

基于音视频信息融合与 Self-Attention-DSC-CNN6 网络的鲈鱼摄食强度分类方法

李道亮^{1,2} 李万超³ 杜壮壮^{1,2}

(1. 中国农业大学信息与电气工程学院, 北京 100083; 2. 国家数字渔业创新中心, 北京 100083;
3. 天津农学院计算机与信息工程学院, 天津 300392)

摘要: 摄食强度识别分类是实现水产养殖精准投喂的重要环节。现有的投喂方式存在过度依赖人工经验判断、投喂量不精确、饲料浪费严重等问题。基于多模态融合的鱼类摄食程度分类能够综合不同类型的数据(如:视频、声音和水质参数),为鱼群的投喂提供更加全面精准的决策依据。因此,提出了一种融合视频和音频数据的多模态融合框架,旨在提升鲈鱼摄食强度分类性能。将预处理后的 Mel 频谱图(Mel Spectrogram)和视频帧图像分别输入到 Self-Attention-DSC-CNN6(Self-attention-depthwise separable convolution-CNN6)优化模型进行高层次的特征提取,并将提取的特征进一步拼接融合,最后将拼接后的特征经分类器分类。针对 Self-Attention-DSC-CNN6 优化模型,基于 CNN6 算法进行了改进,将传统卷积层替换为深度可分离卷积(Depthwise separable convolution, DSC)来达到减少计算复杂度的效果,并引入 Self-Attention 注意力机制以增强特征提取能力。实验结果显示,本文所提出的多模态融合框架鲈鱼摄食强度分类准确率达到 90.24%,模型可以有效利用不同数据源信息,提升了对复杂环境中鱼群行为的理解,增强了模型决策能力,确保了投喂策略的及时性与准确性,从而有效减少了饲料浪费。

关键词: 鲈鱼; 摄食强度分类; 多模态融合; Self-Attention-DSC-CNN6

中图分类号: S951.2

文献标识码: A

文章编号: 1000-1298(2025)01-0016-09

OSID:



Classification Method of Feeding Intensity of Sea Bass Based on Self-Attention-DSC-CNN6 and Multi-modal Fusion

LI Daoliang^{1,2} LI Wanchao³ DU Zhuangzhuang^{1,2}

(1. College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China

2. National Innovation Center for Digital Fishery, Beijing 100083, China

3. College of Computer and Information Engineering, Tianjin Agricultural University, Tianjin 300392, China)

Abstract: Feeding intensity recognition and classification is an important link to realize accurate feeding in aquaculture. Existing feeding methods have problems such as over-reliance on manual experience judgment, imprecise feeding amount, and serious feed waste. Fish feeding degree classification based on multi-modal fusion can synthesize different types of data (e. g., video, sound, and water quality parameters) to provide a more comprehensive and accurate decision basis for fish feeding. Therefore, a multi-modal fusion framework that integrated video and audio data was proposed with the aim of improving the performance of sea bass feeding intensity classification. The preprocessed Mel Spectrogram (Mel) and video frame images were input into the self-attention-depthwise separable convolution-CNN6 (Self-Attention-DSC-CNN6) optimization model for high-level feature extraction, respectively, and the extracted features were further spliced and fused, and finally the spliced features were classified by a classifier. The Self-Attention-DSC-CNN6 optimization model was improved based on the CNN6 algorithm by replacing the traditional convolutional layers with depthwise separable convolution (DSC) to reduce the computational complexity, and the Self-Attention mechanism was introduced to enhance the feature extraction capability. The experimental results showed that the multi-modal fusion framework proposed achieved an accuracy of 90.24% in sea bass feeding intensity classification, and the model can

收稿日期: 2024-11-25 修回日期: 2024-12-10

基金项目: 国家重点研发计划项目(2022YFD2001703)

作者简介: 李道亮(1971—),男,教授,博士生导师,主要从事农业智能信息处理与农业农村信息化研究,E-mail: dliangl@cau.edu.cn

effectively utilize the information from different data sources to improve the understanding of fish behavior in complex environments, enhance the decision-making ability of the model, and ensure the timeliness and accuracy of the feeding strategy, thus effectively reducing the waste of feed. This not only provided strong technical support for the intelligent management of aquaculture, but also laid the foundation for the development of intelligent feeding system.

Key words: sea bass; classification of feeding intensity; multi-modal fusion; Self - Attention - DSC - CNN6

0 引言

水产养殖在应对全球水产品需求的持续增长及保障食物供应可持续性中发挥着核心作用^[1-2]。随着养殖规模的集约化与高效监测、管理系统的快速发展,精准掌握鱼类摄食强度变得日益重要。然而,现有投喂方法大多基于人工经验判断或固定投喂时间,缺乏对鱼类实际摄食需求的动态响应。这种投喂方式往往导致饲料投喂过量或不足,不仅浪费资源,还可能影响鱼类的生长速度和健康状况,甚至加剧水质污染。因此,开展基于多模态融合技术的鱼类摄食强度研究可以实现更精准的投喂管理,从而优化投喂策略,提高饲料利用率,减少环境污染,进一步推动水产养殖业的可持续发展。目前,多数水产养殖监测手段仍局限于单一模态数据驱动决策,如声学监测、视觉观察和传感器数据收集。声学监测受养殖环境噪声影响,视觉观察主观性强且受水体浑浊和环境光线影响,同时,由于鱼群摄食过程中水质参数变化缓慢,传感器监测同样难以精准有效地评估摄食过程。此外,由于水产养殖环境是一个复杂且多变的动态系统,涉及众多相互影响的变量,单一监测模式难以全面捕捉鱼类摄食行为的细节^[3]。相比之下,多模态融合技术通过整合并分析来自不同数据源的信息,为评估鱼类摄食强度提供了更为全面和深入的视角。该技术充分利用多种监测手段的优势,能够更准确地掌握水产养殖环境中鱼类的摄食习性。

