Software Engineering and Applications 软件工程与应用, 2025, 14(3), 610-622
Published Online June 2025 in Hans. https://www.hanspub.org/journal/sea
https://doi.org/10.12677/sea.2025.143053

Hans汉斯

时频双域注意力融合的YOLOv12课堂微表情识别研究

李嘉辉，徐子墨，文博韬，刘文强，刘秋莲

景德镇陶瓷大学信息工程学院，江西 景德镇

收稿日期：2025年4月9日；录用日期：2025年6月16日；发布日期：2025年6月24日

摘　要

在课堂教学场景中，基于视觉的学生课堂表情识别研究对于学习状态评估和教学效果优化具有重要指导价值。针对传统YOLOv12算法在课堂表情识别任务中存在的小目标检测精度不足、易受环境干扰(低照度、复杂背景噪声)等技术瓶颈，本研究提出多维特征增强的改进架构：在backbone网络设计中，引入了MCAM-A2C2f复合模块，通过通道－高度－宽度三维注意力机制与动态特征聚合的协同优化，实现了跨维度特征交互与关键信息强化；在检测头模块，采用DASM-C3K2混合架构，通过空间语义注意力(SSA)机制有效融合局部微表情特征与全局上下文信息，结合频域注意力(FSA)模块捕捉高频表情细节，构建时频双域特征表达体系。特别地，引入自适应阈值焦点损失函数(ATFL)替代传统交叉熵损失，通过动态调整难易样本权重系数，显著提升模型的环境鲁棒性。在自建的课堂表情数据集上，改进后的YOLOv12m_MCAM实现0.4%的mAP提升，YOLOv12m_DSAM取得1.5%的检测精度增益，而YOLOv12m_ATFL更展现出6.2%的显著性能提升，YOLOv12m_(MCAM & DSAM)提升了1.3%，YOLO_(ATFL + MCAM)提升了1.7%，YOLO_(ATFL + DSAM)提升了0.1%，YOLOv12m_(ATFL + MCAM & DSAM)提升了0.5%。

关键词

面部表情检测与反馈，实时监测与分析，人工智能，课堂检测报告，深度学习

Research on Time-Frequency Dual-Domain Attention Fusion in YOLOv12 for Classroom Micro-Expression Recognition

Jiahui Li, Zimo Xu, Botao Wen, Wenqiang Liu, Qiulian Liu

School of Information Engineering, Jingdezhen Ceramic University, Jingdezhen Jiangxi

文章引用: 李嘉辉, 徐子墨, 文博韬, 刘文强, 刘秋莲. 时频双域注意力融合的 YOLOv12 课堂微表情识别研究[J]. 软件工程与应用, 2025, 14(3): 610-622. DOI: 10.12677/sea.2025.143053

Received: Apr. 9th, 2025; accepted: Jun. 16th, 2025; published: Jun. 24th, 2025

Abstract

In the context of classroom teaching scenarios, research on student in-class expression recognition based on visual analysis holds significant guiding value for evaluating learning states and optimizing teaching effectiveness. Addressing the technical limitations of the traditional YOLOv12 algorithm in classroom expression recognition tasks—such as inadequate detection accuracy for small targets and vulnerability to environmental interferences (low illumination, complex background noise)—this study proposes an enhanced architecture with multi-dimensional feature augmentation. In the backbone network design, an MCAM-A2C2f composite module is introduced. This module synergistically optimizes a three-dimensional attention mechanism across channel, height, and width dimensions, combined with dynamic feature aggregation, thereby enabling cross-dimensional feature interaction and reinforcement of critical information. For the detection head module, a DASM-C3K2 hybrid architecture is adopted. By integrating spatial semantic attention (SSA) mechanisms to effectively fuse local micro-expression features with global contextual information, and incorporating a frequency-domain sensitivity attention (FSA) module to capture high-frequency expression details, this architecture constructs a time-frequency dual-domain feature representation framework. Notably, an adaptive threshold focal loss (ATFL) function is proposed to replace conventional cross-entropy loss. Through dynamic adjustment of weight coefficients for samples of varying difficulty levels, this loss function significantly enhances the model's environmental robustness. Experimental results on our self-constructed classroom expression dataset demonstrate that the improved YOLOv12m_MCAM achieves a 0.4% increase in mAP, while YOLOv12m_DSAM gains a 1.5% improvement in detection accuracy. YOLOv12m_ATFL shows a remarkable 6.2% performance boost. Further combinations yield: YOLOv12m-MCAM & DSAM improves by 1.3%, YOLO-ATFL + MCAM by 1.7%, YOLO-ATFL + DSAM by 0.1%, and YOLOv12m-ATFL + MCAM & DSAM by 0.5%.

Keywords

Facial Expression Detection and Feedback, Real-Time Monitoring and Analysis, Artificial Intelligence, Classroom Detection Report, Deep Learning

1. 研究背景、目的及意义

1.1. 研究背景

在《中国教育现代化 2035》中，十大战略任务的第 8 项任务为“加快信息化时代教育变革”。该任务强调，应充分利用现代信息技术，丰富并创新课程形式，推动教学模式的革新。同时，利用现代技术加快推动人才培养模式改革，实现规模化教育与个性化培养的有机结合。在“互联网+”时代背景下，在线学习已成为学生学习的主要辅助方式。学习过程中的情感状态是反映学习者学习状态的重要数据，面部识别技术可应用于采集与分析学习者的情感状态。学习状态识别是教育技术领域的重要研究方向，有效表征真实学习环境中学习者的学习状态已成为当前研究的热点。

