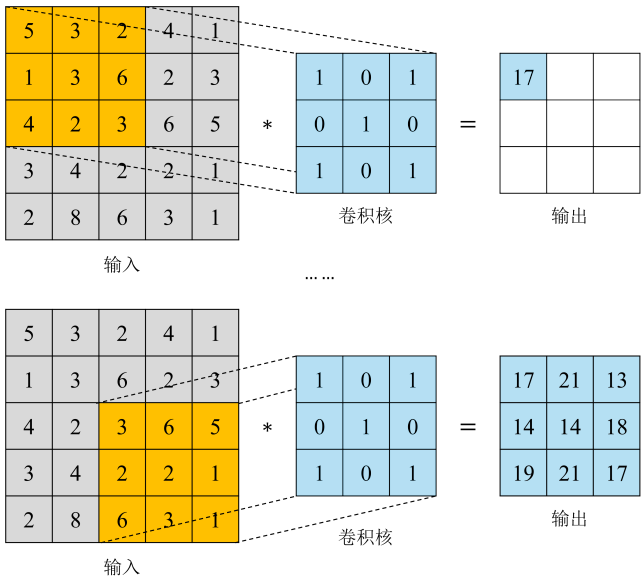


图 2-3 卷积计算过程示意图

2.2.1 图卷积的推导

要定义图上的卷积，就需要先定义图上的傅里叶变换，而定义图的傅里叶变换又需要先引入拉普拉斯矩阵。

对于由顶点集 V 和边集 E 构成的无向图 $G=(V,E)$，其拉普拉斯矩阵定义为：

$$L = D - A, \tag{2-8}$$

其中 D 为度矩阵，是一个对角线元素为顶点度的对角矩阵；A 为图 G 的邻接矩阵。具体而言，L 中的元素为：

$$L_{i,j} = \begin{cases} \deg(v_i) & , v_i = v_j \\ -1 & , v_i \neq v_j \text{ and } (v_i, v_j) \in E \\ 0 & , \text{otherwise} \end{cases}, \tag{2-9}$$

其中 $\deg(\cdot)$ 为顶点的度。以图2-4所示的无向图为例，其有 4 个顶点，图中顶

点上的数字为顶点的编号，其拉普拉斯矩阵为：

$$
\begin{aligned}
L &= D - A \\
&= \begin{bmatrix} 3 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} - \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \\
&= \begin{bmatrix} 3 & -1 & -1 & -1 \\ -1 & 2 & -1 & 0 \\ -1 & -1 & 2 & 0 \\ -1 & 0 & 0 & 1 \end{bmatrix}.
\end{aligned} \tag{2-10}
$$

矩阵的第一行代表 1 号顶点的属性，其中第一个元素代表 1 号顶点的度，剩余元素代表 1 号顶点是否与该位置对应的顶点相连。由于 1 号顶点与 2、3 和 4 号顶点都相连，因此其度为 3，其余元素均为-1。而对于 2 号顶点来说，其与 1 号和 3 号顶点相连，与 4 号顶点不相连，因此其度为 2，矩阵第二行第一个和第三个元素值为-1，而第四个元素值为 0。其余元素值的含义以此类推。

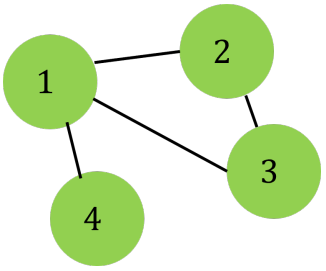


图 2-4 式2-10拉普拉斯矩阵对应的图

除了上述形式的拉普拉斯矩阵之外，还有归一化的拉普拉斯矩阵形式，如随机游走归一化拉普拉斯矩阵：

$$
L^{\text{RandomWalk}} = D^{-1}L = I - D^{-1}A, \tag{2-11}
$$

$L^{\text{RandomWalk}}$ 中的元素为：

$$L_{i,j}^{\text{RandomWalk}} = \begin{cases} 1 & , v_i = v_j \\ -\frac{1}{\text{diag}(v_i)} & , v_i \neq v_j \text{ and } (v_i, v_j) \in E \\ 0 & , \text{otherwise} \end{cases} . \tag{2-12}$$

对称归一化的拉普拉斯矩阵：

$$L^{\text{symmetric}} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}} = I - D^{-\frac{1}{2}} A D^{-\frac{1}{2}}, \tag{2-13}$$

$L^{\text{symmetric}}$ 中的元素为：

$$L_{i,j}^{\text{symmetric}} = \begin{cases} 1 & , v_i = v_j \; and \; \text{diag}(v_i) \neq 0 \\ -\frac{1}{\sqrt{\text{diag}(v_i)\text{diag}(v_j)}} & , v_i \neq v_j \text{ and } (v_i, v_j) \in E \\ 0 & , \text{otherwise} \end{cases} . \tag{2-14}$$

无向图的拉普拉斯矩阵是对称矩阵，可以进行特征值分解，其特征向量之间相互正交；由于拉普拉斯矩阵是半正定矩阵，因此其最小的特征值大于等于 0。对 L 进行特征值分解，可以将 L 写成如下形式：

$$L = U \Lambda U^{-1} = U \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix} U^{-1}, \tag{2-15}$$

其中 U 为正交的特征向量组成的矩阵，Λ 为特征值组成的对角矩阵。由于 U 为正交矩阵，因此又可以写成 $L = U \Lambda U^{\text{T}}$。

拉普拉斯矩阵特征分解后得到的特征向量就是图上的傅里叶变换的基，而特征向量对应的特征值就相当于传统傅里叶变换中的频率，因此图中顶点的嵌入向量也就可以用这一组基进行表示，通过嵌入向量与某个基的点乘即

可以求得该向量在该基上的分量：

$$\hat{f}(\lambda_l) = f \cdot u^l, \tag{2-16}$$

其中 f 为图中顶点的嵌入向量，u^l 为第 l 个特征向量（基），而 $\hat{f}(\lambda_l)$ 则为 f 在特征值 λ_l 代表的基上的分量。将其推广为矩阵形式：

$$\begin{bmatrix} \hat{f}(\lambda_1) \\ \hat{f}(\lambda_2) \\ \vdots \\ \hat{f}(\lambda_N) \end{bmatrix} = \begin{bmatrix} u_1^1 & u_2^1 & \cdots & u_N^1 \\ u_1^2 & u_2^2 & \cdots & u_N^2 \\ \vdots & \vdots & \ddots & \vdots \\ u_1^N & u_2^N & \cdots & u_N^N \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_N \end{bmatrix}, \tag{2-17}$$

即图傅里叶变换的矩阵形式为：

$$\hat{f} = U^{\mathrm{T}} f. \tag{2-18}$$

传统傅里叶逆变换相当于将对频率求积分，而在图中则变为了对特征值 λ_l 求和：

$$f_i = \sum_{l=1}^{N} \hat{f}(\lambda_l) u_i^l, \tag{2-19}$$

推广成矩阵形式为：

$$\begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_N \end{bmatrix} = \begin{bmatrix} u_1^1 & u_1^2 & \cdots & u_1^N \\ u_2^1 & u_2^2 & \cdots & u_2^N \\ \vdots & \vdots & \ddots & \vdots \\ u_N^1 & u_N^2 & \cdots & u_N^N \end{bmatrix} \begin{bmatrix} \hat{f}(\lambda_1) \\ \hat{f}(\lambda_2) \\ \vdots \\ \hat{f}(\lambda_N) \end{bmatrix}, \tag{2-20}$$

即图傅里叶逆变换的矩阵形式为：

$$f = U\hat{f}. \tag{2-21}$$

由卷积定理得知，两函数的傅里叶变换的乘积等于它们卷积后的傅里叶变换。因此，图中顶点的嵌入向量 f 和卷积核 g 的卷积，则等同于对 f 和 g

分别进行图傅里叶变换后的乘积再进行图傅里叶逆变换，即：

$$(f * g)_G = U((U^{\mathrm{T}}g) \cdot (U^{\mathrm{T}}f)). \tag{2-22}$$

在神经网络中，可以把 $U^{\mathrm{T}}g$ 整体当作可以训练的卷积核 g_θ，θ 为卷积核的参数，则图上傅里叶变换最终可以写为：

$$(f * g)_G = U g_\theta U^{\mathrm{T}} f. \tag{2-23}$$

2.2.2 图卷积神经网络的发展

Bruna 等人[42]的工作第一次提出了使用上述图卷积的神经网络。其中图卷积层的定义为：

$$H_j^{k+1} = \sigma(\sum_{i=1}^{c_{k-1}} U\Theta_{i,j}^k U^{\mathrm{T}} H_i^k),\ j \in \{1 \cdots c_k\}, \tag{2-24}$$

其中 $\sigma(\cdot)$ 为激活函数；$H^k \in \mathcal{R}^{N \times c_{k-1}}$ 为第 k 层的输出特征（也即第 $k+1$ 层的输入特征），图顶点数为 N，特征通道数为 c_{k-1}；$H^{k+1} \in \mathcal{R}^{N \times c_k}$ 为第 $k+1$ 层的输出特征，特征通道数为 c_k；$g_\theta = \Theta_{i,j}^k$ 为可训练的卷积核。

在该工作中，g_θ 是一个对角矩阵，即 $g_\theta = \mathrm{diag}(U^{\mathrm{T}}g)$ 用于减小参数量。但是图中顶点数量较多时，参数量 N 依然较大。为了解决这个问题，Defferrard 等人[43]认为 g_θ 是 Λ 的函数，提出用切比雪夫多项式来拟合 g_θ：

$$g_\theta(\Lambda) \approx \sum_{k=0}^{K-1} \theta_k T_k(\tilde{\Lambda}). \tag{2-25}$$

其中 K 为切比雪夫多项式的阶数；T_k 为递归定义的切比雪夫多项式：

$$T_0(x) = 1, \tag{2-26}$$

$$T_1(x) = x, \tag{2-27}$$

$$T_k(x) = 2xT_{k-1}(x) - T_{k-2}(x). \tag{2-28}$$

Λ 被缩放到 $[-1,1]$ 得到 $\tilde{\Lambda}$ 以满足 K 阶切比雪夫多项式展开的条件：

$$\tilde{\Lambda} = \frac{2\Lambda}{\lambda_{\max}} - I_N. \tag{2-29}$$

$\theta \in \mathcal{R}^K$ 为待训练的参数，可以看到参数量从之前的 N 降低为了 K。

Kipf 等人[44]的工作将参数量进一步降低。Kipf 首先将上述的切比雪夫多项式只展开到 1 阶，并假设 $\lambda_{\max} \approx 2$，得到

$$g_\theta * x \approx \theta_0 x + \theta_1 (L - I_N)x = \theta_0 x - \theta_1 D^{-\frac{1}{2}} A D^{-\frac{1}{2}}, \tag{2-30}$$

为了减少参数量及解决过拟合的问题，文章又假设 $\theta = \theta_0 = -\theta_1$，只保留了一个参数，得到

$$g_\theta * x \approx \theta (I_N + D^{-\frac{1}{2}} A D^{-\frac{1}{2}})x. \tag{2-31}$$

此时 $I_N + D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$ 的特征值在 $[0,2]$ 之间，由于图神经网络是若干图卷积层的堆叠，这可能会导致梯度爆炸或者梯度消失而使网络不能收敛。因此该工作又引入了重归一化技巧：

$$I_N + D^{-\frac{1}{2}} A D^{-\frac{1}{2}} \rightarrow \tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}}, \tag{2-32}$$

其中 $\tilde{A} = A + I_N$，$\tilde{D}$ 为 $\tilde{A}$ 的度矩阵。最终就可以得到该工作中图卷积层的公式：

$$H^{k+1} = \sigma(\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} H^k W^k), \tag{2-33}$$

其中 W^k 为待学习的图卷积核矩阵。至此，该公式成为了之后最广泛使用的图卷积形式。

2.3 动作质量评估

动作质量评估方法大致可以分为四种范式，即基于相似度衡量的方法、基于分类的方法、基于回归的方法和基于对比回归的方法，下文将形式化这四种范式，使得任务定义更为清晰：

- **基于相似度衡量的方法**：如图2-5所示，对动作片段打分的任务可以转换为衡量待评估的动作与标准动作之间的相似度，即

$$\text{score} = \text{sim}(\text{action}, \text{std}), \tag{2-34}$$

其中 action 为待评估动作样本，std 为该动作对应的标准动作样本，sim 在传统方法中可以为样本特征间的余弦相似度、L_1 或 L_2 距离以及 DTW 距离[45]等，也可以为深度度量学习方法学习到的距离函数。

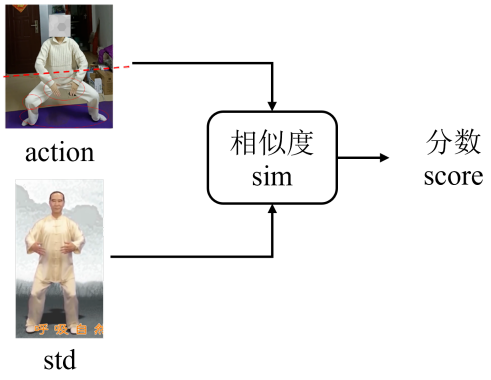


图 2-5 基于相似度衡量的方法示意图

- **基于分类的方法**：如图2-6所示，在动作片段的分数标签为离散值时，可以将所有的动作分数作为类别，直接将预测动作质量分数的任务转化为多分类任务，即

$$\text{score} = \arg\max_i M_\theta(\text{action})_i, \tag{2-35}$$

其中 M_θ 为分类网络，θ 为网络参数，网络的输出是一个向量，每个元素表示对应分数类别的概率。

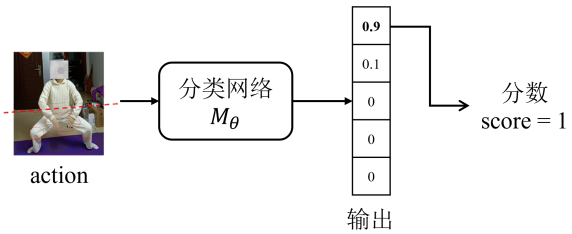


图 2-6 基于分类的方法示意图

- **基于回归的方法**：如图2-7所示，对样本打分天然就是回归任务的一种，对于样本 action 及其标签 score，期望学习到一个好的网络 M_θ，通过最

小化

$$\mathcal{L}(M_\theta(\text{action}), \text{score}) \tag{2-36}$$

使得网络的输出尽量接近真实分数，实现动作质量的评估。同样地，损失函数也可以为各种衡量距离的函数。

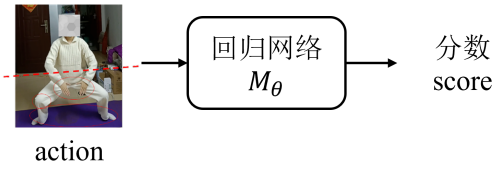


图 2-7 基于回归的方法示意图

- **基于对比回归的方法**：如图2-8所示，一些研究认为直接对单个样本 action_1 回归得到的分数缺乏可解释性，因此又引入了另一个对比样本 action_2。此时网络不再学习样本 action_1 的真实分数，而是学习样本 action_1 相对于 action_2 分数的差值，即最小化

$$\mathcal{L}(M_\theta(\text{action}_1, \text{action}_2), \text{score}_1 - \text{score}_2). \tag{2-37}$$

此时，样本 action_1 的预测质量分数为 $\text{score}_2 + M_\theta(\text{action}_1, \text{action}_2)$。

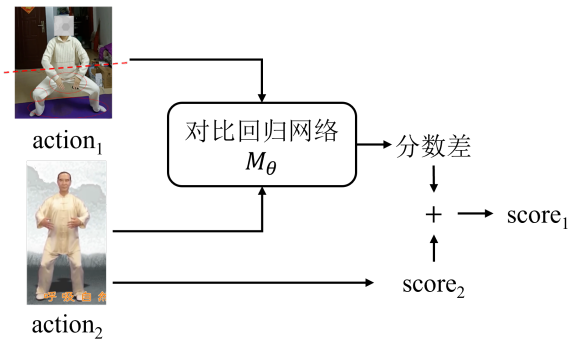


图 2-8 基于对比回归的方法示意图

2.4 本章小结

本章首先介绍了对比学习的基本原理，即学习一个从样本到嵌入特征的映射网络，使得相似的样本在嵌入空间中的表示较近，而不相似的样本在嵌

入空间中的表示较远。InfoNCE 损失同时考虑一对正样本和多对负样本，奠定了后续对比学习的基础，而 SimCLR 则是第一个运用 InfoNCE 的效果较好的自监督对比学习方法。此外，本章还介绍了 triplet 损失，方便在后文中与其它方法的特性进行比较。