在水产养殖领域中,基于单一模态的鱼类摄食强度已取得了广泛研究。水质参数(溶解氧含量,温度,氨氮含量,pH 值等)传感器不仅可以提供实时数据,而且适用于不同的养殖环境中(包括淡水和海水),具有较高的客观性和准确性,能够提供更可靠的数据支持^[4]。水质参数传感器虽然能够提供实时监测的文字或数值型数据,但它只能间接反映鱼类摄食强度,难以直接捕捉鱼类行为的动态特征和具体摄食动作。相较于依靠水质参数判断方法,视觉技术作为一种非侵入性且低成本的监测工具,通过图像和视频数据能够直接记录鱼类摄食行为的动态过程,因此在水产养殖监测中得到了广泛

应用^[5-7]。尽管计算机视觉方法在鱼类识别与监测中展现出了一定的潜力,但受制于光照条件、水体浊度和水下折射等影响,限制了其在复杂水下环境中的应用效果^[8-9]。相比之下,声学技术利用声音数据提供了直接且动态的鱼类摄食强度信息,能够穿透水体减少水下折射干扰,较视觉技术避免了光线条件的限制^[10-11]。尽管声学技术在一定程度上弥补了视觉技术的不足,但其也面临设备成本较高、数据处理复杂以及背景噪声干扰等问题。水听器捕获的声学信号通常为复杂的混合信号,背景噪声与水生生物发声的重叠可能导致鱼类摄食强度评估的误差。因此,基于单一模态的研究方法虽然已经在水产养殖监测中取得了一定的进展,但这些方法仍存在一定的局限性,而且单一模态往往无法全面反映鱼类的生理状态与行为模式。

与单一模态技术相比,多模态融合技术在水产养殖中展现出了诸多优势,其通过整合多种数据类型,例如图像、声音、文本等,从而更全面地了解鱼类生长情况^[12]。每种模态都有助于揭示环境因素和摄食行为的独特视角。通过有效汇总和分析这些多模态数据,养殖户能够实时掌握养殖环境的多维信息,为精准调控水产养殖环境提供可靠的科学依据。在两种模态融合方面,DU 等^[13]提出了一种基于多特征(Mel 频谱图、STFT 特征图和 CQT 特征图)和 LC - GhostNet 轻量级网络的融合方法,用于斑石鲷摄食强度检测。实验结果表明融合特征图准确率为 97.941%,分别比单一特征 Mel、STFT、CQT 高 4.053、7.207、3.003 个百分点。DU 等^[14]又提出了一种 3 种模态数据的融合方法,即基于水听器、光学和图像声呐数据构造的多模态融合方案,测试结果显示,融合多模态信息的准确率为 99.26%,与单一模态(音频、视频和声学数据集)的最佳结果相比,准确率分别提高 12.80、13.77、2.86 个百分点。此外,胡学龙等^[15]提出了一种基于文本、声音、视觉 3 种模态的多模态融合鱼类摄食强度识别算法 Fish - MulT。实验结果表明,相较于 Multimodal transformer 算法的 93.30%,Fish - MulT 算法的鱼类摄食强度识别准确率提高到 95.36%。

尽管多模态融合技术在相关领域取得了显著进

展,但针对鲈鱼摄食强度分类的研究仍存在一定的空白和薄弱环节。胡学龙等^[15]的研究虽然在文本、声音和视觉3种模态的融合方面具有更强的全面性,但其基于水质单一模态的分类准确率不足30%,这表明获取的水质参数等文本模态数据在鱼类摄食强度识别中贡献有限。鉴于此,本文未选择文本模态,而是聚焦于视觉和声学2种模态的数据融合,旨在减少非关键模态可能带来的干扰和噪声。此外,DU等^[13-14]的研究采用了多特征融合方法和轻量化网络,尽管在准确率上表现优异,但其方法未充分利用注意力机制优化不同模态间特征的交互与融合,可能在不同模态特征关系的深度挖掘上存在不足。针对上述问题,本文提出一种基于音视频多模态融合架构的Self-Attention-DSC-CNN6网络模型,通过引入注意力机制强化模态间的信息交互,同时结合深度可分离卷积降低模型复杂度并优化数据融合过程。

1 材料与方法

1.1 数据来源

在广州市南沙区广东省渔业技术推广总站真实工厂化循环水养殖系统中进行实验,如图1所示。实验装置包括循环水养殖系统、音频采集系统和视频采集系统。其中,视频采集系统由一台海康威视彩色摄像机(型号:DS-2SC3Q140MY-TE;分辨率:2560像素×1440像素;帧速率:25 f/s)、一台IEEE802.3af标准POE交换机(型号:DS-3E0105P-E/M)、一台海康威视录像机(型号:DS-7104N-F1)和一台Lecoo显示屏(型号:HX20270FH0)组成。音频采集系统由一个LST-DH01型数字水听器(长沙览声科技有限公司)和一台计算机(戴尔灵越5570;处理器:Intel(R)Core(TM)i5-8250U CPU@1.60 GHz)组成。该系统饲养150条鲈鱼,每条鲈鱼为 (500 ± 50) g。实验过程中溶解氧质量浓度维持在 (7.0 ± 1.0) mg/L,pH值为 8.0 ± 0.5 ,水温维持在 (20 ± 2) °C。在预实验阶段,对鲈鱼实施了为期一周的系统性条件反射训练方案。该训练的核心内容是在预设的时间点和空间位置对鲈鱼进行饵料投喂,旨在诱导其形成特定的条件反应,即习惯于在某一固定区域觅取食物。为保障实验的一致性和精确性,后续的每次投喂过程中均严格遵循相同时间点和空间定位原则。

1.2 数据采集和预处理

1.2.1 数据采集

为确保数据采集的时间一致性,本次实验采用了同步化协议机制。在每次饲喂活动前5 min的精

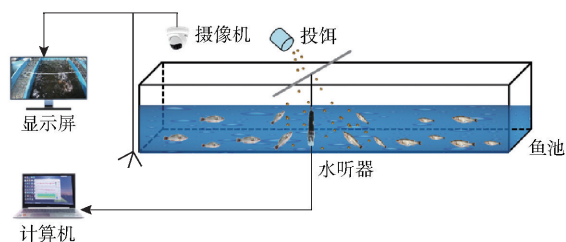


图1 实验系统结构图

Fig. 1 Experimental system structure diagram

确时刻,音频采集系统和视频采集系统被同步激活。5 min 预备期结束后,随即启动饲喂流程,且在此期间,所有相关序列均被详尽记录。在养殖管理环节,严格遵循了“四定”投喂原则,即定时投喂、定点投放、定质饲料、定量供给^[16-18]。投喂频次设定为每日2次,每次的投喂量精确控制在鱼群总体质量的1.5%。投喂作业时间为每日09:00和17:00。