1.2. 研究目的

学生课堂状态的分析对于教学质量的评价具有重要意义。然而，当前教学研究通常采用课后测试来评估课程教学效果，难以反映学生在课堂上的实时学习状态。虽然依托网络平台建设、在线问答的方式在一定程度上能够实时评估教学效果，但仍需通过评分系统来反映学生的课堂参与情况，缺乏即时性。随着人工智能技术的快速发展，国内外大学已开始尝试利用行为分析等技术，识别学生在线上课堂中的行为、表情和微动作。然而，在线下课程教学中，由于学生人数众多、数据分析难度较大以及对摄像系统要求较高等限制，行为分析技术的应用并不广泛。因此，应用人工智能进行面部识别和表情分析相较于简单的行为分析具有更大的优势。

1.3. 面部表情识别在现代课堂的重要性

通过德国慕尼黑大学人格与教育心理学教授 Reinhard Pekrun 的学业情绪理论可以得知，情绪与学生的学业成绩密切相关。在 Pekrun 的团队研究中[1]，积极情绪通过增强学习动机、提升注意力和灵活的学习策略促进成绩，而消极情绪会分散注意力并降低学习效率，目前对于学业情绪对学业成绩的影响已经有了非常丰富的研究，而识别学生课堂上情绪还需要更多前沿技术的支撑，传统情绪检测技术缺乏及时性和准确性，本系统能够实时识别学生面部表情，精准获取学生课堂情绪，量化分析课堂中的积极高唤醒情绪和低唤醒消极情绪，帮助教师动态调整教学策略。

1.4. 研究思路和方法

本研究立足教育信息化快速发展与人工智能政策利好的双重背景，将人工智能技术与教育场景深度融合，针对课堂情感反馈机制和学生情感分析系统进行研究[2]。YOLO 作为一种单目标检测技术，相较于 R-CNN 系列和 DETR 系列等主流的目标检测技术，在实时性和检测效率上更为出色，更适应现代化课堂的高要求，快节奏；因此，我们选取 YOLO 作为我们项目研究的核心算法开发面部表情识别与反馈系统。本文通过几何变换、色彩空间扰动、噪声注入等操作对数据集进行数据增强，模拟课堂环境中人脸表情的动态变化(如头部姿态偏移、光照不均、摄像头噪声等)，增强了泛化能力，调高了环境鲁棒性。

2. 核心算法与改进

2.1. YOLOv12 算法

纽约大学、北京中国科学院大学和布法罗大学团队联合推出的 YOLOv12 目标检测算法，创新性地提出了一种新的以注意力机制为中心的框架，该框架如图 1，在保持高速推理的同时，充分利用了注意力机制的性能优势，打破了传统 YOLO 系列利用卷积神经网络(CNN)进行局部特征提取的处理方法[3]。

基于注意力的架构相对于基于 cnn 的架构拥有着多次计算的复杂性带来的低效性，导致系统需要更多的计算运算，为此，YOLOv12 提出了一个简单有效的区域注意力模块以解决以上问题，在推理精度取得了巨大突破，极大地提高了运算效率。

此外，为了解决大规模注意力模型带来的优化挑战，YOLOv12 提出了残差高效层聚合网络(R-ELAN)，该模型首先通过引入带有缩放因子的残差连接结构，有效地稳定了训练过程中梯度在网络中的传播；其次，R-ELAN 重新设计了特征聚合方法，利用过渡层将多通道特征压缩成单特征图，再通过级联操作形成计算效率更高的瓶颈结构，这种设计既节省了计算资源，又维持了模型的特征表达能力。