本章随后介绍了图卷积神经网络与卷积神经网络的区别以及图卷积的推导，由于本文建立的八段锦动作质量评估数据集样本为人体骨骼的时间序列，而图卷积神经网络非常适合处理这种非欧几里得数据，因此在本文八段锦数据集上训练的也都为图卷积神经网络模型。

本章最后将动作质量评估方法分为四种范式，分别形式化予以介绍。在本文中，由于研究重点为对比学习损失函数的改进，因此同时选用了基于分类与基于回归的方法进行实验和对比。

第三章　八段锦动作质量评估数据集的构建

高质量的数据集是后续工作的基础。本章将详细介绍八段锦动作质量评估数据集的构建过程，包括样本的搜集、标注和处理等，本章还开发了一个数据集辅助标注系统以便于样本的快速标注。本章最后对构建好的数据集的统计特征进行了简单介绍。

3.1　数据搜集

Bilibili 作为中国著名的视频网站之一，成为了许多人上传视频分享自己生活的平台，而其中当然也包括八段锦相关视频。通过在 Bilibili 搜索“八段锦”可以得到各种来源的八段锦视频，其中包括国家体育总局版的八段锦标准教程、八段锦爱好者的练习视频、八段锦比赛视频以及大学八段锦课程的考核作业等。不同来源的视频也贡献了不同水平的八段锦动作，为后续有区分度的标注打下了基础。

并不是所有的八段锦视频都可以放入数据集之中，如图3-1所示，该视频中出现了视角的转换，这不仅会使得姿态估计算法提取出的骨骼的视角不同，也会导致画面中人物的顺序发生改变（如黑衣服的人所处位置从左 3 变为左 2），不利于后续的数据集处理及模型训练；如图3-2所示，在同一帧视频中出现的人数太多且相互遮挡，也会降低姿态估计算法的准确度，造成不同人的关节错分，同时也会导致难以追踪和提取特定个体的骨骼。因此，需要将出现视角转换以及同帧人数太多的视频筛除，选择一镜到底的、人数较少且关键个体无遮挡的视频。

(a) 视角转换前 (b) 视角转换后

图 3-1 视频前后帧出现视角转换

图 3-2 视频中出现太多个体且相互遮挡

3.2 标注

3.2.1 各式动作片段起止时间标注

虽然收集到的很多视频的背景音乐都是国家体育总局版本的口令，但并不是所有人的动作都严格遵守了口令的时间点，经常会有或多或少的提前或延迟；同时，还有一些视频并没有背景音乐，也没有按照口令的时间点进行动作。这就要求对八段锦中每段动作的起止时间进行标注。由于八段锦视频较长，只靠人手动拖动进度条观察动作起止时间费时费力，因此本章设计了算法1以辅助标注者找到每段动作的转折点的大致范围。

通过观察得知，八段锦的每段动作都是由若干个基本动作的组合重复数次构成的，例如“双手托天理三焦”可以粗略地看作“上托”和“下落”动作重复若干次。八段锦的动作是舒缓的，筛选过的数据集中视频一镜到底，背景是几乎不变的，而每一式的动作组合都不相同，也就是视频中人物部分像素点的变化规律不同。

算法 1 动作转折点大致范围寻找算法

输入: 八段锦视频 V，采样率 r
输出: 动作片段起始时间 $T_0, \cdots, T_{K-1}$
$F \leftarrow$sample(Gray(V),r)
for $i = 1$ to length(V) - 1 **do**
$\text{Diff}_i \leftarrow F_i - F_{i-1}$
$N_i \leftarrow \|\text{Diff}_i\|_2$
end for
对 N 进行移动平均滤波
DBSCAN 和遗传算法寻找动作片段起始时间 $T_0, \cdots, T_{K-1}$
return $T_0, \cdots, T_{K-1}$

令

$$F = \text{sample}(\text{Gray}(V), r), \tag{3-1}$$

其中 V 是一个样本视频，r 为采样率，则 F 即是从灰度化的原始视频中以 r 为采样率采样得到的若干帧图像。

令

$$\text{Diff}_i = F_i - F_{i-1}, \tag{3-2}$$

即对采样得到的 F 相邻帧之间作差，背景及人物中前后相同的部分会被减掉，剩下变化的部分（如图3-3所示）。由于不同动作的移动模式不同，改变像素点的位置、数量和颜色也不相同；上述提到八段锦每段动作都会重复若干次，因此这种变化也会呈现一定的周期性。

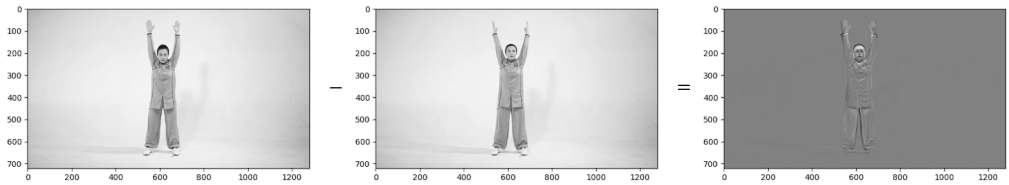
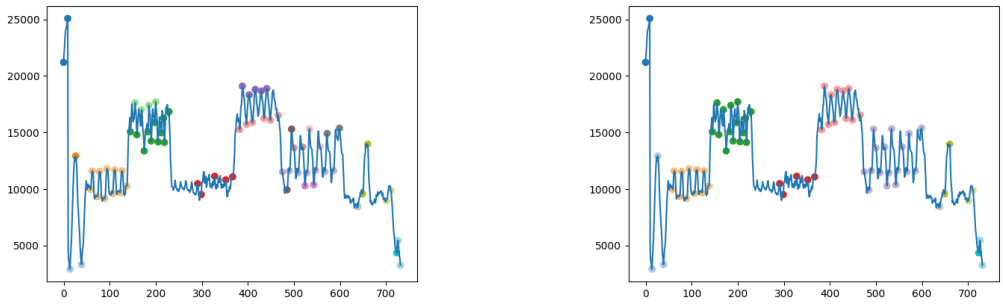


图 3-3 相邻帧作差结果图

可以通过简单地计算 Diff 的模来发现这种周期性，令

$$N_i = \|\text{Diff}_i\|_2, \tag{3-3}$$

之后通过移动平均滤波使其变得更平滑。寻找动作片段起始时间 $T_0, \cdots, T_{K-1}$



(a) 区间合并前聚类结果　　(b) 区间合并后聚类结果

图 3-4 波峰波谷点聚类效果

可以看作是将曲线中的点有序聚类的任务，解决这个问题可以用 Fisher 最优分割法[46]，但基于动态规划的该方法速度较慢。本文则采用两种启发式的算法近似求解。一种为找出曲线 N 的波峰波谷，通过 DBSCAN 算法[47]将其聚类，相比于 K-Means 算法[48]，DBSCAN 不需要预先设定聚类的数量。如图3-4(a)所示，图中绘制了平滑后的曲线 N 及聚类后的波峰波谷点，从中可以看出直接使用 DBSCAN 的聚类结果存在一定的问题，同一段动作波形的波峰波谷点纵坐标距离较远时，聚类算法会将波峰点和波谷点分别聚为一类（如图中浅绿色和绿色的点），因此还需要将区间重合的类进行合并，合并后的聚类结果如图3-4(b)所示。此后将每个类所有点的最大最小横坐标作为动作片段的近似转折点集合 T^{DBSCAN}。

另一种方法则是利用遗传算法[49-50]求解 Fisher 最优分割法的目标函数，即：

$$\arg\min_{\{T'_i\}} \sum_{i} \sum_{j=T'_i}^{T'_{i+1}-1} (N_j - \frac{1}{T'_{i+1}-T'_i} \sum_{k=T'_i}^{T'_{i+1}-1} N_k)^2, \tag{3-4}$$

得到 T^{genetic}。将 T^{DBSCAN} 和 T^{genetic} 合并起来得到最终的动作片段起始时间 $T_0, \cdots, T_{K-1}$。

如图3-5所示，通过可视化模随时间的变化，可以明显发现曲线振荡规律会随动作的改变而改变，图中用红点标注了 DBSCAN 聚类得到的动作转折点，绿色标注了遗传算法得到的动作转折点，从而大大减少了人标注时拖动进度条的时间。

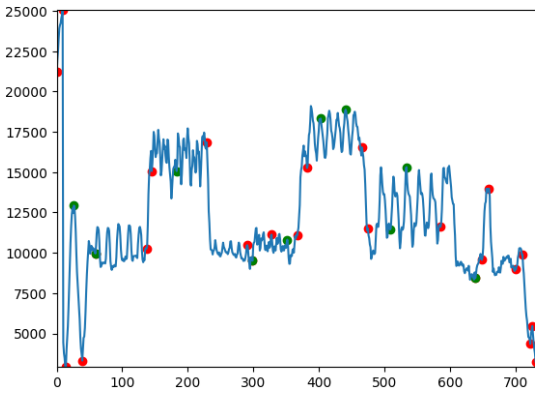


图 3-5 帧差模随时间的变化图

3.2.2 各式动作片段动作质量标注

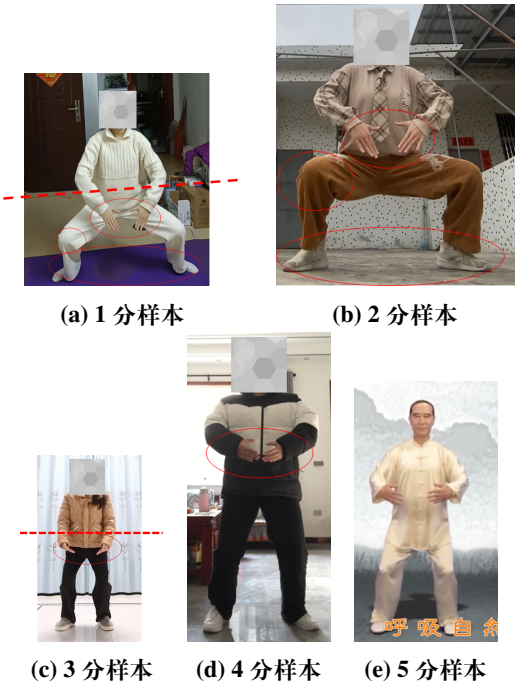
在分割得到各个视频的不同动作片段之后，还需要对这些动作进行质量标注。八段锦不像体操、跳水等比赛动作有明确的动作得分规范，想要详尽地将其得分划分在 0 至 100 之间是非常困难的，因此本文选择将八段锦的质量划分为 5 个等级，1 至 5 分，1 分为最差，5 分为最好。

通过参阅八段锦的动作要领，如表3-1所示，本文为八段锦的前八段动作列举了若干要点或易错点（分左右的动作只列举一个方向，另一方向同理，表中只展示前三段标准）。“背后七颠百病消”和“收式”由于动作较简单，本文不再对这两段动作进行质量评估。评分采用扣分制，每有一个要点没达到（或犯了一个错误），则扣除 1 分，扣除的分数不超过 4 分。这些标准较为客观，标注人员可以比较明确地判断样本中的动作是否达到要求，很大程度上减少了标注的主观性，提高了标注的质量。

以“预备式”为例，图3-6展示了 1 至 5 分不同质量的动作示例。子图3-6(a)中人物手未放在腹前且未指尖相对，两脚分开距离太大，膝关节屈度太大，因此扣除 4 分；子图3-6(b)中人物手放在了腹前，但其他错误与3-6(a)相同，因此扣除 3 分；子图3-6(c)中人物手未放在腹前且未指尖相对，因此扣除 2 分；子图3-6(d)中人物只有两手间距太小这一个问题，因此扣除 1 分；子图3-6(e)为国家体育总局的演示版本，标准无误，因此为 5 分。

表 3-1 八段锦动作质量标注标准表（部分）

动作	要点（易错点）
预备式	1. 两脚与肩同宽，分开距离不能太大 2. 大拇指放平，手掌朝内，指尖相对，两手间距约 10cm 3. 膝关节稍屈，不超过脚尖 4. 手掌放到腹前，不能太低或太高
双手托天理三焦	1. 两掌五指在腹前交叉，掌心向上 2. 两腿伸直，两掌上托于胸前，随后两臂内旋向上托起，掌心向上，抬头目视两掌 3. 两掌继续上托，肘关节伸直，下颌内收，目视前方 4. 膝关节微曲，两臂从两侧下落，两掌捧于腹前 5. 两脚分开距离不能太大
左右开弓似射雕	1. 重心移动开步时曲腿，站立后伸直，两掌交叉于胸前 2. 一掌曲指，另一掌八字拳，与肩同高，看向一侧 3. 随后屈膝



(a) 1 分样本 (b) 2 分样本 (c) 3 分样本 (d) 4 分样本 (e) 5 分样本

图 3-6 “预备式”不同动作质量样本示例

3.2.3 数据集标注系统

为了方便数据集的标注，本文开发了一个数据集标注系统，可以完成八段锦动作的分段起止时间标注及质量标注两项任务，同时可以应用本章的算法1以加快标注速度。数据集标注系统分为前端和后端两部分，其中前端使用 HTML、JS 和 CSS 编写，后端使用 Python 的 Flask 框架。

起止时间标注

图3-7展示了标注系统的起止时间标注界面。左侧为视频文件列表，通过点击“加载”按钮即可读取所有八段锦视频文件并显示在列表中。双击文件列表中的列表项即可通过中间的播放器播放该视频，同时在右侧的时间段列表中显示过往的标注结果（如果有）。点击“计算波形”，后端会通过算法1计算当前视频的帧差模随时间变化的波形图，以辅助标注。视频播放器的下方进度条共有两个滑块，分别用来标注动作片段的起始时间和终止时间。两个滑块均可在全部进度条上移动，在标注完一段动作后只需要把前方滑块拖动到后方滑块的后面，即可记录从上一个动作的终止时间（即下一个动作的开始时间）到下一个动作的终止时间。滑块拖动时，进度条右侧的按钮的内容也会随之改变，显示内容为“起始时间（秒），终止时间（秒）”，点击该按钮，该时间段会被添加到右侧的时间段列表中，并按照起止时间排好序。在标注完一个视频后，可以点击“保存”按钮将标注结果保存至服务器。如果标注结果有错，也可以在时间段列表中选中错误的标注时间段后点击“删除”按钮将其删除。最后，可以点击“下一个”按钮加载下一个视频，重复上述标注过程。

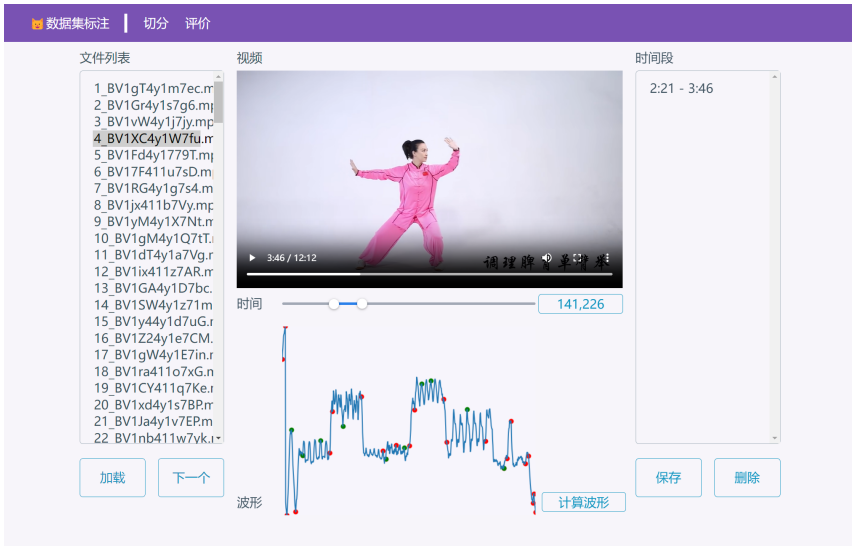


图 3-7 动作片段起止时间标注界面

动作质量标注

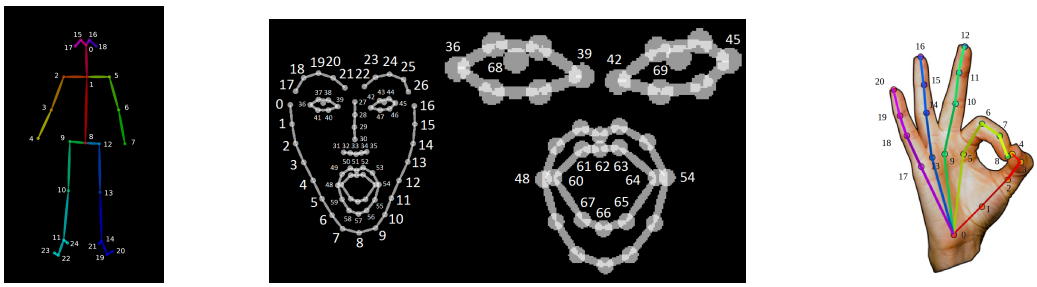
在完成动作片段起止时间标注后，通过顶部导航栏“评价”切换到如图3-8所示的动作质量标注界面。点击“加载”按钮，左侧文件列表将显示读取到的不同类别动作文件夹下的片段视频，每个类别文件夹下均有一个标准动作视频（std.mp4）。双击左侧文件列表中的文件名，右侧的播放器将同时加载标准动作视频和当前选中的待标注视频以便于对比。选择下方五角星评分并点击保存后，即可完成对当前动作片段的质量标注。点击“下一个”按钮加载下一个视频，重复上述标注过程。