1.2.2 数据预处理

首先使用光学摄像机和水听器同时收集鲈鱼摄食期间的数据,随后利用VS Player和EaseUS Video Editor软件同步对齐视频和音频,并将视频和音频分为“强”、“中”、“弱”、“无”4个级别。摄食强度等级的划分依据为:参考专业水产养殖专家的理论知识;参考专业养殖人员的工作经验;参考相关领域文献所提供的分类规则。其中,徐立鸿等^[19-20]将数据集划分为“强”、“弱”、“无”3个级别;唐宸等^[21]将数据集划分为“强”、“弱”、“未”3个级别;刘世晶等^[22]将数据集划分为“强”、“中”、“弱”3个级别;ZHOU等^[23]将数据集划分为“强”、“中”、“一般”、“弱”4个级别;文献[24-26]将数据集划分为“强”、“中”、“弱”、“无”4个级别。尽管3个级别的摄食强度分类方法具有简单易行、易于理解和应用的明显优势,尤其在摄食强度差异不显著或存在重叠时,可以减少分类误差并提高整体分类准确性。然而,随着水产养殖对精细化管理的需求日益提高,3个级别的分类方法在区分摄食强度细节方面的局限性逐渐显现。相比之下,4个级别的摄食强度分类方法能够更精细地量化鱼类摄食强度,不仅可以揭示数据分布特征,更能细致刻画摄食强度的变化规律。通过这种分类方式,可以有效优化投饵策略,提高养殖效率,减少资源浪费。此外,4个级别的分类方法还能够全面反映摄食行为的连续性变化,为精准投饵和环境调控提供更可靠的依据。因此,综合考虑以上因素,本实验采用4个级别的分类方法。强、中、弱、无的鲈鱼鱼群摄食行为特点分别为:①强:所有鲈鱼快速无规律游动吃食或聚集在同一位置抢食,由于抢食激烈,水面上会激起大量水花,形成明显的波动。②中:大部分鲈鱼无规律游动并

食用饵料,但相较于第一种情况,抢食的激烈程度有所降低,水面上激起的水花也较少,水面波动相对较小。③弱:小部分鲈鱼无规律游动并食用饵料,但动作较为温和,几乎没有激起水花,水面波动极小或大部分鲈鱼逆着水流缓慢游动或者静止且不进食。④无:所有鲈鱼逆着水流缓慢游动或者静止且不进食也不产生水花,水面几乎无波动。

1.2.3 数据集划分

数据集经过科学划分,80%用于训练集,以支持模型学习和参数优化,识别数据中的潜在模式和特征。剩余的 20% 被均等分配为验证集和测试集,各占 10%。验证集用于评估模型在训练过程中的表现,并协助调整超参数,以防止过拟合并提高模型的泛化能力;测试集则用于最终评估模型在未知数据上的性能,确保其有效性和鲁棒性。这种划分方法提供了全面的评估框架,确保模型在实际应用中保持稳定性和准确性,具体数据集划分结果如表 1 所示。在本研究中,音频数据和视频数据在预处理阶段进行了严格的同步对齐,因此音频与视频的样本数量完全一致。这种同步处理方式保证了多模态数据的一致性,每条数据均包含音频和视频两种模态信息,为后续多模态融合分析提供了可靠的数据基础。“弱”等级的样本数量约为其他等级的 1/2,主要原因在于鱼类在“弱”摄食阶段的行为持续时间较短,导致数据采集及划分时该等级的样本数量相对较少,后续将扩大“弱”等级的样本数量。

表 1 数据集划分结果
Tab.1 Division of dataset

数据集	分类				总计
	强	中	弱	无	
训练集	2 717	2 203	1 316	2 399	8 635
测试集	339	275	164	300	1 078
验证集	339	275	164	300	1 078
总计	3 395	2 753	1 644	2 999	10 791

1.3 多模态融合方法

1.3.1 多模态融合架构与工作流程

本文提出了一种基于音频与视频融合的多模态分类系统,其核心在于充分整合音频和视频数据的互补特性,实现对鲈鱼摄食程度的精准分类。该系统中首先将音频波形数据经过 Mel 频谱转换器生成 Mel 频谱图以提取低级音频特征;视频信号通过关键帧提取技术获取低级视觉特征。随后,利用改进的 Self - Attention - DSC - CNN6 模块对音频和视频的低级特征分别进行高级特征提取。针对音频和视频模态特征的异质性,模块引入深度可分离卷积,通过分开处理空间和通道信息,有效提升了对模态内

特征的表征能力;针对多模态信息整合中的关键特征捕捉问题,自注意力机制使模型能够更加精准地聚焦于音频和视频中的重要特征区域,从而提升特征表示质量;此外,通过深度可分离卷积和自注意力机制的协同作用,增强了模态间的互补性,实现了更高效的多模态信息整合。在提取高级特征后,系统通过特征拼接的方式融合音频和视频的特征向量,生成多模态综合特征表示,并输入到 Softmax 分类器中,输出“强”、“中”、“弱”、“无”4 种类别的预测结果。相比传统模型,本文提出的多模态融合架构能够更全面地捕捉音视频模态的互补特性,显著提升分类性能的同时保持了较好的鲁棒性。图 2 展示了该多模态融合分类系统整体工作流程及主要组成部分,包括音频与视频信号预处理、特征提取、特征融合以及最终分类结果。

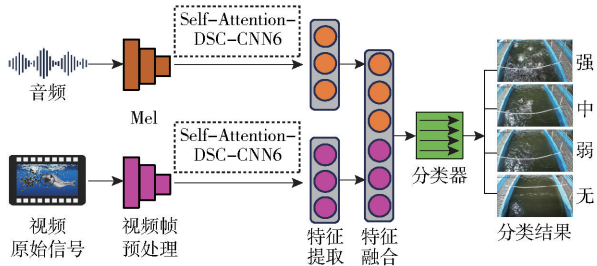


图 2 鲈鱼摄食强度多模态融合框图
Fig.2 Multi-modal fusion framework diagram of sea bass feeding intensity