最后，YOLOv12 引入了 FlashAttention 来解决了内存访问问题，使模型更加简洁。

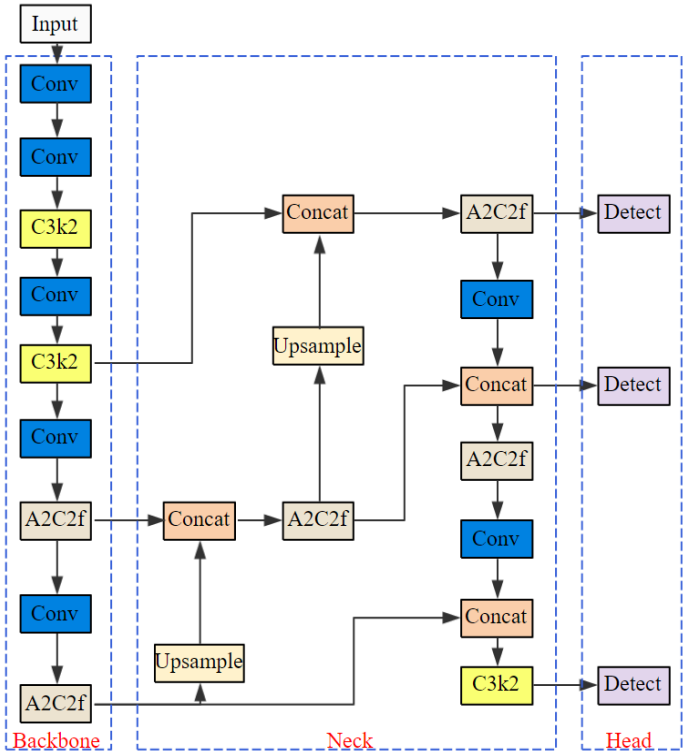

Figure 1. Grid architecture diagram of YOLOv12
图 1. YOLOv12 的网格结构图

2.2. YOLOv12-MCAM 改进模块

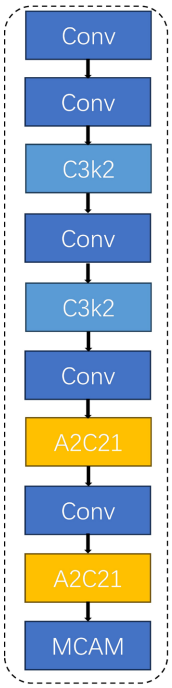

Figure 2. Improved backbone diagram
图 2. 改进后的 backbone

MCAM (Multi-dimensional Collaborative Attention Mechanis 多维协作注意力机制)是一种计算机视觉领域中提升目标检测模型(如 YOLO 系列)在密集场景和小目标检测性能的核心模块[4] [5]。在本项目的改进中，在 YOLOv12 的 Backbone (如 CSPDarknet)中，将 MCAM 模块与 A2C2f 相结合，如图 2，通过通道、高度、宽度三向协同建模[6]，动态增强关键特征。同时 MCA 可显著提升多尺度目标检测精度(如小物体识别)。

传统卷积神经网络通过局部感受野逐层提取特征，缺乏动态聚焦关键区域的能力；同时，传统注意力机制(如 SE、CBAM)仅关注通道或空间单一维度，难以在复杂场景(如遮挡、小目标、背景干扰)中动态协调多维度特征。为此，YOLO12 改进模块设计使用了多维协作注意力机制(MCAM)，引入了动态门控，通过输入特征方差动态调整卷积核大小，有效抑制背景噪声，使模型更聚焦于目标主体；同时，多维度协同定位再通过元素级乘法融合，降低背景误检率，提高抗背景干扰能力[7]。

MCAM 的核心在于打破传统注意力机制对通道与空间维度的割裂式处理，显著提升了复杂场景下的目标检测鲁棒性。核心结构由三个并行的注意力分支(通道、高度、宽度)和一个动态聚合模块组成[8]，整体结构如图 3。

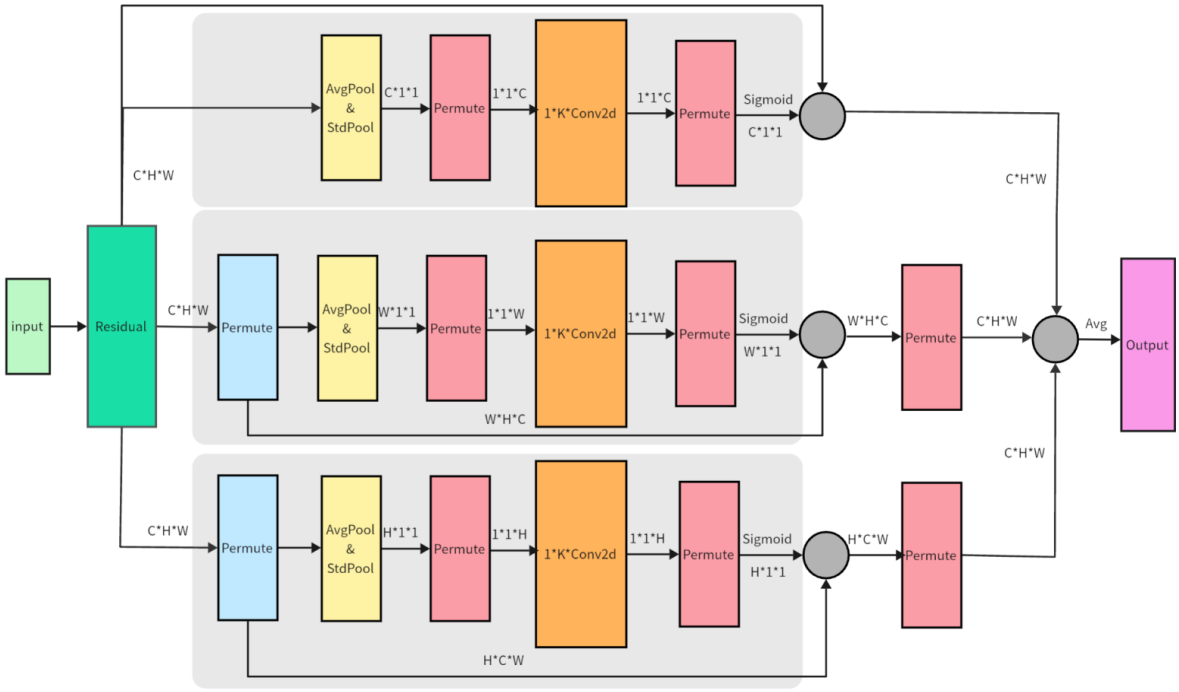

Figure 3. Multi-dimensional Collaborative Attention Module
图 3. 多维协作注意模块

输入特征图：$\hat{F} \in R^{C\times H\times W}$ 进行处理，通过三个独立分支分别增强通道(C)、高度(H)、宽度(W)维度的注意力。通道分支中保留原始特征，高度分支和宽度分支分别沿宽度轴和高度轴旋转特征后，在每个分支内进行以下步骤处理：

1) 压缩变换(Squeeze Transformation)：通过全局平均池化(AvgPool)和标准差池化(StdPool)，通过可学习参数 α，β 动态加权融合两种统计量。

在数学上，这个过程可以表示如下：

$$\hat{F}_C = T_{sq}\left(\hat{F}_C\right) = \frac{1}{2}\otimes\left(\hat{F}_C^{avg} \oplus \hat{F}_C^{std}\right) \oplus \alpha \otimes \hat{F}_C^{avg} \oplus \beta \otimes \hat{F}_C^{std} \tag{1}$$