图 3-8 动作质量标注界面

3.3 姿态估计与数据集处理

为了从 RGB 视频中提取出人体关节骨骼信息，需要使用姿态估计算法，本文选用了 OpenPose[51-54]。OpenPose 是卡内基梅隆大学提出的第一个实时多人身体、手部、面部和脚部关键点检测系统。该系统采用了自底向上的姿态估计算法，算法运行时间不受画面中人数的影响。如图3-9所示，OpenPose 可以识别出身体的 25 个关键点、面部的 70 个关键点和手部的 21 个关键点（单只手），共计 137 个关键点。图3-10展示了八段锦动作质量评估数据集中若干样本的姿态估计结果，每行代表一种动作，三行分别为预备式、双手托天理三焦和左右开弓似射雕，而每一行从左到右分别是质量分数为 1 至 5 分的样本。

姿态估计后，存在一帧画面包含多人关键点的视频，这一方面是因为视频中原本就有多人，另一方面是姿态估计算法出现误差导致单人视频误识别出多人的关键点。对于多人视频，需要挑选出特定一人的关键点放入数据集



(a) 身体关键点 (b) 面部关键点 (c) 手部关键点

图 3-9 OpenPose 识别关键点及编号

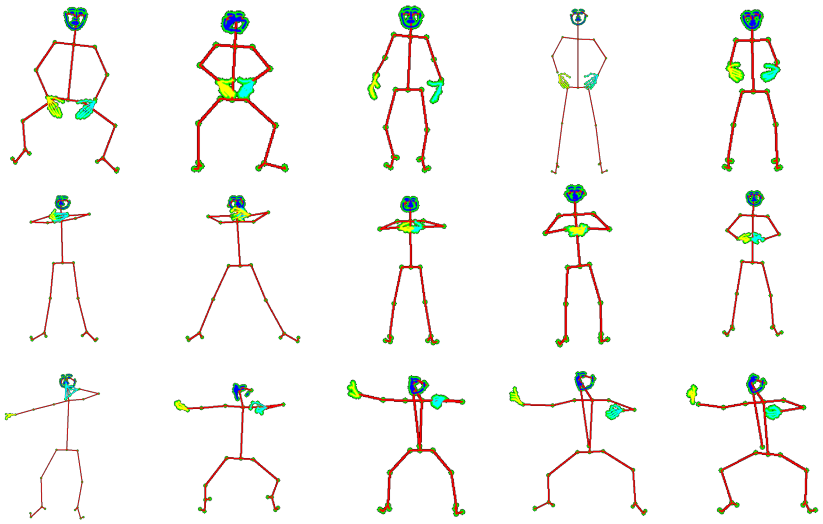


图 3-10 八段锦动作质量评估数据集中样本姿态估计结果

中；对于误识别的视频，需要将误识别的部分去除。姿态估计算法每帧输出的个体的顺序是不固定的，不能简单地通过固定 ID 选出特定的人，因此本章设计了一种基于重心的多人视频挑选特定人关键点算法，其伪代码如算法2所示。

一个视频姿态估计后的关键点时间序列 J 包含了若干帧，而每帧内又有可能有多人。算法需要从一个视频的中间部分选择一帧作为参考帧，其帧编号为 f。在该帧内，需要提取的特定人应该清晰可见，并找到其在画面从上到下、从左到右的顺序编号 i。八段锦动作整体是舒缓的，这意味着在相邻帧间同一个人的动作幅度不会很大，也就是重心的移动距离应该很小。在计算第 f 帧第 i 个人的重心 c 后，从第 f 帧依次向前、向后寻找相邻帧内与重心 c 距离最小的重心所代表的人的编号 j'，并将该编号人的关键点加入到提取出的特定人关键点时间序列 J' 中。同时，需要判断当前处理的帧中的人物数量与参考帧是否一致，如果一致则可以认为该帧是姿态估计结果较好的一

算法 2 多人视频挑选特定人关键点算法

输入:
姿态估计后关键点时间序列 J
参考帧编号 f
参考帧内待提取特定人编号 i
输出: 特定人关键点时间序列 J'
将 J_f 内的各人关键点按照其重心排序
$c \leftarrow \text{center}(J_{f,i})$
$J'_f \leftarrow J_{f,i}$
for $k = f-1$ to 0 **do**
  $j' \leftarrow \arg\min_j \|\text{center}(J_{k,j}) - c\|_2,\ j = 0, 1, \cdots, \text{length}(J_k)$
  $J'_k \leftarrow J_{k,j'}$
  if $\text{length}(J_k) = \text{length}(J_f)$ **then**
    $c \leftarrow \text{center}(J_{k,j'})$
  end if
end for
$c \leftarrow \text{center}(J_{f,i})$
for $k = f+1$ to $\text{length}(J)-1$ **do**
  $j' \leftarrow \arg\min_j \|\text{center}(J_{k,j}) - c\|_2,\ j = 0, 1, \cdots, \text{length}(J_k)$
  $J'_k \leftarrow J_{k,j'}$
  if $\text{length}(J_k) = \text{length}(J_f)$ **then**
    $c \leftarrow \text{center}(J_{k,j'})$
  end if
end for
return J'

帧，可以用该特定人新的重心位置替代之前记录下的参考重心位置。如果不更新参考重心位置，时间跨度较大时特定人可能移动了较大距离，此时与参考重心最近的重心代表的人已经不再是需要提取的特定人。

此外，数据集中的不同视频的分辨率和人物在画面中所占比例各不相同，因此需要对其进行归一化处理。令 $(c_x, c_y) = \text{center}(J'_k)$，即 J'_k 帧内人物的重心，width 和 height 分别是原视频的像素宽度和高度，则归一化后的坐标即为

$$x = \frac{J'_{k,x} - c_x}{\text{width}}, \tag{3-5}$$

$$y = \frac{J'_{k,y} - c_y}{\text{height}}. \tag{3-6}$$

便于模型的训练和收敛。

3.4 数据集统计特征

经过数据搜集筛选后，本文建立的八段锦质量评估数据集共有 178 个视频，总时长约 36 小时，经过姿态估计后总帧数为 3746750 帧。数据集共为八种动作标注了 1 至 5 分的动作质量，表3-2展示了不同动作下动作质量的分布（其中有两个视频缺少“预备式”动作，因此“预备式”动作样本数为 176），总体来看评分为 2 至 5 的样本较多，评分为 1 的样本较少。数据集将每个质量的动作按照 3:1 的比例划分为了训练集和测试集。

表 3-2 不同动作的动作质量分布表

样本数量 / 动作类型 / 动作质量	1	2	3	4	5	6	7	8	总计
5	86	95	62	41	65	49	65	76	539
4	31	19	32	47	29	31	29	40	258
3	9	21	29	30	29	29	41	38	226
2	10	39	41	53	40	34	36	17	270
1	40	4	14	7	15	35	7	7	129
总计	176	178	178	178	178	178	178	178	1422

3.5 本章小结

在这一章节，本文详细阐述了八段锦动作质量评估数据集从搜集筛选，到动作片段起止时间标注及动作质量评估，再到姿态估计提取关节骨骼信息后处理的全过程，并对数据集的统计特征进行了概括。本文制定的八段锦动作质量评估标准较为客观，减少了标注过程中的主观性，标注质量较高。据本文了解，该数据集是第一个同时标注动作片段起止时间和五级动作质量的八段锦动作质量评估数据集，并拥有最多的练习者数量和最长的总时长。同时，该章节还开发了一个动作质量评估数据集的辅助标注系统，该系统不仅可以用于本文数据集的标注，还可适用于其他类似需求的数据集。

第四章　基于利用粗粒度标签的有监督对比学习的八段锦动作质量评估

目前大部分动作质量评估工作聚焦于对模型结构的改进，但越来越复杂的模型也带来了越来越大的计算开销。因此，本章将从改进损失函数的角度出发，探索在几乎不增加计算量的前提下提高模型性能的方法。具体而言，本章提出了利用粗粒度标签的有监督对比学习算法，同时适用于分类和回归形式的八段锦动作质量评估任务。除了在八段锦动作质量评估数据集上的实验外，本章还在一系列分类和回归任务的公开数据集上进行了实验以验证其有效性。

4.1　研究动机

对比学习的思想是使得相似的样本在嵌入空间中的表示较近，而不相似的样本在嵌入空间中的表示较远。而八段锦动作质量评估则可以借鉴这种思想，使质量相近的样本尽量靠近，使质量不同的样本尽量远离，从而达到区分不同质量动作的目的。

如图4-1所示，在自监督对比学习中，对某个锚样本（超球面上的绿色点）来说其正样本（超球面上的橙色点）是它经过不同数据增强后的自身。在优化自监督对比学习损失函数的过程中，图中粉色衣服人本身的与其数据增强后的嵌入特征在超球面上的距离被拉近，与除此之外的负样本（超球面上的红色点）的特征距离被拉远（即使是其他同一式八段锦动作样本）。这是因为自监督对比学习没有用到样本的标签信息，也就没有将同一个类的不同样本之间的关系考虑在内。但是从直觉上来看，同类别样本的特征在超球面上的分布也应该较为接近。

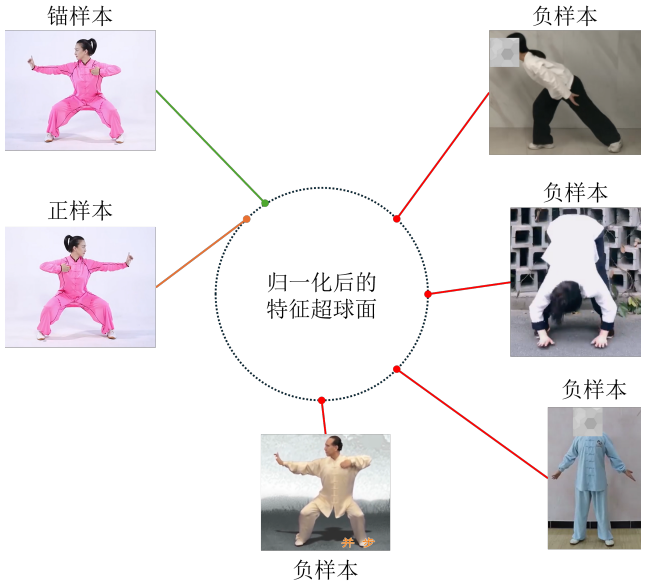


图 4-1 自监督对比学习特征分布

Khosla 等人[55]提出的有监督对比学习则对此做了改进，由于用到了样本的标签信息，因此也从自监督学习变为了有监督学习。相比于自监督对比学习只使用一个正样本，有监督对比学习对于每个锚样本同时考虑了多个正样本和多个负样本。自监督对比学习的正样本是锚样本数据增强后的自身，而有监督对比学习的正样本是与锚样本同类别的其他样本。有监督对比学习还可以看作是 triplet 损失和 InfoNCE 损失的泛化，前者对每个锚样本利用一个正样本和一个负样本，后者则是利用一个正样本和若干负样本。

与自监督对比学习的损失函数相比，有监督对比学习的损失函数为

$$\mathcal{L}^{\text{SupCon}} = \sum_{i \in I} \mathcal{L}_i^{\text{SupCon}} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp\left(z_i \cdot z_p / \tau\right)}{\sum_{a \in A(i)} \exp\left(z_i \cdot z_a / \tau\right)}, \quad (4\text{-}1)$$

其中 $I = \{1, 2, \cdots, 2N\}$，即 N 个样本经过两次不同数据增强后形成的双视图 batch；$P(i)$ 为在一个 batch 中除了锚样本 i 之外与其相同类别的样本集合，即正样本；$A(i)$ 为除了锚样本 i 之外的其他样本。在随机生成的 batch 大小 $2N$ 大于类别数量 C 时，平均每个类别会有 $\frac{2N}{C}$ 个样本存在，也就是说每个锚样本正样本的个数平均为 $\frac{2N}{C} - 1$ 个，这大大地增加了正样本的个数。同时，在分母上依然保留了负样本的部分，这体现了有监督对比同时利用多个正样本和负样本进行学习的特性。

如图4-2所示，经过有监督对比学习后，同种动作类别的样本特征在超球

面上分布更为接近，这不但更符合人类的认知，也可以提高模型在下游任务上的表现。

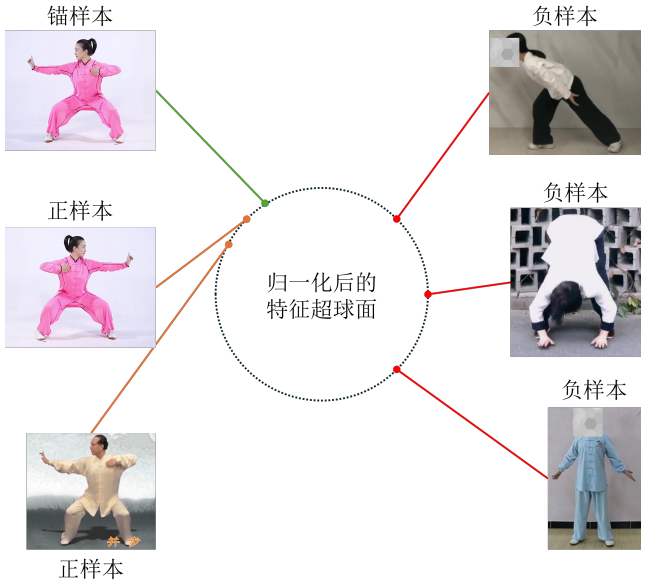


图 4-2 有监督对比学习特征分布

但是在一些数据集中，特别是细粒度分类数据集，样本的标签包含了非常具体的类别。如果只使用有监督对比学习方法，可能又会出现类似于自监督对比学习中的现象——即使属于同一个细粒度类别的样本在超球面上的分布距离较近，但属于同一粗粒度类别的不同细粒度类别样本之间可能距离较远，甚至随机地分布在超球面上。

举例来说，对于八段锦动作质量评估数据集而言，动作质量为细粒度标签而动作种类为粗粒度标签。使用有监督对比学习时，可能会发现预备式得分为 1 的样本之间距离比较近，预备式得分为 5 的样本之间距离也比较近，但是预备式得分为 1 和得分为 5 的样本之间可能距离较远。相对于其他式动作，属于同一式动作的不同质量样本理应距离较近，然而只利用动作质量信息并不能做到这一点。

从直觉上来说，可以参考类似有监督对比学习相对自监督对比学习的改进，同时利用样本粗粒度标签和细粒度标签的信息，拉近属于同一粗粒度类别的各个细粒度类别在超球面上的距离。

近年来已经出现了一些利用粗粒度标签信息的工作，例如 Lu 等人[56]的工作提出了一种粗粒度标签的表示方法，即将一种粗粒度标签表示为其包含

的所有细粒度标签的 one-hot 向量和，并设计了一种损失函数，在训练时同时考虑粗细粒度的损失；Touvron 等人[57]的工作则提出了同时使用自监督对比学习和有监督对比学习的多任务框架；而 Feng 等人[58]的工作提出了 MaskCon，对同属于一种粗粒度标签的样本赋予不同软细粒度标签，同样使用了自监督对比学习和有监督对比学习。后两个工作虽然利用了粗粒度标签信息，但其任务设置与本文不同，探索的是仅有粗粒度标签的情况下如何更好地进行细粒度图像检索任务，因此不能与本文算法进行对比。本文提出的利用粗粒度标签的方法相对简单且有效，仅使用有监督对比学习而不涉及自监督对比学习，可以同时应用于分类和回归任务，下文将会具体介绍并进行实验探究其效果。