1.3.2 音频特征提取

鱼类声音信号是非稳态信号,在时域中难以直接观察其特性^[27-28]。因此,为了有效分析鱼类声音特征,选择基于 Mel 的音频特征提取方法。在提取鲈鱼摄食声音特征的过程中,首先对原始声音信号进行预处理,包括去噪以消除环境噪声、预加重以提升高频能量,接着将信号分帧并施加汉明窗以减少频谱泄漏。随后,对每一帧信号进行短时傅里叶变换(STFT)以获取频谱信息,并将这些频谱信息通过设计的 Mel 滤波器组转换为 Mel 频谱。Mel 滤波器组的设计基于将线性频率 f 转换为 Mel 频率 M 的公式: $M = 2595 \lg(1 + f/700)$,其中在 Mel 频率域内等间隔地设置三角形滤波器,覆盖整个频率范围。最后,对 Mel 滤波器组的输出结果进行对数运算,得到对数 Mel 频谱,为后续特征提取提供准确的频谱特征。选择 Mel 频谱图是因为其能够保留鱼类声音关键信息的同时降低特征维度,更完整地提取鱼类细粒度声音信息。本研究声音特征提取过程如图 3 所示。

1.3.3 视频特征提取

基于海康威视摄像机的视频特征提取过程如

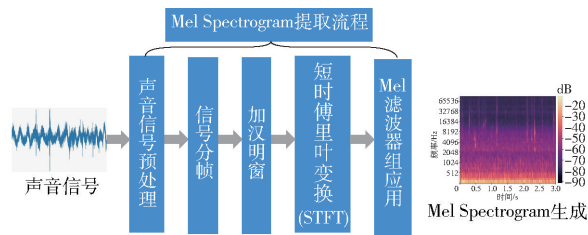


图 3 声音特征提取流程图

Fig. 3 Flowchart of sound feature extraction

图 4 所示。首先,将原始视频输入系统,每段视频时长为 2 s,由 40 帧组成,且每帧分辨率为 2 560 像素 × 1 440 像素。随后,系统进行预处理,具体包括视频帧提取和分辨率缩放。在帧提取过程中,采用下采样技术随机选择一帧,并将其分辨率调整为 224 像素 × 224 像素。此外,在训练阶段结合随机裁剪和颜色增强等数据扩增策略,以提升模型的泛化能力。利用 Self - Attention - DSC - CNN6 模型提取处理后的关键帧所蕴含的高级特征。最终,每个视频帧被转换为一个高维特征向量,进而捕捉视频中的关键信息和模式,为后续的分类任务提供高质量的输入数据。

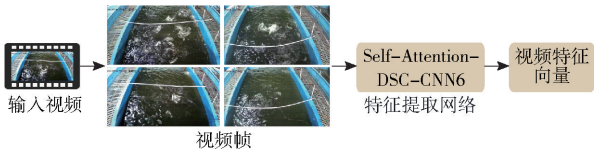


图 4 视频特征提取流程图

Fig. 4 Flowchart of video feature extraction

1.3.4 Self - Attention - DSC - CNN6 模型设计

在鲈鱼摄食程度分类任务中,本文提出了一种基于多模态融合的分类架构,并设计了改进的 Self - Attention - DSC - CNN6 模型作为核心模块。相比传统的 CNN6 模型,Self - Attention - DSC - CNN6 在结构上进行了优化:一方面,采用深度可分离卷积代替传统卷积层,这一设计不仅减轻了模型的计算和存储负担,还提高了模型特征提取能力,尤其在复杂场景下的表现更为突出;另一方面,引入自注意力机制 (Self - Attention),使模型能够动态聚焦于音频和视频数据中的关键特征区域,更高效地提取多模态信

息的深层特征。图 5 用红色字体标出了具体的改进部分。这些优化设计不仅增强了模型对复杂数据的表征能力,还显著提升了音视频模态信息的整合效果,为更精准地实现鲈鱼摄食程度分类提供了有力支持。

1.4 性能评估指标

在评估鲈鱼摄食活动强度分类方法时,采用准确率、精确率、召回率和 F1 分数 4 个核心指标,以深入了解模型性能,优化分类方法,提高分类准确性。

2 实验结果与分析

2.1 实验环境及参数设置

该程序使用 Python 编程语言和 PyTorch 深度学习框架开发。在构建深度学习模型时,采用高度优化且集成化的开发环境。选用 Python 3.9 作为编程语言,使用 PyCharm 作为集成开发环境 (IDE), Anaconda 作为包管理器和虚拟环境工具。所有实验均在搭载 64 位 Windows 11 操作系统的计算机上执行,该计算机配备了强大的硬件配置,包括 13 代 Intel Core i5 - 13400F 处理器 (主频 2.50 GHz)、NVIDIA GeForce RTX 4060 显卡 (数据显存 11 GB) 以及 16.0 GB 的机带 RAM。在训练时,采用 Adam 优化器,初始学习率设为 0.01,随机种子设为 25,迭代次数设为 100,批量大小设为 16。

2.2 实验结果

2.2.1 Self - Attention - DSC - CNN6 模型的可行性

首先验证 Self - Attention - DSC - CNN6 模型在基于视频和音频的多模态融合架构中的可行性。从图 6 中可以看出,随着迭代次数的增加,损失值和准确率均呈现出良好的优化趋势。在训练初期 (前 20 次迭代),损失值快速下降,表明模型学习到了显著特征;随后 (20 ~ 60 次迭代),损失值下降速度趋缓但仍在稳定优化;训练后期 (60 次迭代后),损失值下降趋于平稳,并伴有小幅振荡,反映模型在局部区域进行精细调整。同时,准确率在训练初期迅速上升,表明模型有效捕获了核心特征;在后续训练中,