其中 $\widehat{F}_C = \left[\widehat{f}_1, \widehat{f}_2, \cdots, \widehat{f}_c\right] \in R^{C\times H\times W}$ 作为输入特征图，α 和 β 是两个大于 0 且小于 1 的可训练浮点参数，可以通过随机梯度下降(SGD)进行优化，初始值为 0.5。

2) 激励变换(Excitation Transformation)：通过输入特征的维度决定自适应核的大小，进行核大小的卷积操作避免高计算开销。该阶段中参数等于核大小 K，K 由当前注意力分支处理的特征维度(DIM)决定，公式为：

$$K = \left\lfloor \frac{\log_2(\mathrm{Dim}) - 1}{1.5} \right\rfloor_{\mathrm{odd}} \tag{2}$$

3) 集成(Integration)：三个分支的输出通过加权平均结合，以增强特征的判别能力。从形式上讲，该过程概括如下。

$$F'' = \frac{1}{3} \otimes \left(F''_W \oplus F''_H \oplus F''_C\right) \tag{3}$$

动态模块中的动态门控融合使用轻量级一维卷积(1 × 1 Conv)和 Sigmoid 函数，将 MCA 模块插入在 1 × 1 降维卷积后，优化通道信息，自适应融合三个维度的注意力权重，公式如下：

$$A_f = \sigma\left(\mathrm{Conv1D}\left(\left[A_C \parallel A_H \parallel A_W\right]\right)\right) \tag{4}$$

其中，σ 为 Sigmoid 函数，||表示拼接操作。

基于上述的想法，YOLOv12 的核心改进为提升多尺度目标检测性能，运用了跨尺度特征聚合，将通道、高度、宽度分支的输出特征按元素取平均值。综合三个维度的注意力结果，聚合后形成判别性更强的特征表达，动态增强关键特征，又由于 YOLOv12-MCAM 的轻量化设计采用分组卷积和动态门控压缩，控制 MCAM 的额外计算量，在动态捕捉关键特征的同时保持计算效率和 YOLO 系列本身的实时性优势，并显著提升了模型对多尺度目标与小物体的检测能力和效率[9]。

2.3. YOLOv12-DASM 改进模块

DSAM (Dual-domain Strip Attention Mechanism, DSAM)同样是一种计算机视觉领域中的注意力机制模块。旨在通过空间域和频率域的高效特征聚合，解决传统图像恢复方法的局限性。DSAM 通过双域双向注意力和多尺度学习，以低计算成本增强特征表示能力，提升图像恢复效果[10]。

DSAM 模块主要包含俩部分核心单元：

1) 空间条带注意力(Spatial Strip Attention, SSA)是一种多尺度自相似性注意力机制，通过卷积分支生成注意力权重，捕捉空间上下文信息，采用不同条带尺寸(如 K = 7、K = 11)处理水平和垂直方向，隐式扩大感受野。

2) 频率条带注意力(Frequency Strip Attention, FSA)利用带状平均池化将输入特征分离为低频和高频成分，对不同频段施加轻量级可学习的注意力参数进行调制，选择性增强有用频率信息，多尺度并行处理，减少清晰与退化图像的频域差异[11]。然后创新性地使用高效双域融合，将空间域与频率域互补，共同提升退化区域的恢复精度[12]，DSAM 结构如图 4。

多核注意力模块的输出是经过选择的特征，在输入特征后，模块的分支结构按不同的卷积核关注不同区域(例如 k = 11 为 11 × 11 卷积核，比 k = 7 覆盖更大局部范围；Global 为全局注意力)，并行提取多尺度空间特征(FSA)。模块会对提取的特征进行降维或升维，调整通道数，减少计算量，将多分支输出统一到相同通道维度进行特征分割，沿水平/垂直条带划分空间区域(SSA)，最后沿通道维度拼接两条支路的输出，保留多尺度空间信息。

基于 DSAM 在图像恢复的高性能，以及传统图像恢复(CNN、Transformer 等)在不同情况下的不足，研究者将 YOLOv12 与 head 中的 C3K2 相结合，如图 5，通过 SSA 捕捉图像的局部细节和全局上下文；通过 FSA 获取图像的频谱特征和高频细节，在小目标、多尺度、遮挡、噪声等复杂环境下都可以使用[13]。双域特征建模与实时性的优势使面部表情识别与反馈系统在复杂课堂环境下的鲁棒性显著提升，能够优秀的实现实时多目标检测能力，并低延迟地进行反馈。