4.2 算法设计

数据增强是对比学习的重要组成部分，本章对八段锦动作质量评估数据集样本使用两种数据增强方式，分别为镜像翻转和重采样，以达到扩充数据集、增强模型泛化能力的作用。此外，八段锦动作质量评估数据集中的样本为骨骼时间序列，非常适合采用图卷积神经网络进行处理，本章将采用三种不同的图卷积神经网络模型作为骨干网络进行实验。同时为了改进有监督对比学习的缺陷，本章提出了一种利用粗粒度标签的有监督对比学习算法，同时利用样本的细粒度标签和粗粒度标签信息。在动作质量标注分数是离散值的情况下，动作质量评估任务既可以看作预测分数的多分类问题，也可以看作对质量分数进行回归的问题。因此，本节将分别给出适用于分类任务和回归任务的利用粗粒度标签的有监督对比学习算法形式。

4.2.1 数据增强

数据增强是大部分对比学习方法中不可缺少的一步，图4-3展示了一些图像数据集常用的数据增强方法，其中一些后续将在其他图像数据集的实验中用到。而在八段锦动作质量评估任务中，本章对八段锦动作质量评估数据集样本使用两种数据增强方式，分别为镜像翻转和重采样，以使网络学到更好的样本特征表示。

(a) 原图 (b) 裁剪与缩放 (c) 裁剪、缩放与水平翻转

(d) 灰度化 (e) 颜色失真 (f) 旋转

(g) 遮挡 (h) 高斯模糊 (i) 高斯噪声

图 4-3 图像数据集数据增强方法示例

镜像翻转

由于拍摄相机设置不同，数据集中本来就存在镜像翻转后的视频。因此，通过镜像翻转数据增强后，模型对这部分视频的鲁棒性得到提高。图4-4展示了镜像翻转的效果。

(a) 镜像翻转前 (b) 镜像翻转后

图 4-4 镜像翻转示意图

重采样

完整的八段锦视频时长约 12 分钟，共约两万帧图像，每个动作约有两三千帧的持续时间，如果都输入网络进行处理，无疑会拖慢训练和推理的速度。

八段锦是整体缓慢流畅的，相邻帧之间的变化不会很大，因此可以对原始的帧进行采样。

图4-5展示了采样的方式。图中每个小圆点代表了一帧，向右延伸的横轴为时间轴。将所有帧每隔 d 帧划分为一段，再在每一小段中均匀采样一帧作为该小段动作的代表帧（即图中黑色小圆点），最后所有代表帧连接在一起作为最终的采样结果。这样做的好处是每次采样得到的帧序列都是不同的，变相地极大扩充了数据集，提高了模型的泛化能力；同时采样后的帧数大大减少，也提高了模型对长时间动作样本的处理能力。图4-6展示了真实数据集中的一个样本经过两次重采样后得到的不同帧序列，可以看出它们之间存在一些差别。

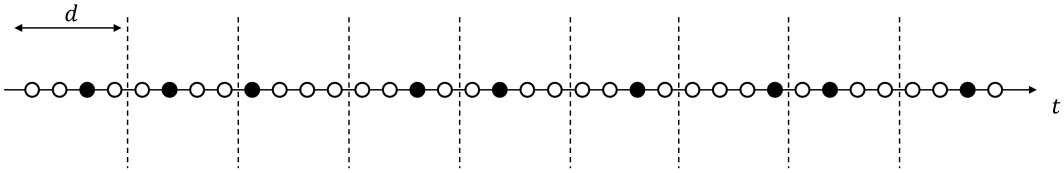
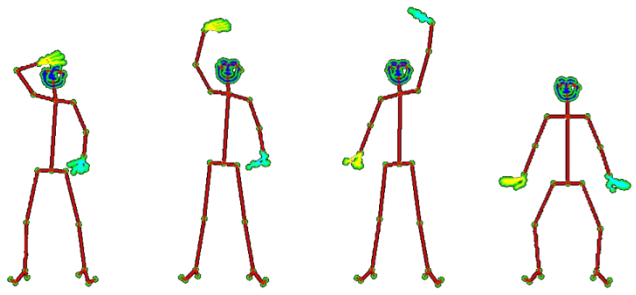
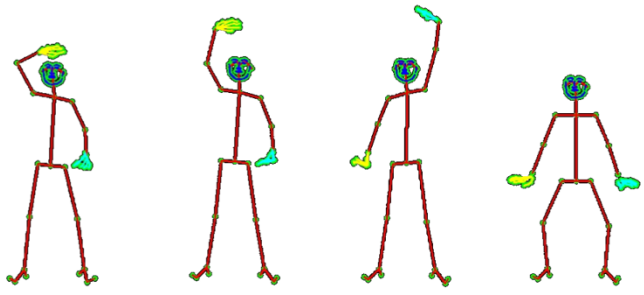


图 4-5 重采样过程示意图



(a) 采样帧序列 1



(b) 采样帧序列 2

图 4-6 数据集中同一样本两次不同采样帧序列

4.2.2 骨干网络

本文提出的数据集中的样本为人体骨骼的时间序列，非常适合用图卷积神经网络进行处理。为了验证本文提出的利用粗粒度标签的有监督对比学习方法的有效性，本章通过若干骨干网络提取特征后使用一层简单的全连接层作为分类器或者回归器，对动作质量分数进行预测。

ST-GCN

Yan 等人提出的 ST-GCN（Spatial Temporal Graph Convolutional Networks）[26]是本节选用的第一个骨干网络。该工作是将图卷积神经网络用于基于骨骼的动作识别任务的开山之作，其亦可以用在回归任务之中[27]。

基于骨骼的动作质量评估任务的输入为人体关节坐标的时间序列，即 $x \in \mathcal{R}^{T\times V\times C}$，其中 T 为时间序列的长度，V 为人体关节的数量，而 C 则是每个关节的特征长度，本文使用的 OpenPose 姿态估计的结果中 $C=3$，分别为关节的二维坐标及置信度。在一帧内，各关节按照人体自然的骨骼进行连接；在相邻帧之间，相同的关节在时间维度上进行连接。

ST-GCN 由若干层堆叠的图空间卷积和时间卷积层构成。其中图空间卷积层用于在同一帧内聚合特征，ST-GCN 对本文第二章介绍的图卷积公式做了改进：

$$H^{k+1}=\sigma(\sum_{j=0}^{2}D_j^{-\frac{1}{2}}(A_j\odot M)D_j^{-\frac{1}{2}}H^kW_j^k), \tag{4-2}$$

其中 $A+I=\sum_{j=0}^{2}A_j$，D_j 为 A_j 的度矩阵，该工作将图本身的连接加上自环后形成的图分成了 3 部分，3 个子图分别代表了图中每个关节到自身的连接、每个关节到远离自身重心的关节的连接和每个关节到靠近自身重心的关节的连接，在每个子图上进行图卷积后再对特征求和，充分利用了身体骨骼的空间局部特征。考虑到并不是身体中的每个关节连接都对识别动作或评估动作质量有帮助，该工作又引入了可学习的掩码 M 与原邻接矩阵 A_j 对应元素相乘，M 中某位置的值越大，代表了这条边在识别或评估中的作用越大。

而时间卷积的运算较为简单，由于在时间维度上相同关节在相邻帧上有连接，因此可以在时间维度直接使用传统卷积，聚合若干帧之间的特征。

AGCN

ST-GCN 仅仅利用了人体骨骼之间自然连接的信息，即使加入了可学习的掩码 M，由于使用的是对应元素乘法，原本邻接矩阵中为 0 的位置永远不会改变，这意味着不能产生新的连接。为了解决这个问题，Shi 等人提出了 AGCN（Adaptive Graph Convolutional Networks）[59]，对图卷积层做了进一步的改进：

$$H^{k+1} = \sigma(\sum_{j=0}^{2} D_j^{-\frac{1}{2}}(A_j + B_j + C_j)D_j^{-\frac{1}{2}}H^k W_j^k), \tag{4-3}$$

其中的不同在于卷积使用的邻接矩阵由三部分组成：A_j、B_j 和 C_j。

A_j 依然代表了人体骨骼的自然连接。

B_j 则是一个没有限制的可学习的矩阵，也就是说，B_j 中的元素完全从数据之中学来，通过数据驱动的方式，B_j 可以学习到人体骨骼自然连接之外的有利于下游任务的连接。同时由于 B_j 中元素大小不受限制，因此也可以表征连接的强度。B_j 是直接加到 A_j 之上的，与 ST-GCN 中的 M 相比可以产生新的连接，更加灵活。

C_j 同样为数据驱动的矩阵，但与 B_j 不同，每个样本的 C_j 矩阵均不相同。具体来说，AGCN 使用自注意力计算了特定样本中人体两两关节之间是否应该存在连接，以及连接的强度：

$$C_j = \text{softmax}(H^{\mathrm{T}} W_\theta^{\mathrm{T}} W_\phi H). \tag{4-4}$$

其中 W_θ 和 W_ϕ 代表了两个不同的卷积投影层，对原输入特征分别进行两次卷积后得到投影后的特征 $W_\phi H$ 和 $W_\theta H$，将 $W_\theta H$ 转置后与 $W_\phi H$ 相乘可以得到一个 $N \times N$ 的矩阵，通过 softmax 将其归一化到 $[0, 1]$ 之间，此时矩阵中的权重即可代表人体关节间的连接强度。

A_j、B_j 和 C_j 分别从不同的方面刻画了人体关节之间的连接关系，在下游任务上取得了较好的效果。

AAGCN

虽然 AGCN 可以灵活地建立人体关节之间的空间连接，但对时间、特征等维度并没有特殊处理。此后，Shi 等人又提出了 AAGCN（Attention-enhanced Adaptive Graph Convolutional Networks）[60]，引入了更多的注意力机制，分别为空间注意力、时间注意力和通道注意力，这些注意力均应用于图卷积后的特征，以对特征进行调整。

空间注意力模块可以帮助模型在不同的关节上赋予不同的注意力，其计算方式为：

$$M_s = \text{Sigmoid}(g_s(\text{AvgPool}(f_{\text{in}}))), \tag{4-5}$$

其中 $f_{\text{in}} \in \mathcal{R}^{C_{\text{in}} \times T \times V}$ 为输入特征，在时间维度平均池化后，通过 1×1 的卷积 g_s 将通道数调整为 1，经过 Sigmoid 激活函数后得到元素在 $[0, 1]$ 之间的 $M_s \in \mathcal{R}^{1 \times 1 \times V}$。

时间注意力模块与空间注意力模块相似，其计算方式为：

$$M_t = \text{Sigmoid}(g_t(\text{AvgPool}(f_{\text{in}}))), \tag{4-6}$$

计算得到的 $M_t \in \mathcal{R}^{1 \times T \times 1}$。

通道注意力模块可以帮助模型加强特征中更重要的通道，其计算方式为：

$$M_c = \text{Sigmoid}(W_2(\text{ReLU}(W_1(\text{AvgPool}(f_{\text{in}}))))), \tag{4-7}$$

在时间和空间维度对输入特征平均池化后，经过 W_1 和 W_2 两个全连接层，得到 $M_c \in \mathcal{R}^{C_{\text{in}} \times 1 \times 1}$。

如图4-7所示，计算完图卷积后的特征经过空间注意力模块后，得到空间注意力与原特征对应元素相乘后以类似残差连接的形式加到原特征上，再以相同的方式依次通过时间注意力和通道注意力模块，得到注意力调整后的输出。AAGCN 相比 AGCN 在各个维度都使用了注意力机制，效果更好。

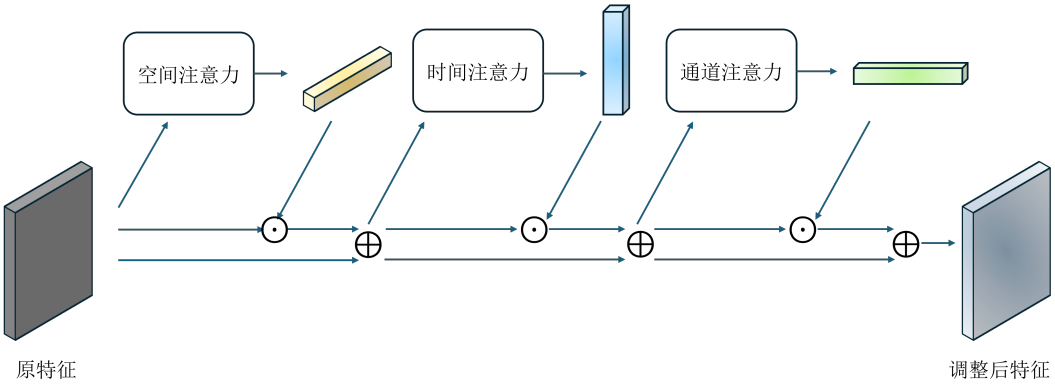


图 4-7　AAGCN 注意力模块流程

4.2.3　利用粗粒度标签的有监督对比学习分类算法

借鉴有监督对比学习损失函数相对于自监督对比学习损失函数的改进，本章提出的新损失函数为利用粗粒度标签的有监督对比损失（**Sup**ervised **Con**trastive loss **W**ith **C**oarse-grained labels，SupConWC），令

$$\mathcal{L}^{\text{finegrained}}=\sum_{i\in I}\mathcal{L}_i^{\text{finegrained}}=\sum_{i\in I}\frac{-1}{|P(i)|}\sum_{p\in P(i)}\log\frac{\exp\left(g_i\cdot g_p/\tau\right)}{\sum_{a\in A(i)}\exp\left(g_i\cdot g_a/\tau\right)}, \tag{4-8}$$

$$\mathcal{L}^{\text{coarse}}=\sum_{i\in I}\mathcal{L}_i^{\text{coarse}}=\sum_{i\in I}\frac{-1}{|Q(i)|}\sum_{q\in Q(i)}\log\frac{\exp\left(h_i\cdot h_q/\tau\right)}{\sum_{a\in A(i)}\exp\left(h_i\cdot h_a/\tau\right)}, \tag{4-9}$$

则

$$\mathcal{L}^{\text{SupConWC}}=\lambda\cdot\mathcal{L}^{\text{finegrained}}+(1-\lambda)\cdot\mathcal{L}^{\text{coarse}}, \tag{4-10}$$

其中 $I=\{1,2,\cdots,2N\}$，同样为 N 个样本经过两次不同数据增强后形成的双视图 batch；$P(i)$ 为在一个 batch 中除了锚样本 i 之外与其细粒度标签相同的样本集合，$Q(i)$ 为在一个 batch 中除了锚样本 i 之外与其粗粒度标签相同的样本集合；$A(i)$ 为除了锚样本 i 之外的其他样本。如图4-8所示，$g_l=\text{proj}(\text{Enc}(\tilde{x}_l))$，$h_l=\text{proj}'(\text{Enc}(\tilde{x}_l))$，即经过数据增强的样本 $\tilde{x}_l$ 通过编码器网络后的输出，通过两个不同的投影头得到的投影特征。两个投影头不共享权重，将编码器输出特征投影到不同的特征空间中，减少粗粒度损失和细粒度损失之间的相互干扰。$\lambda\in(0,1)$ 用于平衡粗粒度损失和细粒度损失，极端情况下，当 $\lambda=0$

时，该损失完全退化成只使用粗粒度标签信息的有监督对比学习；当 $\lambda = 1$ 时，该损失完全退化成只使用细粒度标签信息的有监督对比学习。

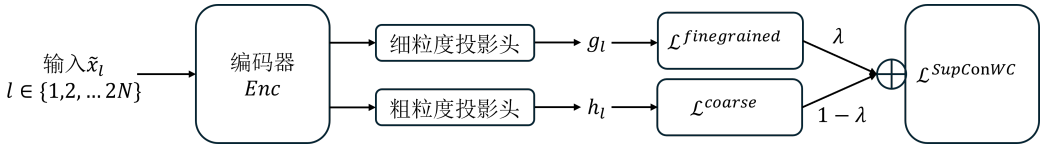


图 4-8 利用粗粒度标签的有监督对比学习损失计算结构图

图4-9展示了利用粗粒度标签的有监督对比学习损失的设计原理，图中不同颜色代表了不同细粒度标签的样本，不同的形状代表了不同粗粒度标签的标签。其原理是通过将编码器输出特征空间中的特征通过细粒度投影头投影到细粒度投影特征空间中，在此空间中，细粒度有监督对比学习损失将相同细粒度标签样本间的距离拉近；而在粗粒度投影特征空间中，粗粒度有监督对比学习损失将相同粗粒度标签样本间的距离拉近。在两个投影空间中的约束会共同间接地优化原编码器输出特征空间中的样本特征，使其达到更合理的分布。