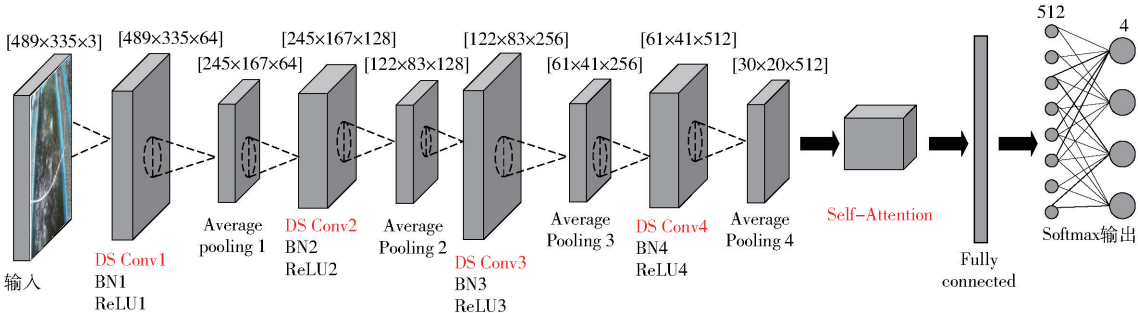


图 5 Self - Attention - DSC - CNN6 模型结构图

Fig. 5 Self - Attention - DSC - CNN6 model structure diagram

准确率逐步提升并接近饱和状态,说明模型对训练数据的拟合效果不断提高,分类能力逐步增强。整体来看,损失值持续下降和准确率稳步上升相互印证,表明当前算法结构合理,适用于鲈鱼摄食强度分类任务。

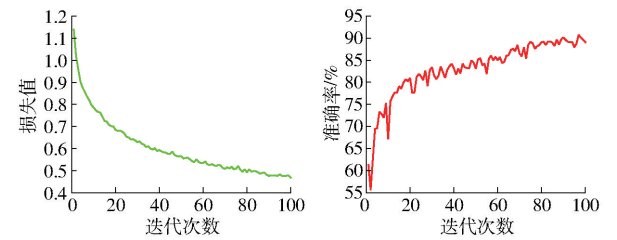


图6 损失值和准确率变化曲线

Fig.6 Change curves of loss value and accuracy rate

2.2.2 多模态融合架构中不同算法对比

为了评估本文提出的基于音频和视频的多模态融合架构性能,在相同的鲈鱼摄食强度数据集上与其他主流多模态分类模型(如 ResNet18、ResNet50 和 MobileNetV3 - Small)进行了对比实验,所有模型均使用相同的数据集、训练参数和迭代次数,实验结果如图 7 所示。从结果可以看出,MobileNetV3 - Small 作为轻量化网络,虽然计算效率较高,但由于其模型复杂度较低,融合多模态信息的能力有限,因此分类准确率最低;ResNet18 和 ResNet50 在多模态特征提取能力上有所提升,尤其是 ResNet50 通过更深的网络结构表现出更高的分类性能。然而,与本文提出的多模态融合架构相比,这些算法在多模态信息的深度融合与关键特征的提取上仍存在一定局限性。相比之下,本文提出的多模态融合架构充分利用了音频和视频两种模态的互补特性,通过优化的多模态特征融合策略,更全面地捕捉鱼类摄食行为的多维特征。在架构设计中,引入 Self - Attention 注意力机制增强了对关键特征的聚焦能力,并结合深度可分离卷积提高了特征提取的表现力,从而进一步提升了分类性能。实验结果表明,本文提出的多模态融合架构在分类性能上显著优于对比模型,验证了其在复杂行为分析任务中的优势,也展现了其在水产养殖智能化监测中的应用潜力。

2.2.3 注意力机制消融实验

为了验证本文提出的多模态融合框架中自注意力机制(Self - Attention)在提升模型性能上的优越性,本文设计了一系列消融实验,基于相同的数据集和 CNN6 算法,分别引入多头注意力机制(Multi-head Self - Attention)、通道注意力(Channel - Attention)、SEBlock 模块(Squeeze-and-Excitation Block)与本文所使用的自注意力机制进行了单独训练和测试,实验结果见表 2。从表 2 可以看出,基于

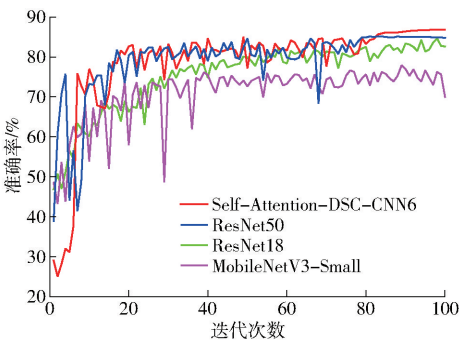


图7 不同算法准确率变化曲线

Fig.7 Accuracy change curves of different algorithms

CNN6 的基础模型在引入注意力机制后,各项性能指标均有所提升。其中,CNN6 + Self - Attention 准确率最高,为 88.60%,与引入的多头注意力机制、通道注意力及 SEBlock 模块相比,分别提高 0.50、1.46、0.15 个百分点。同时,进一步结合深度可分离卷积的 Self - Attention - DSC - CNN6 模型表现最优,其准确率、精确率、召回率和 F1 分数分别达到 90.24%、91.17%、89.14% 和 90.45%,显著高于其他模型。这表明,本文提出的自注意力机制及其与深度可分离卷积的融合能够更有效地提取特征并提升模型性能。

表2 不同注意力机制性能对比

Tab.2 Performance comparison of different

模型	attention mechanisms				%
	准确率	精确率	召回率	F1 分数	
CNN6	87.78	87.21	87.79	87.44	
CNN6 + Multi-head Self - Attention	88.10	87.51	88.46	87.85	
CNN6 + Channel - Attention	87.14	88.10	87.69	87.74	
CNN6 + SEBlock	88.45	88.06	88.55	88.21	
CNN6 + Self - Attention	88.60	88.04	88.49	88.21	
Self - Attention - DSC - CNN6	90.24	91.17	89.14	90.45	