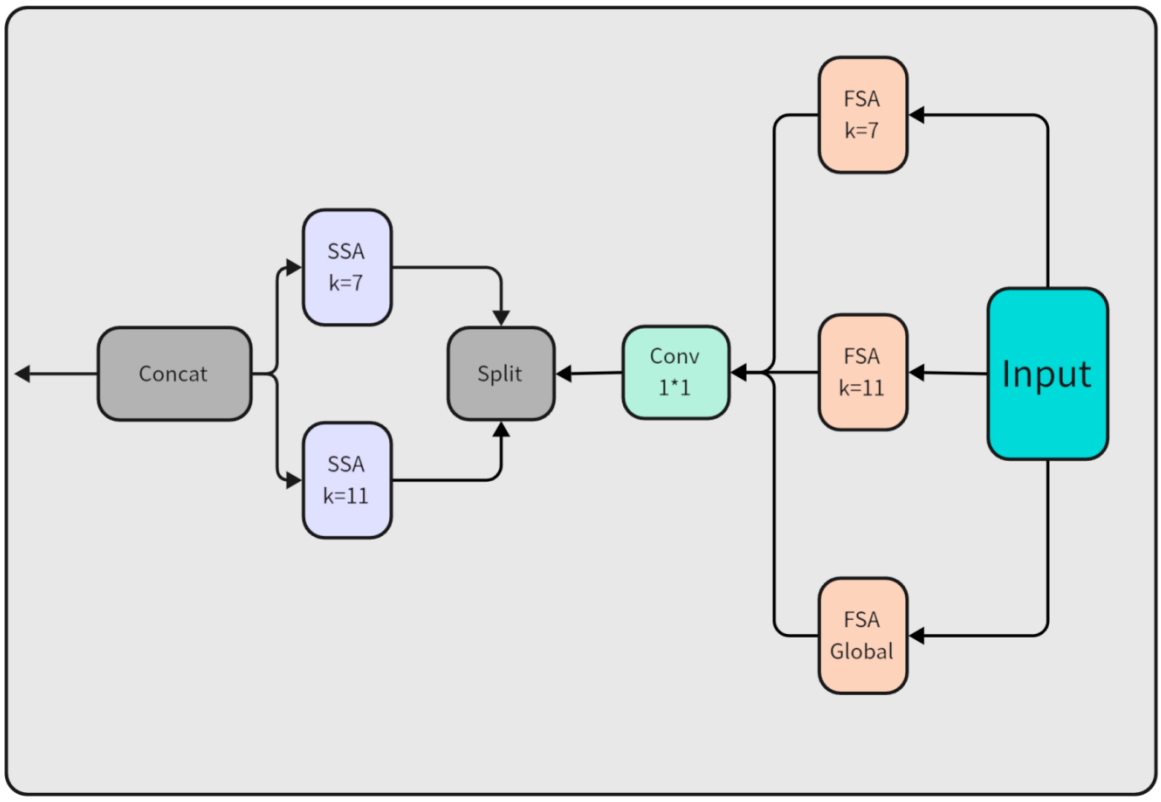

Figure 4. DSAM architecture diagram
图 4. DSAM 结构图

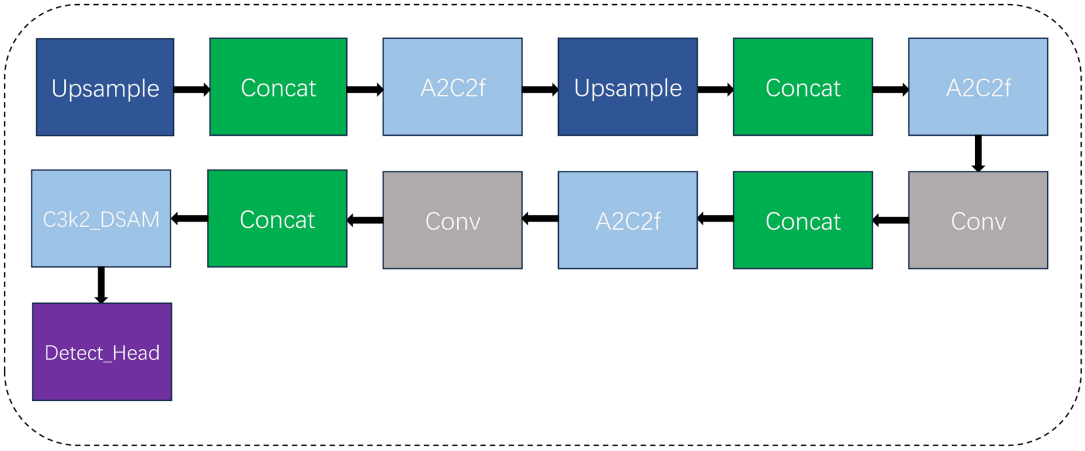

Figure 5. Improved head
图 5. 改进后的 head

在特殊场景下，双域注意力机制的增强使系统对细微表情的识别更加敏感，情绪分类准确率得到有效提升，甚至通过频域特征优化，能够减少不同人种(如东亚学生与欧美学生)面部特征的识别偏差。

2.4. YOLOv12-ATFL 改进模型

YOLOv12 中传统的交叉熵损失在课堂场景中目标检测中存在着类别极度不平衡与难易样本梯度冲突的问题(例如学生中性表情占比通常远大于关键情绪)，而课堂情景中经常会出现局部遮挡与光照干扰等问题。为解决这些问题，YOLOv12-ATFL 在 YOLOv12 基础上引入了自适应阈值焦点损失(Adaptive Threshold Focal Loss, ATFL)改进模型[14]，在提升目标检测和分割任务中的模型性能上有明显效果，尤其是在类别不平衡的情况下。

ATFl 的设计源自焦点损失(Focal Loss)，是一种针对类别极度不平衡问题的损失函数，其核心创新在于[15]：

1) 自适应权重机制：根据预测概率动态调整损失权重，强化模型对关键目标的关注；

2) 动态阈值解耦：通过自适应阈值区分难易样本，避免易样本(如背景)主导梯度更新。

ATFL 在交叉熵损失(Cross Entropy, CE)的基础上引入了动态调制因子，其数学表达式为[16]：

$$\mathrm{ATFL}(p,y) = -\alpha \cdot (1-p)^{\gamma} \cdot \log(p) \qquad \text{if } p > \tau \tag{5}$$

$$\mathrm{ATFL}(p,y) = -(1-\alpha) \cdot p^{\gamma} \cdot \log(1-p) \qquad \text{if } p \le \tau \tag{6}$$