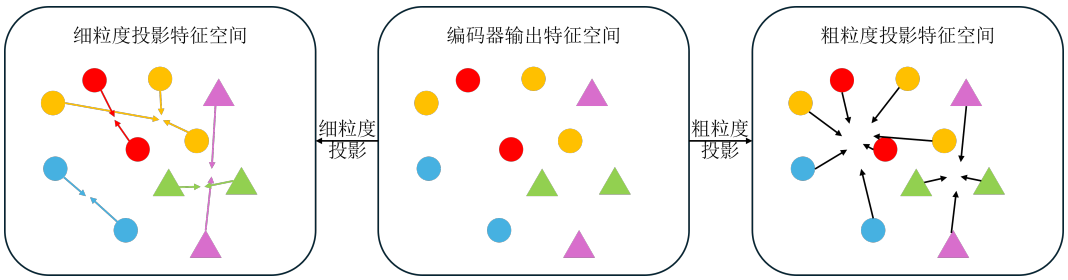


图 4-9 利用粗粒度标签的有监督对比学习损失原理示意图

4.2.4 利用粗粒度标签的有监督对比学习回归算法

不同于分类任务，回归任务的标签是连续的，这也就意味着很难再像 SupCon 那样对于某个锚样本找到细粒度标签与其完全一致的样本使它们之间的距离靠近。为了解决这个问题，Zha 等人[61]提出了 RnC（Rank-N-Contrast）损

失函数：

$$\mathcal{L}^{\text{RnC}} = \frac{1}{2N}\sum_{i=1}^{2N}\mathcal{L}_i^{\text{RnC}} \tag{4-11}$$

$$= \frac{1}{2N}\sum_{i=1}^{2N}\frac{1}{2N-1}\sum_{j=1,j\neq i}^{2N} -\log\frac{\exp(\text{sim}(v_i,v_j)/\tau)}{\sum_{v_k\in S_{i,j}}\exp(\text{sim}(v_i,v_k)/\tau)}. \tag{4-12}$$

其中 $v_l = \text{Enc}(\tilde{x}_l)$ 为数据增强后的样本 $\tilde{x}_l$ 经过编码器后的输出特征；$S_{i,j} = \{v_k \mid k \neq i, d(\tilde{y}_i,\tilde{y}_k) \geqslant d(\tilde{y}_i,\tilde{y}_j)\}$ 代表了对于锚样本 i 来说，标签距离大于样本对 (i,j) 的其他样本对集合，其中 $d(\cdot,\cdot)$ 用来计算两个样本的标签距离（如 L_1 距离）。

举例来说，假设在动作质量评估任务中，5 个样本的动作质量分别为 1、2、3、4 和 5 分，锚样本为 2 分样本，锚样本将与除自身外其他所有样本构成正样本对。对于正样本对（2 分，1 分）来说，其对应的负样本对为分数差大于等于 1 的（2 分，3 分）、（2 分，4 分）和（2 分，5 分）；对于正样本对（2 分，3 分）来说，其对应的负样本对为分数差大于等于 1 的（2 分，1 分）、（2 分，4 分）和（2 分，5 分）；对于正样本对（2 分，4 分）来说，其对应的负样本对为分数差大于等于 2 的（2 分，5 分）；没有比（2 分，5 分）分数差更大的负样本，因此这对不计算在损失内。通过优化损失函数使正样本对间的相似程度大于负样本对间的相似程度，最后将使得嵌入特征的序与它们在标签空间中的序相对应，即标签距离大的样本，其在特征空间中的表示也应该距离较大。

RnC 损失函数同样没有利用到样本的粗粒度标签信息，对于回归任务的样本来说，其粗粒度标签仍然可以是离散的，因此可以将其做与 SupConWC 相同的改进，称作 RnCWC（**R**ank-**N**-**C**ontrast loss **W**ith **C**oarse-grained labels）：

$$\mathcal{L}^{\text{RnCWC}} = \lambda\cdot\mathcal{L}^{\text{RnC}} + (1-\lambda)\mathcal{L}^{\text{coarse}}. \tag{4-13}$$

其中 $\mathcal{L}^{\text{coarse}}$ 依然是使用粗粒度标签信息的 SupCon 损失。

4.3 实验与分析

4.3.1 实验细节

本节实验均在 4 张 Tesla V100 32GB 显卡上进行。在将八段锦动作质量评估看作分类任务时，在预训练阶段 batch 大小选择为 64，学习率为 0.5，使用余弦退火算法调节学习率并进行学习率预热，温度超参数 τ 选择 0.1，λ 为 0.7，训练 200 个 epoch[1]；在分类阶段 batch 大小选择为 64，学习率为 0.01，训练 100 个 epoch。在训练阶段对训练集进行镜像翻转和重采样增强，其中镜像翻转的概率为 0.5，重采样帧数为 200。而在将八段锦动作质量评估看作回归任务时，温度超参数 τ 选择 2，其余超参数相同。

此外，本节还在另一些同样具有粗粒度标签和细粒度标签的分类和回归数据集上进行了实验，以验证本章算法的有效性。对于分类任务的数据集，在预训练阶段 batch 大小选择为 1024，学习率为 0.5，使用余弦退火算法调节学习率并进行学习率预热，温度超参数 τ 选择 0.1，λ 为 0.7，训练 200 个 epoch；在分类阶段 batch 大小选择为 1024，学习率为 5，训练 100 个 epoch。在训练集上进行相同的数据增强操作，即随机缩放裁剪成大小为 32×32 的图片，以及随机水平翻转。对于回归任务的数据集，本文遵循与 Zha 等人[61]工作中相同的实验设置，即预训练阶段训练 400 个 epoch，batch 大小为 256，学习率为 0.5，温度 τ 为 2，使用 L_1 距离作为标签距离，使用 L_2 距离作为特征间的距离；回归训练阶段的 batch 大小为 64，学习率为 0.05，训练 100 个 epoch。

4.3.2 对比实验

分类任务

八段锦动作质量评估数据集

表4-1展示了使用三个骨干网络分别在预训练阶段使用 SimCLR、SupCon 和 SupConWC 损失训练得到编码器后，冻结编码器参数后只训练全连接层分类器的对比实验，各模型预测分数的准确率如表所示。

1 根据 Khosla 等人[55]在论文中的实验得知，200 个 epoch 就足以使得模型收敛获得较好的表现，因此本文也选择训练 200 epoch。

ST-GCN 的图卷积的邻接矩阵是相对固定的，不能捕捉到未被人体骨骼直接连接的关节之间的联系，因此其表现相对最差，损失函数为 SimCLR 时准确率为 34.82%。使用 SupCon 损失训练后，准确率提升了 20.17%，而使用本文提出的 SupConWC 损失训练相较 SupCon，准确率进一步提升了 0.24%。

而 AGCN 由于引入了使用注意力机制的自适应图，可以建立不直接相连的关节之间的联系，表现相对 ST-GCN 有了一定的提升，损失函数为 SimCLR 时的准确率为 36.73%，相较 ST-GCN 提升了 1.91%。而使用 SupCon 损失训练后，准确率相较 SimCLR 提升了 20.64%，使用 SupConWC 后准确率进一步提升了 2.95%。

AAGCN 同时在空间、时间和通道维度上引入了注意力机制，效果在三个骨干网络模型中最佳。损失函数为 SimCLR 时准确率为 37.53%，而使用 SupCon 和 SupConWC 训练后准确率分别逐次提升了 23.33% 和 2.68%。

表 4-1 八段锦动作质量评估数据集分类实验结果

骨干网络	SimCLR[41]	SupCon[55]	SupConWC
ST-GCN[26]	34.82	54.99	**55.23**
AGCN[59]	36.73	57.37	**60.32**
AAGCN[60]	37.53	60.86	**63.54**

从以上实验中可以得出结论，本文提出的 SupConWC 在八段锦动作质量评估数据集上有出色的表现，仅需要替换掉损失函数而不改动网络结构即可获得较为显著的提升，这说明了利用粗粒度标签的有监督对比学习算法在分类任务上的有效性。

CIFAR-100

CIFAR-100 数据集共有 100 个细粒度类别，每个细粒度类别有 600 张图片，其中 500 张用作训练集，100 张用作测试集。此外，数据集将 100 个细粒度类别聚合成了 20 个超类，例如“人”超类有“婴儿”、“男孩”、“女孩”、“男人”和“女人”五个细粒度子类。这也为本文测试利用粗粒度标签的有监督对比学习效果提供了便利。

表4-2展示了三个骨干网络分别使用三种不同的损失函数在 CIFAR-100 数据集上预训练后再在数据集上训练分类器后的分类准确率。从表中可以看出，SimCLR（自监督对比学习）效果最差，本文提出的 SupConWC 在使用

表 4-2 CIFAR-100 数据集实验结果

骨干网络	SimCLR[41]	SupCon[55]	SupConWC
ResNet-18	59.15	73.08	**73.18**
ResNet-50	64.71	72.97	**73.67**
ResNet-101	65.50	77.27	**77.33**

ResNet-18 作为编码器时，相比 SupCon 准确率提升 0.1%；在使用 ResNet-50 作为编码器时，相比 SupCon 准确率提升 0.7%；在使用 ResNet-101 作为编码器时，相比 SupCon 准确率提升 0.06%，均取得了最好的效果。

Stanford Cars 数据集

Stanford Cars 是另一个在图像分类领域广泛使用的基准测试数据集，其中包含 196 种不同的车，但是该数据集并没有提供官方的粗粒度标签，本文遵循 Feng 等人工作[58]中的划分方式，将 196 个类别映射到 8 个粗粒度标签，分别为“0：敞篷车（Cab）”、“1：轿车（Sedan）”、“2：运动型多功能车（SUV）”、“3：敞篷轿车（Convertible）”、“4：双门轿车（Coupe）”、“5：掀背式轿车（Hatchback）”、“6：旅行车（Wagon）”和“7：货车（Van）”。

表 4-3 Stanford Cars 数据集数据集实验结果

骨干网络	SimCLR[41]	SupCon[55]	SupConWC
ResNet-18	11.40	44.57	**46.86**
ResNet-50	10.11	31.08	**40.16**
ResNet-101	6.85	30.00	**32.86**

表4-3展示了三个骨干网络分别使用三种不同的损失函数在 Stanford Cars 数据集上预训练后再在数据集上训练分类器后的分类准确率。从表中可以看出，SimCLR 的表现远远不如 SupCon 及 SupConWC，这可能是因为数据集被缩放为 32×32，损失了较多的信息，仅靠自监督对比学习并不能很好地学到好的样本表示以供下游分类任务使用。本文提出的 SupConWC 在使用 ResNet-18 作为编码器时，相比 SupCon 准确率提升 2.29%；在使用 ResNet-50 作为编码器时，相比 SupCon 准确率提升 9.08%；在使用 ResNet-101 作为编码器时，相比 SupCon 准确率提升 2.86%，均取得了最好的效果。值得注意的一点是，在 Stanford Cars 数据集上并不是参数量越大的网络效果越好，这可能与数据集较小，参数量大的网络容易过拟合难以学习有关。

Stanford Online Products 数据集

Stanford Online Products 数据集共有 22634 类，120053 张商品图片。该数据集有 12 种粗粒度标签，如“自行车”和“水壶”等，而细粒度标签则是用来区分不同的自行车和水壶等同类不同种商品，每一个细粒度标签下有若干张不同视角的同种商品图。本文依然遵循 Feng 等人工作[58]中的划分方式，从数据集中选出 1498 个细粒度类别，每种类别有 12 张图片，其中 10 张用作训练集，2 张用作测试集。

表 4-4 Stanford Online Product 数据集数据集实验结果

骨干网络	SimCLR[41]	SupCon[55]	SupConWC
ResNet-18	47.20	51.57	**53.81**
ResNet-50	41.32	48.43	**55.67**
ResNet-101	48.70	45.49	**52.64**

表4-4展示了三个骨干网络分别使用三种不同的损失函数在 Stanford Online Product 数据集上预训练后再在数据集上训练分类器后的分类准确率。从表中可以看出，SimCLR 的表现在使用 ResNet-18 和 ResNet-50 时依然不如 SupCon 及 SupConWC，但在使用 ResNet-101 时相较 SupCon 却提高了 3.21%。本文提出的 SupConWC 在使用 ResNet-18 作为编码器时，相比 SupCon 准确率提升 2.24%；在使用 ResNet-50 作为编码器时，相比 SupCon 准确率提升 7.24%；在使用 ResNet-101 作为编码器时，相比 SupCon 准确率提升 7.15%，均取得了最好的效果。

从在八段锦动作质量评估数据集上的实验及 CIFAR-100、Stanford Cars 和 Stanford Online Products 等其他分类数据集上的实验可以看出，SimCLR 的效果差于 SupCon，这验证了有监督信息的重要性；而 SupConWC 的效果又优于 SupCon，这验证了粗粒度标签信息的重要性以及 SupConWC 的有效性。

回归任务

八段锦动作质量评估数据集

表4-5展示了使用三个骨干网络分别进行端到端直接回归和预训练阶段使用 RnC 或 RnCWC 损失训练得到编码器后，冻结编码器参数训练回归器参数等三种设置的对比实验后各模型预测分数的 L_1 损失。

ST-GCN 由于受限于固定的图卷积邻接矩阵，其表现依然是最差的，直接回归的 L_1 损失为 0.644。使用 RnC 损失训练后，L_1 损失降低了 0.057，而本文提出的 RnCWC 损失训练相较 RnC，L_1 损失进一步降低了 0.043。

引入注意力机制建立不直接相连的关节间的联系后，AGCN 表现相对 ST-GCN 有了一定的提升，直接回归的 L_1 损失为 0.580，相较 ST-GCN 下降了 0.064。而使用 RnC 训练后，损失下降了 0.061，使用 RnCWC 后损失进一步降低了 0.018。

AAGCN 较 AGCN 更完备的注意力机制使其在三个骨干网络模型中表现最佳，直接回归的 L_1 损失为 0.570，而使用 RnC 和 RnCWC 训练后损失分别又下降了 0.049 和 0.023。

表 4-5 八段锦动作质量评估数据集回归实验结果

骨干网络	直接回归	RnC[61]	RnCWC
ST-GCN[26]	0.644	0.587	**0.544**
AGCN[59]	0.580	0.519	**0.501**
AAGCN[60]	0.570	0.521	**0.498**

从以上实验中可以得出结论，本文提出的 RnCWC 在八段锦动作质量评估数据集的回归任务上也有出色的表现，这说明了利用粗粒度标签的有监督对比学习算法同时适用于分类及回归任务。

为了对模型打分效果有更直观的了解，表4-6展示了 AAGCN 模型在从测试集随机抽样的 40 个样本（各动作各质量分数均抽样 1 个样本）上的预测分数，表中括号内数字为模型预测分数与真实分数的差。从表中可以看出，模型对 5 分样本的预测更准确，而对于 1 分和 2 分样本的预测偏差相对较大，这可能是由数据集中低分样本数量较少导致的。但整体而言，模型可以很好地学习到不同动作质量样本之间的差异，将低分样本与高分样本区分开。此外，相较于将八段锦动作质量评估任务看作分类任务，将其看作回归任务得到的分数是连续的，可以给人以更精细的反馈。

AgeDB 数据集

本文选取了 AgeDB 数据集[62]来验证利用粗粒度标签的有监督对比学习用于其他回归任务的有效性。AgeDB 是一个通过人脸图像预测年龄的数据集，共包含 568 个人的 16488 张图像，平均每个人有各个年龄的 29 张图像。

表 4-6 八段锦动作质量评估测试集随机抽样推理结果

动作	1 分样本	2 分样本	3 分样本	4 分样本	5 分样本
预备式	0.79(-0.21)	1.61(-0.39)	2.68(-0.32)	4.20(0.20)	4.58(-0.42)
双手托天理三焦	2.33(1.33)	2.48(0.48)	3.19(0.19)	4.45(0.45)	5.00(0.00)
左右开弓似射雕	1.32(0.32)	2.04(0.04)	3.36(0.36)	4.37(0.37)	4.97(-0.03)
调理脾胃须单举	2.39(1.39)	2.26(0.26)	2.74(-0.26)	3.66(-0.34)	4.94(-0.06)
五劳七伤往后瞧	1.45(0.45)	3.11(1.11)	3.65(0.65)	4.64(0.64)	4.98(-0.02)
摇头摆尾去心火	1.44(0.44)	1.67(-0.33)	3.12(0.12)	4.44(0.44)	4.97(-0.03)
双手攀足固肾腰	2.73(1.73)	2.24(0.24)	2.93(-0.07)	4.35(0.35)	4.97(-0.03)
攒拳怒目增气力	2.65(1.65)	3.35(1.35)	3.37(0.37)	3.50(-0.50)	4.94(-0.06)

在该数据集中，年龄是细粒度标签，而不同的个体则是粗粒度标签。值得注意的是，在回归任务中粗粒度标签与细粒度标签之间不一定是包含关系，例如若干年龄不同的样本同属于一个个体，但不同个体间会存在年龄相同的样本。也就是说，在不指定样本的情况下，仅知道细粒度的年龄是无法将其对应到粗粒度个体标签上的。此时粗粒度标签体现的是若干细粒度样本之间的共性，而并非是若干细粒度标签的并集。