图 8 为本文提出的 Self - Attention - DSC - CNN6 模型在鲈鱼摄食程度分类任务中 4 个关键评估指标(准确率、精确率、召回率和 F1 分数)的训练过程及表现。实验结果表明,该模型在迭代次数内 4 项指标均逐步提升,并在接近 100 次迭代时趋于稳定,显示了良好的收敛性和鲁棒性。具体而言,各指标间保持了显著的平衡,模型不仅在准确率上表现卓越,同时在精确率、召回率和 F1 分数上也展现出强劲的性能。这一结果进一步验证了 CNN6 模型在整合 Self - Attention 与深度可分离卷积后对特征提取与分类任务的增强能力。

图 9 通过混淆矩阵直观展示了多模态融合模型在鲈鱼摄食强度分类任务中的性能。混淆矩阵的标签分别对应“0”(无摄食)、“1”(弱摄食)、“2”(中等

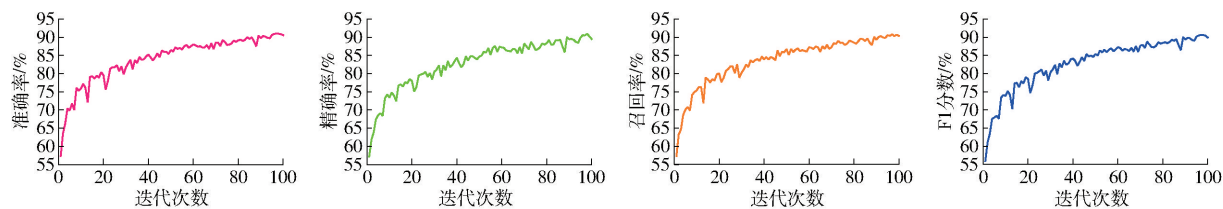


图8 优化算法在验证集上4个评估指标变化曲线

Fig. 8 Optimization algorithm change curves for four evaluation metrics on validation set

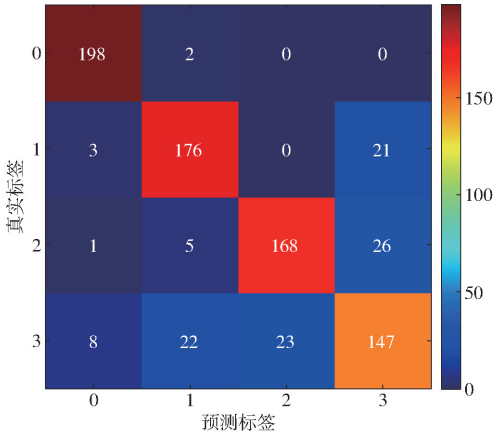


图9 混淆矩阵

Fig. 9 Confusion matrix

摄食)和“3”(强摄食)。对角线上的值表示正确分类的样本数量,颜色的深浅反映了分类准确性高低。可以看到,模型对“0”(无摄食)和“3”(强摄食)的识别表现较好,其准确率分别为99%和81%。这是因为“0”(无摄食)和“3”(强摄食)的特征更加显著:“无摄食”时鱼类活动明显减少,水面波动基本为零;“强摄食”时则伴随剧烈水面波动和明显的鱼体跃出水面现象,音频和视频模态的信息更为明显。然而,“1”(弱摄食)和“2”(中等摄食)的准确率相对较低,误判率较高,特别是“弱摄食”部分被误判为“中等摄食”。这种误判主要源于“1”(弱摄食)和“2”(中等摄食)摄食行为特征的相似性。在弱摄食阶段,鱼类摄食动作较为缓慢且不连贯,水面波动变化差距较小且不易察觉,但仍存在一定的摄食声音特征;而在中等摄食阶段,鱼类的摄食动作较为活跃,但水面波动和声音特征介于弱摄食与强摄食之间,呈现一定的模糊性。因此,多模态融合通过综合多种模态信息使得模型得以从多个维度理解摄食行为,减小单一模态特征中可能存在的信息相似度高或模糊的影响。这种信息的互补性在弱摄食和中等摄食难以区分的情况下显得尤为重要,弥补了单一模态的不足。通过实验结果可以验证多模态融合的有效性,显著提升了模型在复杂行为分类任务中的鲁棒性和准确性。

综上所述,通过损失值与准确率的分析、注意力机制的消融实验、多模态融合架构中不同算法的对

比实验,以及所提算法在4项评估指标与混淆矩阵结果中的表现,充分验证了本文所提的多模态融合架构在鲈鱼摄食程度分类任务中的有效性。

2.3 多模态融合与单一模态对比

通过对比分析,揭示多模态融合架构在提升分类精度及增强模型泛化能力方面的重要意义。为实现这一目标,本研究基于已训练完成的多模态融合架构模型,分别以仅视频数据、仅音频数据以及两者融合的数据作为输入,利用性能评估指标对模型进行全面评估。通过对比实验结果,可以清晰地看到多模态数据融合在分类任务中的优越性。与单一模态输入相比,多模态融合显著提高了模型在复杂行为分类场景中的性能。综上所述,多模态融合技术为复杂行为分类提供了新的解决方案,其在提升模型性能和处理复杂任务方面展现出独特的优势。通过多模态信息的引入,多模态融合架构能够在行为分类任务中更好地提取有效特征,体现出多模态技术在摄食程度分类领域的巨大潜力。

2.3.1 准确率对比

图10展示了在相同算法下,单一模态(音频或视频)与多模态融合(音频+视频)的分类准确率变化趋势。实验结果表明,虽然单一模态模型在各自任务中表现较为出色,但其整体分类准确率和预测能力均低于多模态融合模型。具体而言,音频数据由于容易受到采集环境噪声的干扰,其分类准确率呈较大波动;相比之下,视频数据质量和预测准确性均明显优于音频。然而,多模态融合架构通过整合

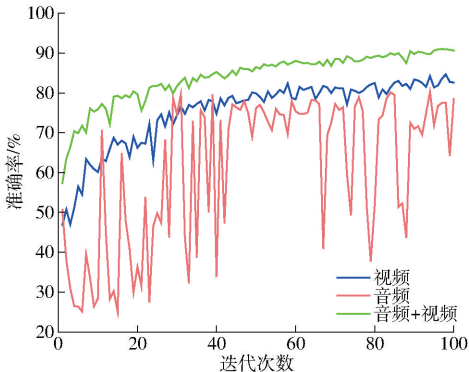


图10 多模态融合与单一模态准确率变化曲线

Fig. 10 Multi-modal fusion and single-modal accuracy change curves

视频与音频两种模态的数据,充分发挥了模态间信息的互补性和协同性,显著提升了分类准确率,达到了更高的性能水平。这一结果表明,多模态信息融合能够更全面地捕捉数据特征,有效增强模型的分类能力和稳健性,为实现更精确的分类任务提供了可靠的技术支持。