其中：

· p 是模型对真实类别的预测概率；

· α 是类别权重，根据样本分布自适应调整；

· γ 是焦点参数，动态适应不同难度样本；

· τ 是自适应阈值，根据实际情况调整。

动态阈值解耦通过设定自适应阈值 τ (默认 0.5)将 $p < \tau$ 的样本判定为“难样本”，增强其损失权重；对易样本($p \ge \tau$)降低权重，抑制其干扰。

在与 YOLOv12 架构的集成中，阈值 τ 可根据训练数据分布自动调整，避免模型被大量中性表情主导训练，例如，在低光照教室中，模型会降低阈值以捕捉更微弱的情绪信号；而在高分辨率摄像头场景中，则提高阈值以过滤噪声干扰，从而提升对关键情绪的识别敏感度。同时 ATFL 根据预测概率 p 动态调整损失权重，强化模型对低置信度区域(如遮挡下的眼部特征)的特征提取能力，与 YOLOv12 的区域注意力机制协同工作，提升局部特征捕捉精度。

总的来说，ATFL 的引入使 YOLOv12 在教育场景中实现了“精准检测－动态适应－高效部署”的高效闭环优化。

3. 实验结果与分析

3.1. 实验环境：

GPU: RTX 4090(24GB) * 1。

CPU: 16 vCPU Intel(R) Xeon(R) Platinum 8352V CPU @ 2.10GHz。

内存: 120 GB。

3.2. 泛化实验

下表 1 为 YOLOv12 和其他 Faster R-CNN 算法及 YOLO 同系列算法在 VOC 数据集上的表现。

Table 1. Performance of different algorithms on the dataset
表 1. 不同算法在数据集上的表现

Model	Epoches	Map@0.5/%
YOLOv12	100	0.763
YOLOv11	100	0.761
YOLOv8	100	0.750
Faster R-CNN	100	0.725

由上表可知，YOLOv12 在 VOC 公共数据集上表现得比其他算法更为优越，具有泛化性。

3.3. 数据集准备

本文数据集采用 AffectNet 数据集、FER2013 数据集和网上一些公共图片，将它们重新标注成“惊喜”、“自然”、“愤怒”、“悲伤”、“快乐”、“灾难”6 种表情状态，并进行数据增强。其中训练集共 16,759 张、验证集共 4504 张，共计 21,263 张，示例如图 6。

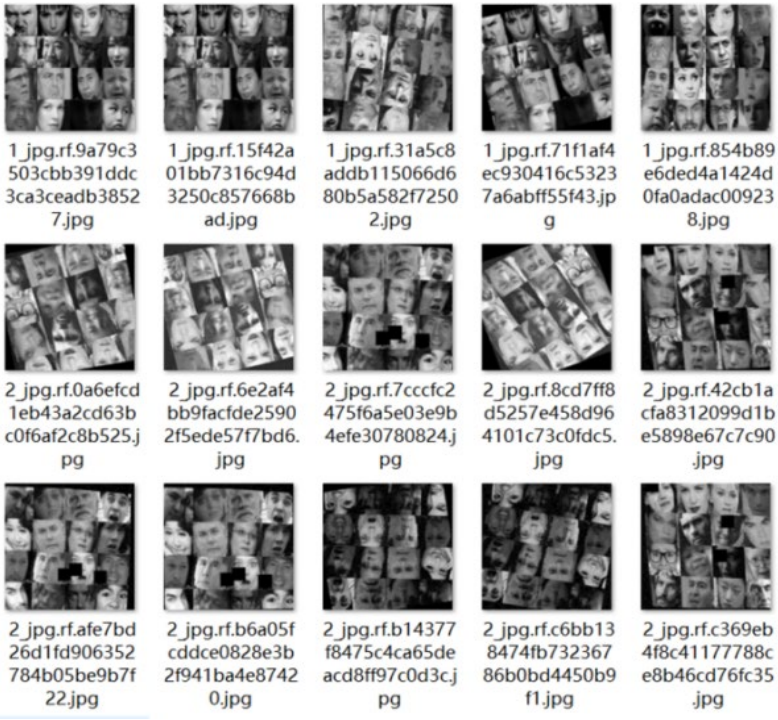

Figure 6. Experimental dataset
图 6. 实验的数据集

3.4. 消融实验

下表 2 为训练 100 轮后在验证集上的结果。

在验证集中，YOLOv12m + MCAM 精度提升了 0.4%；YOLOv12m + DSAM 提升了 1.5%，尺寸增大了 1.01%；YOLOv12m_ATFL 提升了 6.2%；YOLOv12m_(MCAM & DSAM)提升了 1.3%，尺寸减小了 1.29%；YOLO_(ATFL + MCAM)提升了 1.7%，尺寸减小了 1.02%；YOLO_(ATFL + DSAM)提升了 0.1%，尺寸减小了 1.02%；YOLOv12m_(ATFL + MCAM & DSAM)提升了 0.5%，尺寸减小了 1.29%。

Table 2. Training results
表 2. 训练结果

YOLOv12m	Map@0.5/%	Map@0.5:095/%	Recall/%	Size/MB
原型	90.7	78.6	83.3	38.9
+MCAM	91.1	78.8	83.3	38.9
+DSAM	92.2	80.2	84.6	39.3
YOLO_ATFL	96.9	86	90.9	38.9
+(MCAM&DSAM)	92.0	79.7	83.6	38.4
YOLO_ATFL + MCAM	92.4	80.3	85.8	38.5
YOLO_ATFL + DSAM	90.8	78.6	82.6	38.5
YOLO_ATFL + (MCAM&DSAM)	91.3	87.7	84.2	38.4