表4-7展示了使用 ResNet-18 作为骨干网络，各种学习方法在 AgeDB 数据集上的 L_1 损失（带星号的是本文重新进行实验得出的结果）。其中表示学习方法（Linear Probing）指的是冻结预训练的编码器，只优化回归器的参数；表示学习方法（Fine-tuning）指的是不冻结预训练的编码器，其参数与回归器的参数一同优化。从表中可以看出，本文方法 RnCWC 取得了最好的 L_1 损失 6.12，比直接使用 L_1 回归降低 0.42，比改进前的 RnC 方法降低 0.14。这进一步说明了在回归任务上引入粗粒度标签信息同样可以提升模型的表现，本文提出的利用粗粒度标签的有监督对比学习损失同时适用于分类和回归任务，这体现了其优越性。

4.3.3 表征分析

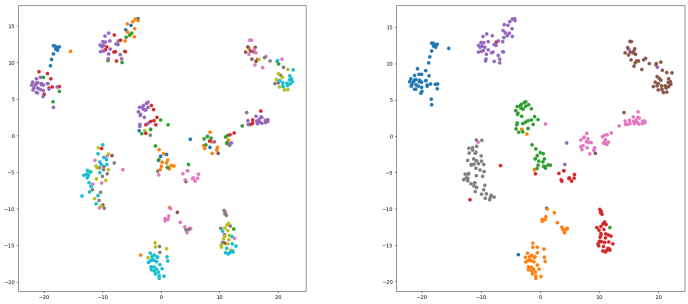
为了验证利用粗粒度标签的有监督对比学习是否能够使得样本在编码器输出的特征空间中更合理地分布，本节使用 t-SNE 算法[67]将编码器输出特征降维至二维空间，并分别以不同的颜色来标示细粒度标签和粗粒度标签。

如图4-10和4-11所示，子图4-10(a)和4-11(a)分别为使用 SupCon 和 SupConWC 损失函数在八段锦动作质量评估数据集上训练后，编码器降维特征

表 4-7 AgeDB 数据集实验结果

方法	算法	L_1 损失
表示学习方法（Linear Probing）	SimCLR[41]	9.59
	DINO[63]	10.26
	SupCon[55]	8.13
表示学习方法（Fine-tuning）	SimCLR[41]	6.57
	DINO[63]	6.61
	SupCon[55]	6.55
回归学习方法	L_1	6.54*
	LDS+FDS[64]	6.45
	RankSim[65]	6.51
	Ordinal Entropy[66]	6.57
	RnC(L_1)[61]	6.26*
本文方法	RnCWC	6.12

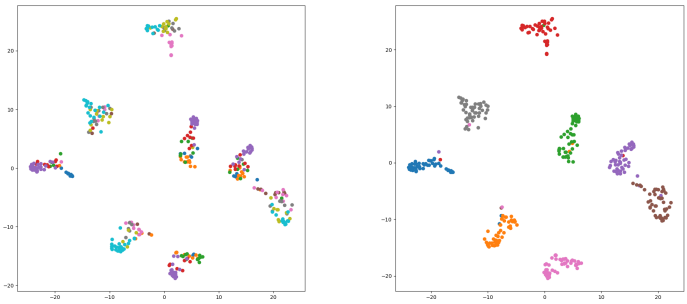
的分布，不同的颜色代表了不同的细粒度标签；而子图4-10(b)和4-11(b)的颜色代表了不同的粗粒度标签。子图4-10(a)和4-10(b)数据点的分布是相同的，子图4-11(a)和4-11(b)数据点分布也是相同的，区别仅在于颜色代表了不同粒度的标签。通过对比子图4-10(b)和4-11(b)，可以发现相比于 SupCon，SupConWC 算法相同粗粒度标签的各个细粒度样本之间的距离被拉近，数据点聚合程度更高。例如原本有一部分红色样本点与绿色样本点更为接近，而改进后的红色样本点与绿色样本点之间有了明显的距离。这说明利用粗粒度标签的有监督对比学习确实可以使得同一粗粒度的各细粒度样本之间的相对距离拉近，分布更合理。



(a) 八段锦数据集细粒度标签 SupCon 特征可视化　　**(b) 八段锦数据集粗粒度标签 SupCon 特征可视化**

图 4-10 八段锦数据集细粒度和粗粒度标签 SupCon 特征可视化图

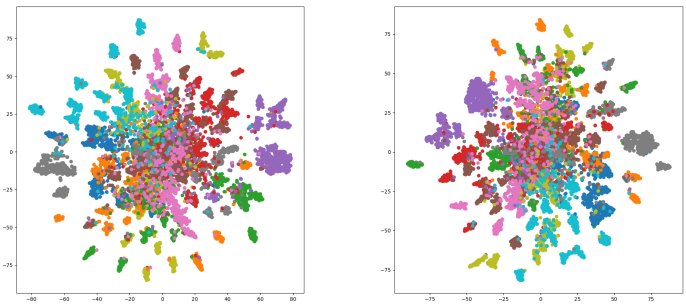
4-12展示了 SupCon 和 SupConWC 在 CIFAR-100 数据集上的 t-SNE 降维结果。通过对比子图4-12(a)和4-12(b)，可以发现4-12(b)中的深蓝色、紫色数



(a) 八段锦数据集细粒度标签 SupConWC 特征可视化　(b) 八段锦数据集粗粒度标签 SupConWC 特征可视化

图 4-11　八段锦数据集细粒度和粗粒度标签 SupConWC 特征可视化图

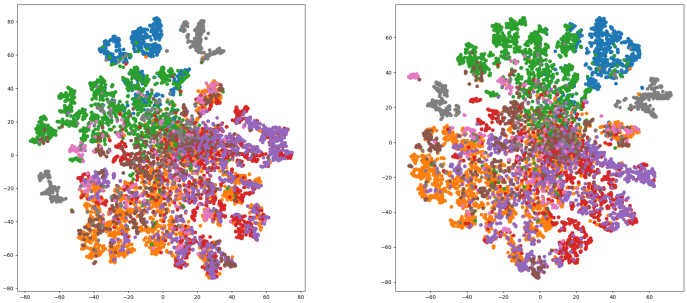
据点聚合程度更高。



(a) CIFAR-100 数据集粗粒度标签 SupCon 特征可视化　(b) CIFAR-100 数据集粗粒度标签 SupConWC 特征可视化

图 4-12　CIFAR-100 数据集粗粒度标签 SupCon 与 SupConWC 特征可视化图

图4-13则展示了 SupCon 和 SupConWC 在 Stanford Cars 数据集上的 t-SNE 降维结果。子图4-13(b)与4-13(a)相比，其中的红色、绿色等数据点相对集中。

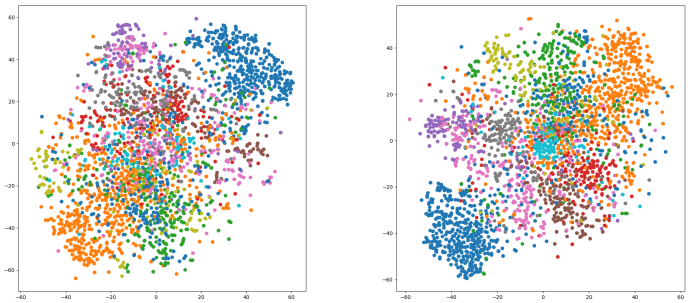


(a) Stanford Cars 数据集粗粒度标签 SupCon 特征可视化　(b) Stanford Cars 数据集粗粒度标签 SupConWC 特征可视化

图 4-13　Stanford Cars 数据集粗粒度标签 SupCon 与 SupConWC 特征可视化图

图4-14展示了 SupCon 和 SupConWC 在 Stanford Online Products 数据集上的 t-SNE 降维结果。在该数据集上，SupConWC 相对于 SupCon 的优势更为明显，子图4-14(a)整体是比较散乱的，而子图4-14(b)的各个颜色标签基本上聚合在了一起，尤其是浅蓝色、红色和灰色等。

通过对四个数据集上分别使用 SupCon 和 SupConWC 损失函数训练出的编码器的输出特征降维可视化的观察，可以得出结论，SupConWC 使得编码器输出特征分布更符合人类认知，进一步说明了其有效性。



(a) Stanford Online Products 数据集粗粒度标签 SupCon 特征可视化

(b) Stanford Online Products 数据集粗粒度标签 SupConWC 特征可视化

图 4-14 Stanford Online Products 数据集粗粒度标签 SupCon 与 SupConWC 特征可视化图

4.3.4 消融实验

数据增强消融实验

表4-8展示了 AAGCN 作为骨干网络，RnCWC 作为损失函数，使用不同数据增强方法实验的结果。其中只使用镜像翻转训练出的模型 L_1 损失为 0.529，相比于同时使用镜像翻转和重采样升高了 0.031；只使用重采样的模型 L_1 损失为 0.519，升高了 0.021；而不使用数据增强的模型 L_1 损失为 0.530，升高了 0.032。从而可以得出结论，镜像翻转和重采样等数据增强方法对于利用对比学习算法的八段锦动作质量评估的性能是十分重要的。镜像翻转和重采样都可以变相地扩增数据集，提高模型的鲁棒性；而重采样还可以提高模型对长时间段样本的处理能力。

表 4-8 八段锦动作质量评估数据集数据增强消融实验结果

数据增强方法	L_1 损失
只镜像翻转	0.529
只重采样	0.519
无数据增强	0.530
镜像翻转 & 重采样	**0.498**

重采样帧数的影响

图4-15展示了重采样帧数分别为 10、30、50、100、150、200、250 和 300 时训练得到的模型的 L_1 损失，可以看出在重采样帧数达到了一定数量（100 帧）后，再增加帧数并不一定能提高模型的性能，甚至会使模型性能下降。这可能是因为采样到的帧中已经有足够的信息，盲目地增加帧数反而会使得训练和推理的效率降低，同时增加训练和收敛的难度而使性能下降。但是重采样帧数较低时，采样到的帧中信息不足，模型不能充分地学习到质量评估相关的知识。

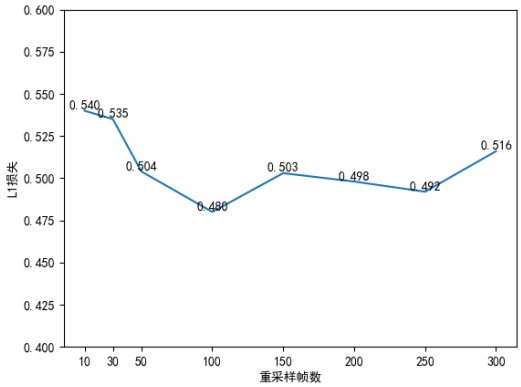


图 4-15 不同重采样帧数对模型 L_1 损失的影响图

串行投影头

图4-8的利用粗粒度标签的有监督对比学习损失计算结构图中两个投影头的输入均为编码器的输出，这两个投影头是并行的。但如果如图4-16所示，将粗粒度投影头的输入改为细粒度投影头的输出，使得两个投影头是串行的，结果如表4-9所示，其展示了使用串行投影头和并行投影头在四个数据集上，使用三个不同骨干网络时的效果。

在八段锦动作质量评估数据集上，ST-GCN 网络使用串行投影头的准确率相比并行投影头提高 0.80%，AGCN 网络降低 3.75%，AAGCN 网络则降低了 5.90%；在 CIFAR-100 数据集上，ResNet-18 网络使用串行投影头的准确率相比并行投影头降低 0.01%，ResNet-50 网络则降低 0.04%，然而 ResNet-101 网络使用串行投影头的准确率相比并行投影头提高了 0.19%；在 Stanford Cars 数据集上，ResNet-18 网络使用串行投影头的准确率相比并行投影头降低 1.93%，ResNet-50 网络降低 4.51%，ResNet-101 网络降低 0.66%；在 Stanford Online Products 数据集上，ResNet-18 网络使用串行投影头的准确率相比并行投影头降低 0.21%，ResNet-50 网络降低 1.86%，ResNet-101 网络降低 6.21%。也就是说，除了在八段锦动作质量评估数据集上训练的 ST-GCN 以及 CIFAR-100 数据集上训练的 ResNet101 网络之外，其他的实验都说明并行投影头的效果优于串行投影头，而在八段锦动作质量评估任务中 AGCN 和 AAGCN 网络串行投影头的效果比 SupCon 更低，这说明串行投影头的效果并不稳定，所以在一般情况下选择并行投影头即可。

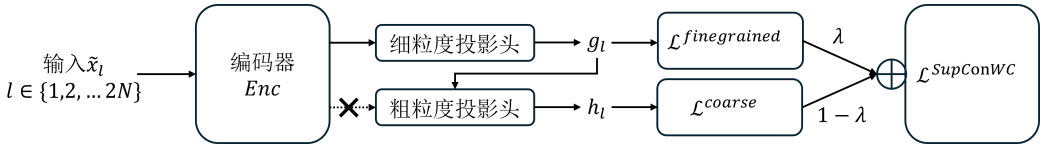


图 4-16 串行投影头结构

表 4-9 串行投影头对比实验结果

数据集	骨干网络	SupCon[55]	SupConWC（并行）	SupConWC（串行）
八段锦动作质量评估	ST-GCN[26]	54.99	55.23	**56.03**
	AGCN[59]	57.37	**60.32**	56.57
	AAGCN[60]	60.86	**63.54**	57.64
CIFAR-100	ResNet-18	73.08	**73.18**	73.17
	ResNet-50	72.97	**73.67**	73.63
	ResNet-101	77.27	77.33	**77.52**
Stanford Cars	ResNet-18	44.57	**46.86**	44.93
	ResNet-50	31.08	**40.16**	35.65
	ResNet-101	30.00	**32.86**	32.20
Stanford Online Products	ResNet-18	51.57	**53.81**	53.60
	ResNet-50	48.43	**55.67**	53.81
	ResNet-101	45.49	**52.64**	46.43

超参数 λ 的影响

上文提到超参数 λ 决定了粗粒度损失和细粒度损失之间的平衡，为了探究其影响，本文在 Stanford Online Products 数据集上使用 ResNet-18 作为骨干网络，分别测试 λ 取值为 0.1、0.3、0.5、0.7、0.9 和 1 时的表现。图4-17则展示了实验的结果。从图中可以看出，λ 取值为 0.1 时，在预训练阶段粗粒度损失几乎贡献了所有的最终损失，而下游任务是细粒度的分类，因此此时取得了最低的准确率 47.60%；而随着 λ 增加为 0.3、0.5 和 0.7，准确率也依次上升了 4.64%、0.73% 和 0.84%；当 λ 进一步增大为 0.9 和 1 时，在预训练阶段细粒度损失的贡献逐渐增大，而 $\lambda = 1$ 时损失完全退化为细粒度损失，准确率也开始下降。从该实验中可以得出，λ 选择 0.7 较为合适，本文中所有数据集上的实验均采用此设置，此时细粒度损失的贡献占大部分，有利于下游的细粒度分类或回归任务，而适当比例的粗粒度损失也可以引导样本特征在空间中的分布更为合理。此外，实验还验证了超参数 λ 的选择较为鲁棒，其取值在 0.3-0.7 之间时准确率均优于 SupCon，这进一步体现了本章提出的损失的优越性。

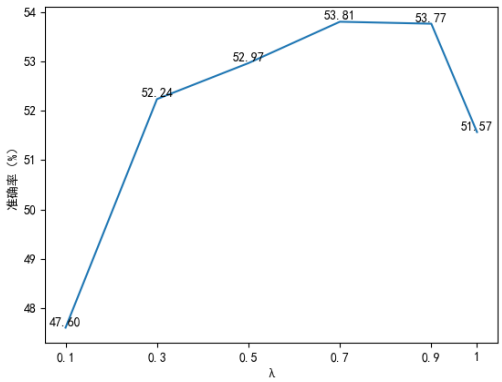


图 4-17 λ 对分类准确率的影响曲线图

4.4 本章小结

本章从有监督对比学习及人类认识事物的规律得到启发，提出了利用粗粒度标签的有监督对比学习（SupConWC 和 RnCWC），分别将八段锦动作质量评估任务看作分类和回归任务进行实验，验证了本章提出算法的有效性。

此外，本章还在 CIFAR-100、Stanford Cars 和 Stanford Online Products 三个分类任务数据集及 AgeDB 等回归任务数据集上进行了对比实验，同样取得了显著的效果提升，进一步验证了本章算法的有效性。本章还对使用 SupConWC 训练出的编码器输出特征进行了表征分析，说明了利用粗粒度标签确实可以使特征分布更符合人类的认知。此外，本章还探究了并行串行投影头的影响，得出了一般情况下并行投影头效果更好的结论。而超参数 λ 的取值较为鲁棒，在 0.3-0.9 之间时都远远优于 SupCon，这进一步体现了该算法的优越性。