2.3.2 其他性能指标对比

对比了多模态融合与单一模态在精确率、召回率和 F1 分数方面的性能。从表 3 可以看出,基于音频和视频的多模态融合架构在所有评估指标上均优于单一模态模型。首先,就准确率而言,多模态融合模型达到 90.24%,这一结果相较于视频模态的 82.62% 和音频模态的 70.00%,分别提升 7.62、20.24 个百分点。这表明,通过整合来自视频和音频两种模态的信息,模型能够更准确地识别和理解输入数据,减少了误判的可能性。在精确率方面,多模态融合模型同样表现出色,达到 91.17%,远超视频模态的 83.09% 和音频模态的 70.74%,由此进一步证明了本文提出的多模态融合方法在识别任务中的准确性和可靠性。在召回率方面,多模态融合模型也达到 89.14%,高于视频模态和音频模态的 82.62% 和 70.00%。最后,多模态融合模型 F1 分数为 90.45%,同样显著优于单一模态的 82.51% 和 69.43%。这些数据进一步证明了多模态融合模型在综合利用不同模态信息上的优势,能够更全面地

捕捉和融合各种信号,提高模型的鲁棒性和适应性。

3 结束语

提出了一种基于音频和视频的多模态融合架构,旨在实现鲈鱼摄食程度的精准分类。传统养殖管理方法多依赖人工观察与经验判断,难以实现高效、精准的投喂。为此设计了一种融合声学 with 视觉信息的模型,以增强对摄食行为的监测与分析能力。该多模态融合架构将音频和视频数据视为同等重要的信息源,通过特征拼接策略有效整合两种模态信息,充分发挥其在特征提取与分类中的互补优势。音频信息提供了鲈鱼摄食过程中的声音特征,如水面的波动声和吞咽饲料的声音,而视频信息则捕捉鲈鱼的动态表现,包括游动行为、摄食动作以及与周围环境的互动等视觉特征。这种多模态信息输入使模型能够在复杂的水产养殖环境中准确识别与分类鱼类的摄食情况。在基于音频和视频的多模态融合架构中,本文通过将传统卷积层替换为深度可分离卷积和引入自注意力机制来提取更为高级的特征。深度可分离卷积有助于有效捕捉音频和视频中的关键特征;而自注意力机制则增强了模型对重要信息区域的聚焦能力,从而提升了特征的表示质量。实验结果表明,该架构分类准确率达到 90.24%,较单一音频和视频模态模型分别提升 20.24、7.62 个百分点,显著提升了分类性能。在与主流分类模型如 ResNet18、ResNet50 及轻量化模型 MobileNetV3 - Small 进行比较时,Self - Attention - DSC - CNN6 同样展现出了卓越的性能。基于音频和视频的多模态融合架构不仅为鲈鱼摄食行为的精准分类提供了有效解决方案,也为智能化养殖技术的发展开辟了新的途径。

表 3 多模态融合与单一模态性能对比
Tab.3 Comparison of multi-modal fusion and single-modal performance %

模态	准确率	精确率	召回率	F1 分数
视频	82.62	83.09	82.62	82.51
音频	70.00	70.74	70.00	69.43
视频 + 音频	90.24	91.17	89.14	90.45

参 考 文 献

[1] BONDAD-REANTASO M G, MACKINNON B, KARUNASAGAR I, et al. Review of alternatives to antibiotic use in aquaculture [J]. Reviews in Aquaculture, 2023, 15(4): 1421 - 1451.

[2] 谷立帅. 基于时空特征融合的鱼类摄食行为量化系统研究与实现[D]. 大连: 大连海洋大学, 2024.

GU Lishuai. Research and implementation of fish feeding behavior quantification system based on spatiotemporal feature fusion [D]. Dalian: Dalian Ocean University, 2024. (in Chinese)

[3] LI W, DU Z, LI D, et al. A review of aquaculture: from single modality analysis to multimodality fusion[J]. Computers and Electronics in Agriculture, 2024, 226: 109367.

[4] 陈浩. 基于多源数据融合的圈养鲈鱼智能投饵决策系统研究[D]. 武汉: 华中农业大学, 2023.

CHEN Hao. Intelligent baiting decision system for captive bass based on multi-source data fusion[D]. Wuhan: Huazhong Agricultural University, 2023. (in Chinese)

[5] 冯双星, 王丁弘, 潘良, 等. 基于轻量型 S3D 算法的鱼类摄食强度识别系统设计与试验[J]. 渔业现代化, 2023, 50(3): 79 - 86.

FENG Shuangxing, WANG Dinghong, PAN Liang, et al. Implementation of fish feeding intensity identification system using light-weight S3D algorithm[J]. Fishery Modernization, 2023, 50(3): 79 - 86. (in Chinese)

[6] WU Y, WANG X, SHI Y, et al. Fish feeding intensity assessment method using deep learning-based analysis of feeding