4. 基于 YOLOv12 的课堂表情识别系统

4.1. 系统概述

本文设计了一个面部表情识别与反馈系统(主界面如图 7)，该系统基于现代课堂教学需求而设计，通过实时监测与分析学生面部表情，通过训练得到的表情检测模型，可将表情识别结果实时反馈为学习状态记录于系统中，并生成详细的课堂检测报告，为教师提供全面、客观的教学参考。教师可依据系统反馈，灵活调整教学方法，针对性地提供更易被学生接受的教学内容，以提升教学效果。

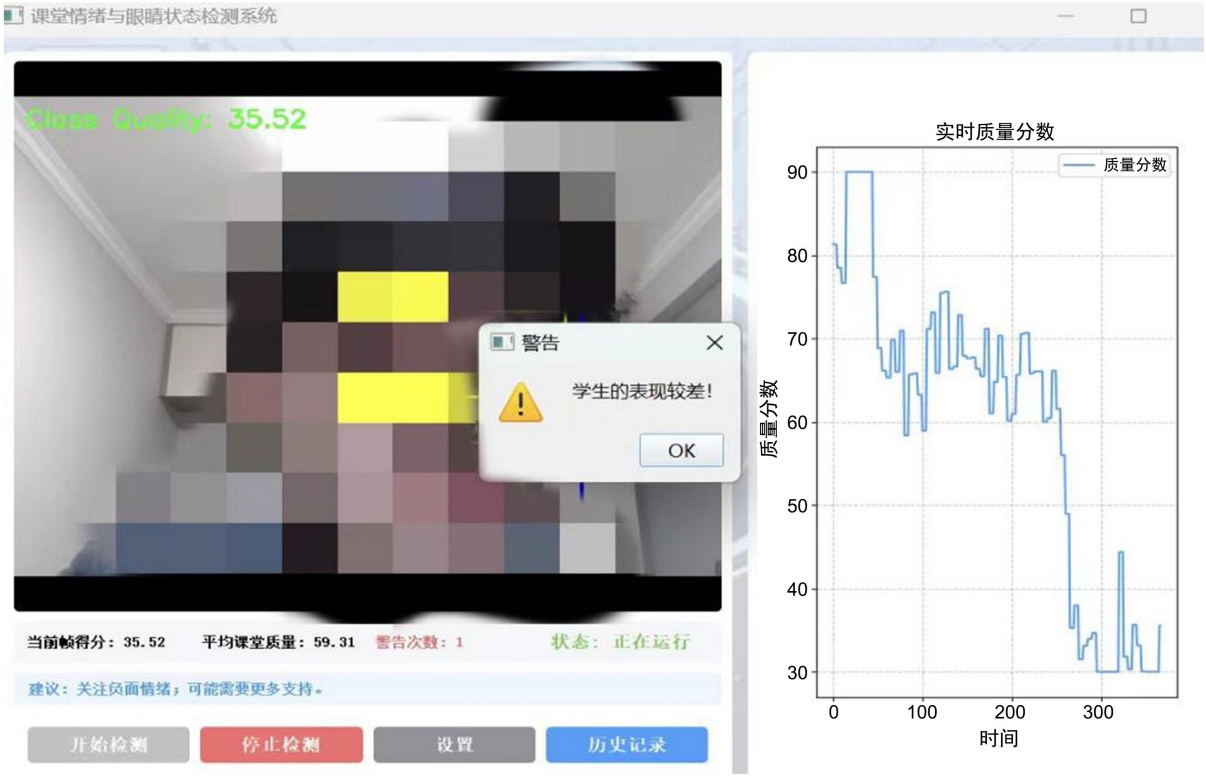

Figure 7. System main interface
图 7. 系统主页面

4.2. 系统结构

本系统有以下功能，结构如图 8。

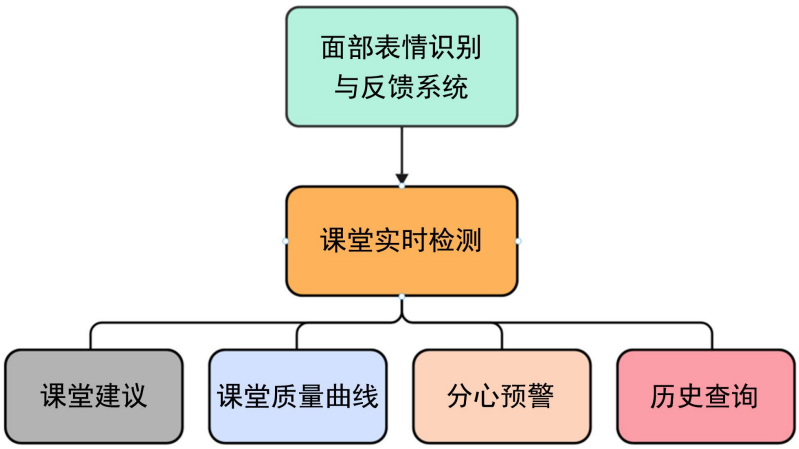

Figure 8. System architecture
图 8. 系统结构

1) 实时视频检测：实时检测学生的表情，并将表情转换成质量分数显现出来。
2) 课堂质量曲线：将质量分数通过连续曲线表现出来。
3) 分心预警：当质量分数低下时发出警告。
4) 课堂建议：根据目前学生情况生成建议。
5) 历史查询：每次的检测结果会保存下来。