第五章　八段锦动作质量评估系统

为了验证本文建立的八段锦动作质量评估数据集及本文提出的利用粗粒度标签的有监督对比学习算法在真实场景中的有效性，本章搭建了一个八段锦动作质量评估系统，并将在本文建立的数据集上使用本文提出的算法训练出的八段锦动作质量评估模型应用于其中。本章的系统旨在满足用户足不出户就可以有效地进行八段锦锻炼身体的需求。通过该系统，用户可以通过自己的摄像头进行八段锦动作的评估和反馈，从而帮助他们改善动作的准确性和流畅性，提高锻炼效果。本章节将对该系统的用户需求、系统整体设计、各个模块的实现细节和系统展示等方面进行详细介绍。

5.1　需求分析

为了达成帮助用户足不出户就能改善八段锦动作，同时可以提高锻炼积极性的目的，本系统需要支持以下需求：

- 八段锦动作完整跟随练习：用户可以在屏幕上同时看到八段锦标准动作视频及自身当前的动作，通过模仿标准动作进行从头到尾的练习，在结束后得到评分反馈。
- 八段锦动作分解跟随练习：用户在本系统的帮助下了解到某个动作自己做不好，或者想加强对某个动作的练习，可以对照屏幕上的特定动作的标准视频与自身动作进行针对性的练习和纠正，在结束后得到评分反馈。
- 八段锦动作自由练习：比较熟练的用户可能不需要跟随标准八段锦动作的示范，因此在该模式下用户只能看到自身的动作，但是可以听到八段锦的动作口令，在结束后得到评分反馈。
- 用户练习视频上传：系统需要将用户从摄像头采集到的视频上传至服

务器，以便于后续的分析处理。

- 姿态估计：在系统将用户的八段锦练习视频上传至服务器后，系统可以将练习视频中人体的各个关键点识别并提取出来，以供后续动作质量评估。
- 动作质量评估：系统将人体的各个关键点输入到动作质量评估模型中，得到反馈评分。
- 练习记录：用户可以查看过去所有的练习视频及对应评分，并且可以看到自身练习评分随时间的变化，也可以看到不同动作的平均分数，以便对薄弱动作进行针对性的练习。

5.2 系统设计

5.2.1 整体设计

图5-1展示了本系统的整体拓扑结构，图中箭头代表了数据流动的方向。本系统采用前后端分离的架构，前端为HTML网页，后端使用轻量级的Python Flask框架开发。前端负责展示用户界面，与用户交互，并调用后端提供的算法接口和数据接口。算法接口包括姿态估计和动作质量评估，在进行模型推理时，也需要调用数据接口获取输入及保存输出；数据接口则是对数据库操作的抽象封装，包括对系统数据（如标准动作视频、模型权重等）及用户数据（如用户上传视频、用户练习记录等）进行操作的接口。

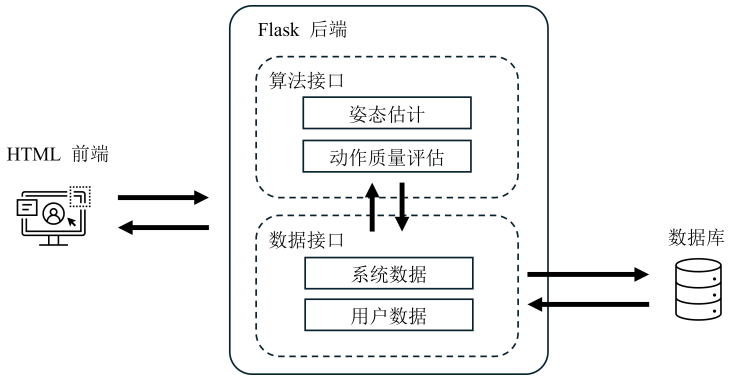


图 5-1 系统拓扑结构图

图5-2展示了本系统的整体操作流程。用户在打开本系统之后，首先会看到选择菜单，用户可以在“完整跟练”、“分解跟练”、“自由练习”和“练习记

录”之间选择想要的功能。当用户选择“完整跟练”，或者选择“分解跟练”并选择要练习的动作之后，系统会开始播放标准动作视频；当用户选择“自由练习”，系统只会播放八段锦的口令音频。此后系统将记录用户的动作并上传至服务器进行动作质量评估，之后展示在前端。而当用户选择“练习记录”后，会进入二级菜单，选择“查看记录”以查阅过往的练习记录视频及评分；选择“得分曲线图”可以查看自己的评分随练习八段锦次数的变化曲线图；选择“薄弱动作”则可以查看八段锦各段动作分别统计的平均分，得知自身的薄弱动作以进行针对练习。

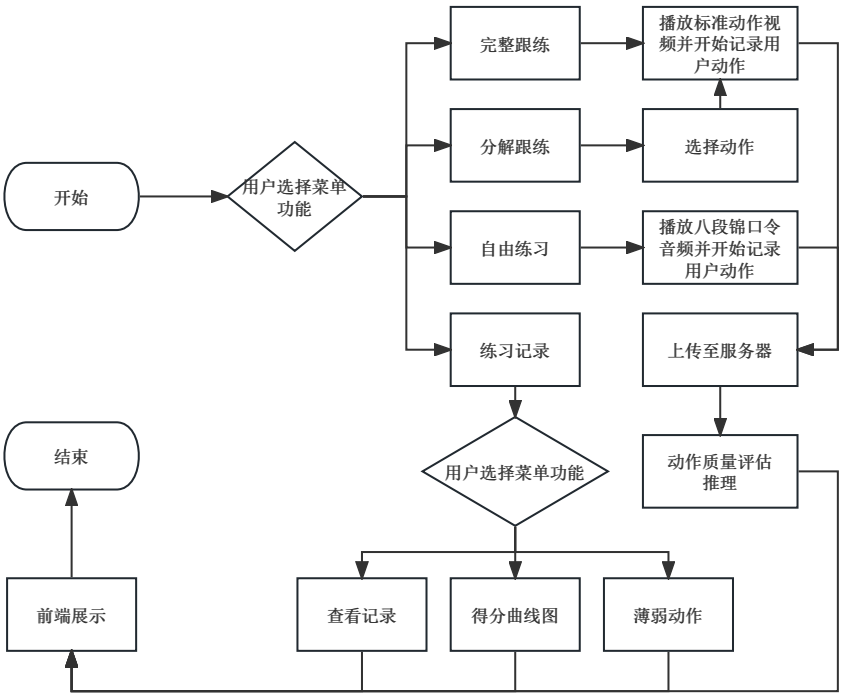


图 5-2 系统流程图

5.2.2 完整跟练模块

图5-3展示了完整跟练模块的流程图，流程图中同时展示了前端和后端的运行流程。需要注意的是，用户在点击“开始练习”后，摄像头开始记录用户的练习动作，并每隔一段时间从中截取一帧，将其上传到服务器。为了与第四章训练动作质量评估模型时用到的数据保持一致，每段动作都会被采样 200 帧，由于不同动作的持续时间不同，因此不同动作的采样间隔也不同。

表5-1展示了标准八段锦动作视频各段动作的起止时间、持续时间及相应的采样间隔。简便起见，系统前端会严格按照表5-1中的时间点切换当前动作及采样间隔，将每段动作采样得到的 200 帧图像分别保存。服务器在接收到视频帧后，会立即调用姿态估计算法识别视频帧中人体各个关键点并将其保存到服务器中。这样做的好处是在用户正在练习的同时服务器便已经开始同步进行姿态估计的处理，而不是等到结束后一起发送到服务器再进行处理，减少了用户的等待时间。在标准视频播放完毕后，前端将会调用后端的动作质量评估接口，将之前处理好的人体关键点信息作为输入推理得到动作质量的评分并展示给用户。同时，评分也会保存到服务器中，以供后续“练习记录”模块的使用。

表 5-1 标准八段锦动作视频各段动作的起止时间、持续时间及相应的采样间隔

动作	开始时间（秒）	结束时间（秒）	持续时间（秒）	采样间隔（秒）
预备式	22	45	23	0.115
双手托天理三焦	45	130	85	0.425
左右开弓似射雕	130	223	93	0.465
调理脾胃须单举	223	295	72	0.36
五劳七伤往后瞧	295	371	76	0.38
摇头摆尾去心火	371	470	99	0.495
双手攀足固肾腰	470	596	126	0.63
攒拳怒目增气力	596	662	66	0.33

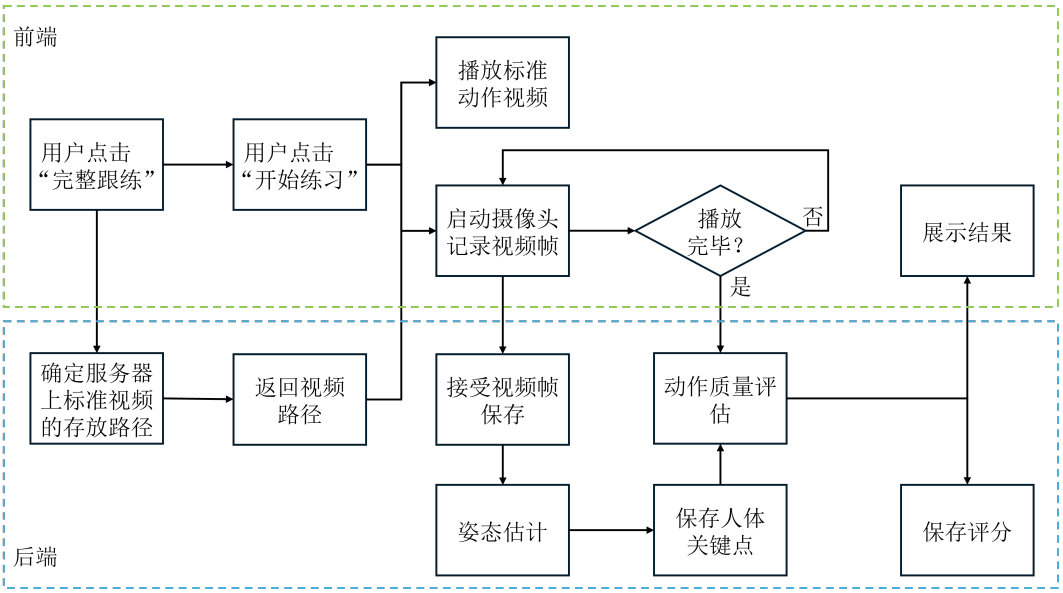


图 5-3 完整跟练模块流程图

“分解跟练”和“自由练习”模块的结构与“完整跟练”模型类似，区别

在于“分解跟练”模块还需要再选择要练习的动作，而“自由练习”模块不再播放标准动作视频，因此其流程图在此不再赘述。

5.2.3 练习记录模块

“练习记录”模块包含查看练习记录、查看得分曲线图和发现薄弱动作三个功能。图5-4展示了查看练习记录子模块的流程图，图5-5展示了发现薄弱动作的流程图，查看得分曲线图的流程类似。得分曲线图和发现薄弱动作功能中的各动作平均得分图均由前端使用 Chart.js 提供的接口绘制。相较于在后端绘图并返回给前端，这样做既可以节省传输时间（只需要传输若干分数而不是一张图像），又可以使图响应式地根据设备调整大小而不模糊。此外，使用 Chart.js 绘制的图表在默认设置下也非常美观，便于用户识读。

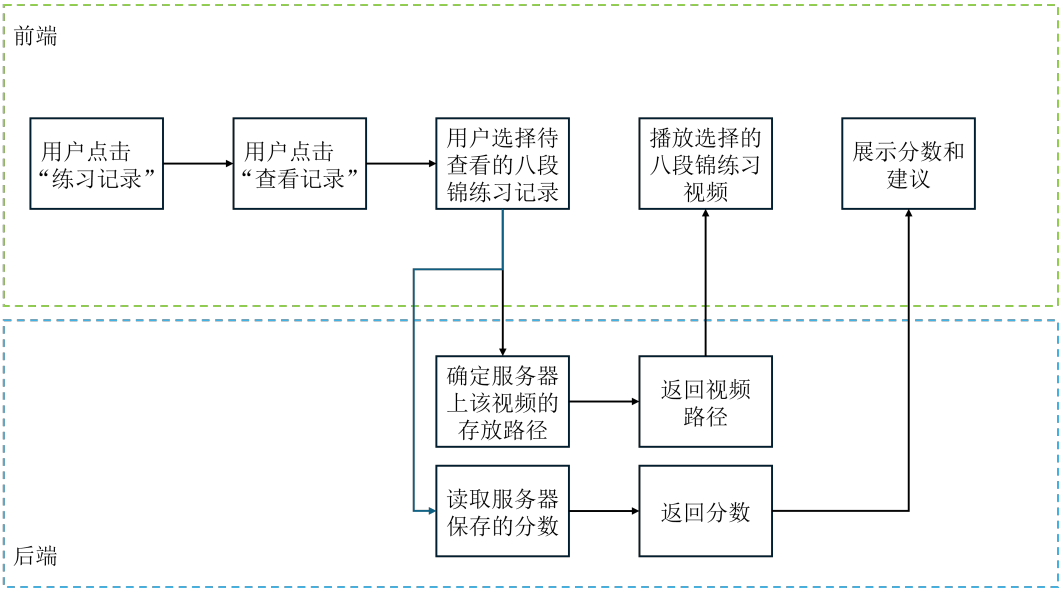


图 5-4 查看练习记录子模块流程图

5.3 系统展示

5.3.1 运行环境

本系统的前端对性能要求较低，几乎可以运行在任何有浏览器的系统中；本系统的后端对性能要求较高，最好运行在有 GPU 的服务器中。在本文的部署环境中，系统后端的 CPU 为 Intel(R) Xeon(R) Gold 6248，GPU 为

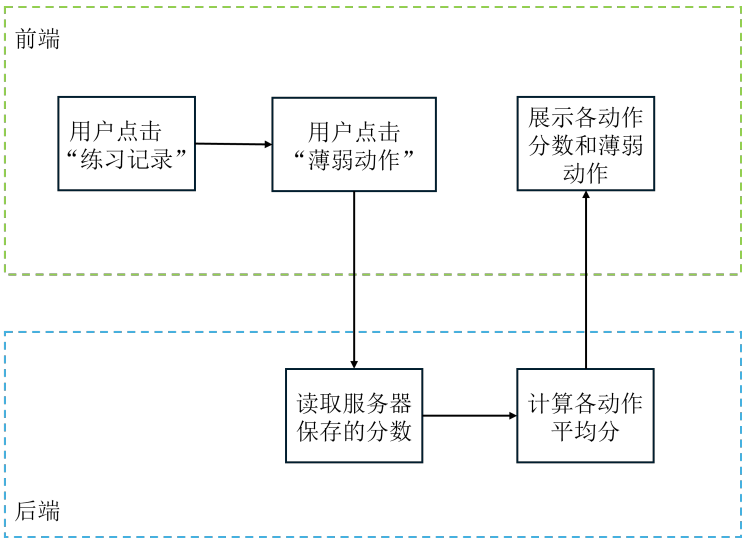


图 5-5 发现薄弱动作子模块流程图

Nvidia Tesla V100 32GB。Python 的版本为 3.10，PyTorch 的版本为 1.10.0。本文将 OpenPose 姿态估计模型和本系统后端部分打包为了 Docker 镜像，方便在不同机器上进行部署。

5.3.2 各模块运行效果展示

图5-6展示了八段锦动作质量评估系统的主界面，用户除了可以在界面中央选择功能外，还可以在左侧导航栏选择，且导航栏在各个功能界面均显示，方便用户随时切换。在用户选择了“完整跟练”后，就会进入图5-7所示的界面，除了练习结束得到评分外，在练习过程中还可以点击“提前结束”按钮，此时系统会调用后端接口，评估已完成动作的质量。

图 5-6 八段锦动作质量评估系统主界面

当用户选择“分解跟练”后，通过图5-8所示的菜单选择要练习的动作，

图 5-7 完整跟练界面

之后进入与完整跟练类似的界面；而“自由练习”的界面则如图5-9所示，此时界面中只能看到练习者自身的动作。在这两个功能中，开始练习后同样可以点击按钮提前结束练习。不过对于“分解跟练”功能来说，由于每次只练习一段动作，提前结束会导致采样帧数不足，此时并不会将已采样的部分进行动作质量评估。