- splashes[J]. Computers and Electronics in Agriculture, 2024, 221: 108995.
- [7] CAI K, YANG Z, GAO T, et al. Efficient recognition of fish feeding behavior: a novel two-stage framework pioneering intelligent aquaculture strategies[J]. Computers and Electronics in Agriculture, 2024, 224: 109129.
- [8] LIU H, MA X, YU Y, et al. Application of deep learning-based object detection techniques in fish aquaculture: a review[J]. J. Mar. Sci. Eng., 2023, 11(4): 867–888.
- [9] 胥婧雯,于红,张鹏,等. 基于声音与视觉特征多级融合的鱼类行为识别模型 U-FusionNet-ResNet50+SENet[J]. 大连海洋大学学报,2023,38(2):348–356.
- XU Jingwen, YU Hong, ZHANG Peng, et al. A fish behavior recognition model based on multi-level fusion of sound and vision U-FusionNet-ResNet50+SENet[J]. Journal of Dalian Ocean University, 2023, 38(2): 348–356. (in Chinese)
- [10] DU Z, CUI M, WANG Q, et al. Feeding intensity assessment of aquaculture fish using Mel spectrogram and deep learning algorithms[J]. Aquacultural Engineering, 2023, 102: 102345.
- [11] 郑金存,叶章颖,赵建,等. 融合摄食过程声像特征鱼类摄食强度量化方法研究[J]. 海洋与湖沼,2024,55(3):577–588.
- DENG Jincun, YE Zhangying, ZHAO Jian, et al. Sensitive detection of fish feeding intensity by using sound and image features of feeding process[J]. Oceanologia et Limnologia Sinica, 2024, 55(3): 577–588. (in Chinese)
- [12] 杨雨欣. 复杂条件下声音与视觉特征融合的鱼类行为识别方法研究[D]. 大连:大连海洋大学,2024.
- YANG Yuxin. Fish behavior recognition method based on acoustic and visual features fusion under complex conditions[D]. Dalian: Dalian Ocean University, 2024. (in Chinese)
- [13] DU Z, XU X, BAI Z, et al. Feature fusion strategy and improved GhostNet for accurate recognition of fish feeding behavior[J]. Computers and Electronics in Agriculture, 2023, 214: 108310.
- [14] DU Z, CUI M, XU X, et al. Harnessing multimodal data fusion to advance accurate identification of fish feeding intensity[J]. Biosystems Engineering, 2024, 246: 135–149.
- [15] 胡学龙,朱文韬,杨信廷,等. 基于水质-声音-视觉融合的循环水养殖鱼类摄食强度识别[J]. 农业工程学报,2023, 39(10): 141–150.
- HU Xuelong, ZHU Wentao, YANG Xinting, et al. Identification of feeding intensity in recirculating aquaculture fish using water quality-sound-vision fusion[J]. Transactions of the CSAE, 2023, 39(10): 141–150. (in Chinese)
- [16] 沈海霞,刘晓倩,张永旺. 高效养鱼投饵方法要点[J]. 渔业致富指南,2014(13):27–28.
- [17] 焦世璋. 水产饵料投喂注意事项[J]. 渔业致富指南,2014(7):31–33.
- [18] 沈海霞,刘晓倩,张永旺. 鱼塘投饵有技巧[J]. 农村百事通,2015(3):51–53.
- [19] 徐立鸿,黄志尊,龙伟,等. 基于 MobileViT-CBAM-BiLSTM 的开放式养殖环境鱼群摄食强度分类模型[J]. 农业机械学报,2024,55(11):147–153.
- XU Lihong, HUANG Zhizun, LONG Wei, et al. Classification model of fish feeding intensity based on MobileViT-CBAM-BiLSTM[J]. Transactions of the Chinese Society for Agricultural Machinery, 2024, 55(11): 147–153. (in Chinese)
- [20] 徐立鸿,黄薪,刘世晶. 基于改进 LRCN 的鱼群摄食强度分类模型[J]. 农业机械学报,2022,53(10):236–241.
- XU Lihong, HUANG Xin, LIU Shijing. Recognition of fish feeding intensity based on improved LRCN[J]. Transactions of the Chinese Society for Agricultural Machinery, 2022, 53(10): 236–241. (in Chinese)
- [21] 唐宸,徐立鸿,刘世晶. 基于光流法的鱼群摄食状态细粒度分类算法[J]. 农业工程学报,2021,37(9):238–244.
- TANG Chen, XU Lihong, LIU Shijing. Fine-grained classification algorithm of fish feeding state based on opticalflow method[J]. Transactions of the CSAE, 2021, 37(9): 238–244. (in Chinese)
- [22] 刘世晶,涂雪滢,钱程,等. 基于帧间光流特征和改进 RNN 的草鱼摄食状态分类[J]. 水生生物学报,2022,46(6):914–921.
- LIU Shijing, TU Xueying, QIAN Cheng, et al. Feeding state classification of grass carp based on optical flow and improved RNN[J]. Acta Hydrobiologica Sinica, 2022, 46(6): 914–921. (in Chinese)
- [23] ZHOU C, XU D, CHEN L, et al. Evaluation of fish feeding intensity in aquaculture using a convolutional neural network and machine vision[J]. Aquaculture, 2019, 507: 457–465.
- [24] 周超,徐大明,吝凯,等. 基于近红外机器视觉的鱼类摄食强度评估方法研究[J]. 智慧农业,2019,1(1):76–84.
- ZHOU Chao, XU Daming, LIN Kai, et al. Evaluation of fish feeding intensity in aquaculture based on near-infrared machine vision[J]. Smart Agriculture, 2019, 1(1): 76–84. (in Chinese)
- [25] 朱明,张镇府,黄凤,等. 基于轻量级神经网络 MobileNetV3-Small 的鲈鱼摄食状态分类[J]. 农业工程学报,2021, 37(19):165–172.
- ZHU Ming, ZHANG Zhenfu, HUANG Huang, et al. Classification of perch ingesting condition using lightweight neural network MobileNetV3-Small[J]. Transactions of the CSAE, 2021, 37(19): 165–172. (in Chinese)
- [26] 邓皓. 基于深度学习的鲈鱼目标检测和摄食强度分类模型研究[D]. 广州:仲恺农业工程学院,2023.
- DENG Hao. Research on deep learning-based bass object detection and feeding intensity classification model[D]. Guangzhou: Zhongkai University of Agriculture and Engineering, 2023. (in Chinese)
- [27] 杨雨欣,于红,杨宗轶,等. 基于 Mel 声谱图与改进 SEResNet 的鱼类行为识别[J]. 渔业现代化,2024,51(1):56–63.
- YANG Yuxin, YU Hong, YANG Zongyi, et al. Fish behavior recognition based on Mel spectrogram and improved SEResNet[J]. Fishery Modernization, 2024, 51(1): 56–63. (in Chinese)
- [28] 沈昕昊. 基于深度学习的声学场景分类方法研究[D]. 无锡:江南大学,2023.
- SHEN Xinhao. Research on classification of acoustic scenes based on deep learning[D]. Wuxi: Jiangnan University, 2023. (in Chinese)