图 5-8 分解跟练选择动作界面

图 5-9 自由练习界面

图5-10展示了练习记录功能选择界面。在图5-11所示的得分曲线图界面中，用户可以看到得分随练习时间的变化。在图5-12所示的薄弱动作界面中，用户可以看到各个动作的平均得分，系统同时也会指出得分最低的两个动作以便用户针对性地练习。在图5-13所示的查看记录界面中，用户可以看到所有的练习历史，选中某次练习可以查看练习回放，同时在右侧可以看到各个动作及整体的分数。

图 5-10 练习记录功能选择

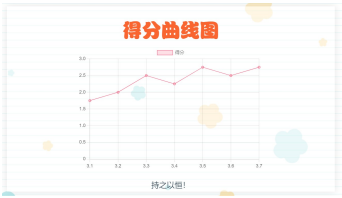


图 5-11 得分曲线图界面

图 5-12 薄弱动作界面

图 5-13 查看记录界面

5.4 本章小结

本章详细介绍了八段锦动作质量评估系统前后端的设计方案和实现细节。首先对八段锦动作质量评估系统的需求进行汇总，介绍了该系统的几项基本功能；之后介绍了系统整体的拓扑结构和流程图，并对其各个组成模块详细介绍了其实现细节，最后展示了该系统的真实运行效果。该系统可以帮助练习者了解自身动作的质量，发现薄弱动作加以练习，最终达到更好的八段锦锻炼效果。

第六章　总结与展望

本文围绕着八段锦动作质量评估任务，依照数据集建立、算法设计和应用系统的行文逻辑对本文重点部分进行介绍。

目前国内对八段锦动作质量评估的研究较少，缺少相应的数据集。本文通过在社交媒体网站上爬取八段锦视频并使用姿态估计算法提取其中的骨骼，建立了第一个同时对动作片段起止时间和质量进行标注的八段锦动作质量评估数据集。为了方便标注，本文还开发了辅助标注系统，该系统同样适用于类似任务的标注需求。本文建立的数据集总时长约 36 小时，为后续的训练奠定了基础。

当前动作质量评估任务的研究主要集中在网络结构的设计上，而对损失函数的研究较少。受到人类对事物的认知规律启发，本文提出了利用粗粒度标签的有监督对比学习损失 SupConWC 和 RnCWC，并分别在八段锦动作质量评估数据集及 CIFAR-100、Stanford Cars 和 Stanford Online Products 等分类任务数据集及 AgeDB 等回归任务数据集上进行了对比实验，取得了优异的表现。

在前面研究的基础上，本文开发了八段锦动作质量评估系统，该系统可以对用户的练习动作打分，帮助用户了解自身的薄弱动作并加以针对性的练习，最终提高用户的八段锦水平，达到更好的锻炼效果。

本文的研究仍有几点可以继续深入：

- 在数据集的构建方面，本文选取的都是正面视角遮挡较少或无遮挡的八段锦视频，但在实际情况中，由于摄像头的摆放位置及锻炼空间的限制，画面中的人物的视角可能不同，身体的某些部分也可能被遮挡，本文并没有对这些情况进行特别的处理，在未来的研究中可以收集相应的样本，并提出解决方案。
- 本文模型训练仅使用了原始视频姿态估计后得到的骨骼信息，而原始

视频中其实还有大量的额外信息，在未来可以以 RGB 视频加骨骼信息多模态的方式训练动作质量评估模型，进一步提升模型的表现。

- 由于未对模型进行蒸馏和量化等操作，本文开发的系统后端对运行配置要求较高，不利于高效推理，在未来可以针对性地优化，甚至可以将部分运算转移到边缘端设备。
- 在未来可以更细粒度地标注数据集，例如标注每个样本动作的错误之处，改进模型以实现指出练习者具体不足的功能。

参考文献

[1] 张韫哲. 八段锦促进大学生健康睡眠的应用研究[D]. 首都体育学院, 2023.

[2] 张鹏. 八段锦对大学生颈肩健康的影响研究[D]. 牡丹江师范学院, 2023.

[3] 余豫敏. 八段锦结合 PNF 对老年膝骨关节炎患者步行能力干预效果的研究[D]. 辽宁师范大学, 2023.

[4] 齐永昊. 健身气功八段锦与六字诀对中老年 T2DM 患者的干预效果研究[D]. 西安体育学院, 2023.

[5] 吴鹏, 王晓芬, 赵林梁, 等. 运动模式评估在蹦床运动员中的应用[J]. 湖北体育科技, 2019, 38(11): 1010-1012+1017.

[6] 马俊阳, 朱厚伟, 潘慧炬. 基于功能性运动筛查的职业排球运动员体能训练评估研究[C]//第十八届全国运动生物力学学术交流大会（CABS 2016）论文集. 2016: 137.

[7] 冯晨, 刘光曹. 面向体育教学的高效动作质量评估算法[J]. 福建电脑, 2024, 40(01): 27-32.

[8] 艾新龑. 基于深度学习的跆拳道智慧教学算法与应用[D]. 河南师范大学, 2022.

[9] 马烨星. 基于 Kinect 的康复训练运动动作质量评估系统[D]. 中国科学院大学 (中国科学院深圳先进技术研究院), 2021.

[10] GORDON A S. Automated video assessment of human performance[C]// Proceedings of AI-ED: vol. 2. 1995.

[11] ILG W, MEZGER J, GIESE M. Estimation of skill levels in sports based on hierarchical spatio-temporal correspondences[C]//Pattern Recognition: 25th DAGM Symposium, Magdeburg, Germany, September 10-12, 2003. Proceedings 25. 2003: 523-531.

[12] ÇELIKTUTAN O, AKGUL C B, WOLF C, et al. Graph-based analysis of physical exercise actions[C]//Proceedings of the 1st ACM International Workshop on Multimedia Indexing and Information Retrieval for Healthcare. 2013: 23-32.

[13] PIRSIAVASH H, VONDRICK C, TORRALBA A. Assessing the quality of actions[C]//Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13. 2014: 556-571.

[14] LECUN Y, BOSER B, DENKER J S, et al. Backpropagation applied to handwritten zip code recognition[J]. Neural Computation, 1989, 1(4): 541-551.

[15] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[C]//Advances in Neural Information Processing Systems: vol. 25. 2012.

[16] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[17] LI Y, CHAI X, CHEN X. End-to-end learning for action quality assessment[C]//Pacific Rim Conference on Multimedia. 2018: 125-134.

[18] TRAN D, BOURDEV L, FERGUS R, et al. Learning spatiotemporal features with 3d convolutional networks[C]//Proceedings of the IEEE International Conference on Computer Vision. 2015: 4489-4497.

[19] GAO J, ZHENG W S, PAN J H, et al. An asymmetric modeling for action assessment[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16. 2020: 222-238.

[20] DONG L J, ZHANG H B, SHI Q, et al. Learning and fusing multiple hidden substages for action quality assessment[J]. Knowledge-Based Systems, 2021, 229: 107388.

[21] WANG S, YANG D, ZHAI P, et al. Tsa-net: Tube self-attention network for action quality assessment[C]//Proceedings of the 29th ACM International Conference on Multimedia. 2021: 4902-4910.

[22] ZHANG Z. Microsoft kinect sensor and its effect[J]. IEEE Multimedia, 2012, 19(2): 4-10.

[23] LI H, LEI Q, ZHANG H, et al. Skeleton-based deep pose feature learning for action quality assessment on figure skating videos[J]. Journal of Visual Communication and Image Representation, 2022, 89: 103625.

[24] PAN J H, GAO J, ZHENG W S. Action assessment by joint relation graphs[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 6331-6340.

[25] PAN J H, GAO J, ZHENG W S. Adaptive action assessment[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(12): 8779-8795.

[26] YAN S, XIONG Y, LIN D. Spatial temporal graph convolutional networks for skeleton-based action recognition[C]//Proceedings of the AAAI Conference on Artificial Intelligence: vol. 32: 1. 2018.

[27] BRUCE X, LIU Y, CHAN K C. Skeleton-based detection of abnormalities in human actions using graph convolutional networks[C]//2020 Second International Conference on Transdisciplinary AI (TransAI). 2020: 131-137.

[28] BRUCE X, LIU Y, CHAN K C, et al. Skeleton-based human action evaluation using graph convolutional network for monitoring Alzheimer's progression[J]. Pattern Recognition, 2021, 119: 108095.

[29] YU B X, LIU Y, ZHANG X, et al. EGCN: An ensemble-based learning framework for exploring effective skeleton-based rehabilitation exercise assessment[C]//Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22. 2022: 3681-3687.

[30] LEI Q, LI H, ZHANG H, et al. Multi-skeleton structures graph convolutional network for action quality assessment in long videos[J]. Applied Intelligence, 2023, 53(19): 21692-21705.

[31] LI M, ZHANG H B, LEI Q, et al. Pairwise contrastive learning network for action quality assessment[C]//European Conference on Computer Vision. 2022: 457-473.

[32] XU J, RAO Y, YU X, et al. Finediving: A fine-grained dataset for procedure-aware action quality assessment[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 2949-2958.

[33] WANG X, LI J, HU H. Skeleton-Based action quality assessment via partially connected LSTM with triplet losses[C]//Chinese Conference on Pattern Recognition and Computer Vision (PRCV). 2022: 220-232.

[34] SCHROFF F, KALENICHENKO D, PHILBIN J. Facenet: A unified embedding for face recognition and clustering[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 815-823.

[35] 戴志豪. 基于图卷积神经网络的太极拳动作识别与评价[D]. 电子科技大学, 2022.

[36] 叶倩文. 基于时空图卷积的太极拳动作质量评估研究[D]. 西安工业大学, 2023.

[37] 叶倩文, 肖秦琨, 李梦茹. 基于 ST-GCN 的形体动作质量评估算法分析[J]. 集成电路应用, 2023, 40(01): 98-99.

[38] 陈静. 面向慢阻肺患者的肺功能评估模型与康复训练动作智能评估算法研究[D]. 中国科学技术大学, 2023.

[39] GUTMANN M, HYVÄRINEN A. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models[C]//Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. 2010: 297-304.

[40] HE K, FAN H, WU Y, et al. Momentum contrast for unsupervised visual representation learning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 9729-9738.

[41] CHEN T, KORNBLITH S, NOROUZI M, et al. A simple framework for contrastive learning of visual representations[C]//International Conference on Machine Learning. 2020: 1597-1607.

[42] BRUNA J, ZAREMBA W, SZLAM A, et al. Spectral networks and locally connected networks on graphs[J]. ArXiv preprint arXiv:1312.6203, 2013.

[43] DEFFERRARD M, BRESSON X, VANDERGHEYNST P. Convolutional neural networks on graphs with fast localized spectral filtering[C]//Advances in Neural Information Processing Systems: vol. 29. 2016.

[44] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[J]. ArXiv preprint arXiv:1609.02907, 2016.

[45] MÜLLER M. Dynamic time warping[J]. Information Retrieval for Music and Motion, 2007: 69-84.

[46] FISHER W D. On grouping for maximum homogeneity[J]. Journal of the American Statistical Association, 1958, 53(284): 789-798.

[47] ESTER M, KRIEGEL H P, SANDER J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise[C]//KDD: vol. 96: 34. 1996: 226-231.

[48] MACQUEEN J, et al. Some methods for classification and analysis of multivariate observations[C]//Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability: vol. 1: 14. 1967: 281-297.

[49] DE JONG K A. An analysis of the behavior of a class of genetic adaptive systems.[M]. University of Michigan, 1975.

[50] HOLLAND J H. Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence[M]. MIT press, 1992.

[51] CAO Z, HIDALGO MARTINEZ 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.

[52] SIMON T, JOO H, MATTHEWS I, et al. Hand keypoint detection in single images using multiview bootstrapping[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2017.

[53] CAO Z, SIMON T, WEI S E, et al. Realtime multi-person 2D pose estimation using part affinity fields[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2017.

[54] WEI S E, RAMAKRISHNA V, KANADE T, et al. Convolutional pose machines[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2016.

[55] KHOSLA P, TETERWAK P, WANG C, et al. Supervised contrastive learning[C]//Advances in Neural Information Processing Systems: vol. 33. 2020: 18661-18673.

[56] LU C, ZOU Y. Using coarse label constraint for fine-grained visual classification[C]//MultiMedia Modeling: 25th International Conference, MMM 2019, Thessaloniki, Greece, January 8–11, 2019, Proceedings, Part II 25. 2019: 266-277.

[57] TOUVRON H, SABLAYROLLES A, DOUZE M, et al. Grafit: Learning fine-grained image representations with coarse labels[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 874-884.

[58] FENG C, PATRAS I. MaskCon: Masked contrastive learning for coarse-labelled dataset[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 19913-19922.

[59] SHI L, ZHANG Y, CHENG J, et al. Two-stream adaptive graph convolutional networks for skeleton-based action recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 12026-12035.

[60] SHI L, ZHANG Y, CHENG J, et al. Skeleton-based action recognition with multi-stream adaptive graph convolutional networks[J]. IEEE Transactions on Image Processing, 2020, 29: 9532-9545.

[61] ZHA K, CAO P, SON J, et al. Rank-N-Contrast: Learning continuous representations for regression[C]//Advances in Neural Information Processing Systems: vol. 36. 2023: 17882-17903.

[62] MOSCHOGLOU S, PAPAIOANNOU A, SAGONAS C, et al. Agedb: The first manually collected, in-the-wild age database[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2017: 51-59.

[63] CARON M, TOUVRON H, MISRA I, et al. Emerging properties in self-supervised vision transformers[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 9650-9660.

[64] YANG Y, ZHA K, CHEN Y, et al. Delving into deep imbalanced regression[C]//International Conference on Machine Learning. 2021: 11842-11851.

[65] GONG Y, MORI G, TUNG F. Ranksim: Ranking similarity regularization for deep imbalanced regression[J]. ArXiv preprint arXiv:2205.15236, 2022.

[66] ZHANG S, YANG L, MI M B, et al. Improving deep regression with ordinal entropy[J]. ArXiv preprint arXiv:2301.08915, 2023.

[67] VAN DER MAATEN L, HINTON G. Visualizing data using t-SNE.[J]. Journal of Machine Learning Research, 2008, 9(11).

致　　谢

三年的时光匆匆而逝，我还记得入学时我独自一人拖着行李，下飞机后又坐了两个小时地铁才到学校。那是晴朗的一天，小蓝鲸气球飘荡在空中。如今，我却要离开了，心中不禁泛起了许多感慨。

首先，我要衷心感谢我的导师申富饶老师。在这三年中，申老师不仅在学术上给予我指导，还在个人成长方面提供了宝贵的建议。尽管申老师工作繁忙，但他每周都会抽出时间与我们一对一交流，细心了解我们的科研进展，解答我们的疑问，鼓励我们不断前行。我的论文选题正是在申老师的启发下逐步明确和完善的，我深表感激。

其次，我要感谢我的家人和朋友们。感谢你们始终如一的支持和鼓励，每当我遇到困难时，你们总是第一时间站在我身后，为我加油打气。特别是在我疲惫不堪时，你们的关心和鼓励使我重新振作，继续前行。你们偶尔的打岔和调侃，也为我的繁忙生活带来了不少欢乐，让我能够在紧张的科研工作中找到一丝放松和平衡。

向所有给我带来小小欢乐的事物致敬！它们或许微不足道，但正是这些细碎的幸福，组成了我这三年中的美好回忆。

最后，再次感谢所有在我这段求学之路上给予我帮助和支持的人，感谢你们让我在这里度过了一段充实而难忘的时光。我将带着这份感恩和收获，继续前行，迎接未来的挑战。

简历与科研成果

基本信息

张耕，男，汉族，1998 年 7 月出生，山东省莱州人。

教育背景

2021 年 9 月—2024 年 6 月 南京大学人工智能学院 硕士

2017 年 9 月—2021 年 6 月 电子科技大学计算机科学与工程学院 本科

攻读硕士学位期间完成的学术成果

1. Suorong Yang, **Geng Zhang**, Jian Zhao, Furao Shen. A simple geometric-aware indoor positioning interpolation algorithm based on manifold learning. arXiv preprint arXiv:2311.15583, 2023.

攻读硕士学位期间完成的专利成果

1. 申富饶, **张耕**, 赵健. “一种基于帧间差模波形图的动作质量评估数据集辅助标注系统”（202410304715.X）

攻读硕士学位期间参与的科研课题

1. 科技部重大项目 “基于神经可塑性的脉冲网络高效学习机制与类脑智能系统”（参与课题年限 2021 年 9 月—2024 年 6 月），负责神经网络模型相关研究。
2. 国家电网 “基于多维巡检影像匹配和对比技术的变电设备缺陷分析技术研究”（参与课题年限 2021 年 9 月—2022 年 12 月），负责目标检测相关研究。