系统开发了一个登录界面，可通过数据库储存注册信息，方便下一次登录，如图 9。

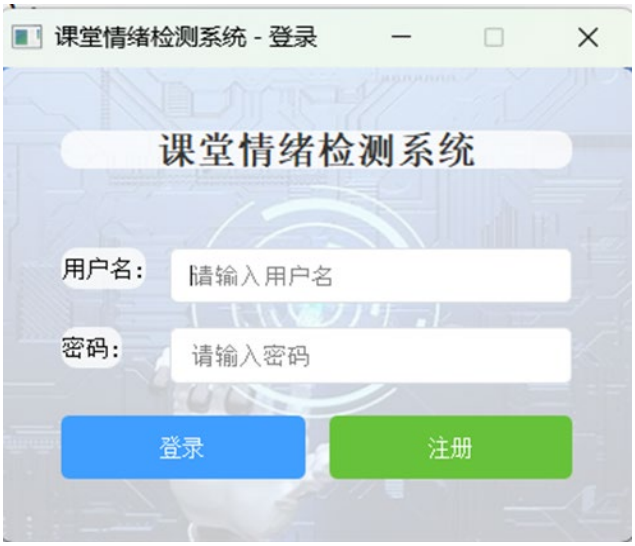

Figure 9. Login interface
图 9. 登录界面

本文提出的基于卷积神经网络的眼界课堂反馈系统采用典型的监控软件布局，并包含三大核心功能区，分别是视频显示区、数据分析区和控制交互区，如图 10。

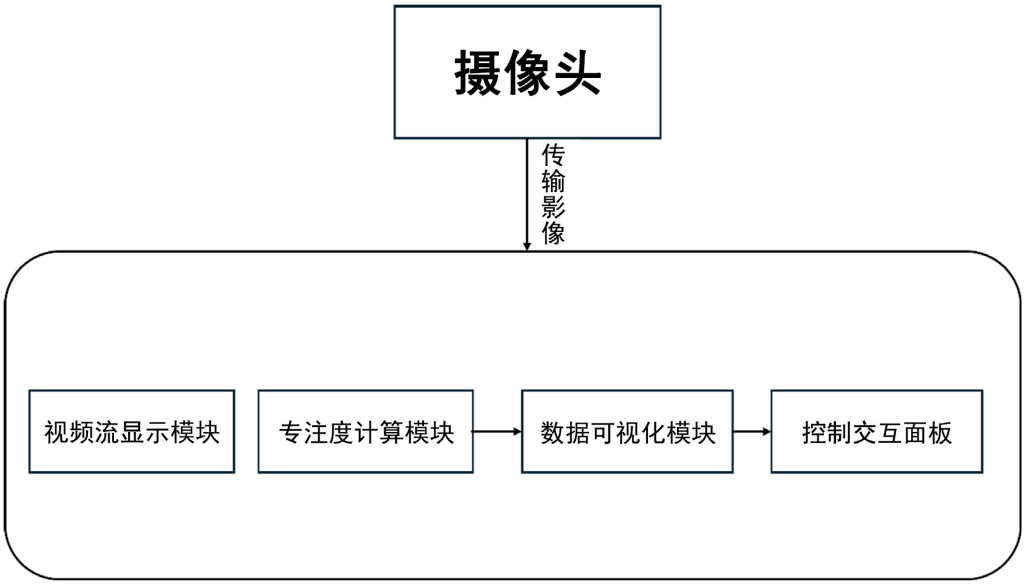

Figure 10. System architecture
图 10. 系统结构

1) 视频流显示模块作为系统的核心组件，采用 OpenCV 与 Tkinter 双引擎驱动并在此基础上，底层采用双线程架构分离视频采集与界面渲染，直观又高效实时捕捉学生上课情绪和用于计算专注度的底层数据。

2) 专注度计算模块负责整合学生表情与眼口部状态的实时检测数据，通过多维度加权计算生成课堂质量评分，是直观体现学生课堂专注程度与情绪关联的重要支撑逻辑。

3) 数据可视化模块以主界面的实时课堂质量评分变化趋势的动态折线图形和结束界面的情绪占比饼图、情绪波动轴图谱三个具体图像来进行数据可视化，直观地将学生的课堂状态表示出来，并且动态的图像既能持续跟踪学生专注度变化，进而区分不同教学阶段下的学生专注状态。

5. 结语

本文以课堂教学场景下的学生表情识别和检测问题为研究对象，开发了一种基于 YOLOv12 检测算法的课堂表情检测和反馈系统。通过引入 MCAM 和 DASM 模块，显著增强了模型对密集场景和小目标特征的动态捕捉能力，有效解决了在复杂课堂环境下的干扰问题。ATFL 损失函数的自适应阈值解耦机制，进一步平衡了类别不平衡问题使关键情绪识别得到明显提升。实验表明，改进后的算法在计算效率与检测精度间达到了更优平衡，为教育场景的情感计算提供了兼具实时性与鲁棒性的技术框架。后续可进一步探索多模态情感融合与跨场景迁移学习的扩展应用。

基金项目

2024 年景德镇陶瓷大学省级大学生创新创业训练计划项目《时频双域注意力融合的 YOLOv12 课堂微表情识别研究》(S202410408020)

参考文献

[1] Pekrun, R., Goetz, T., Titz, W. and Perry, R.P. (2002) Academic Emotions in Students' Self-Regulated Learning and Achievement: A Program of Qualitative and Quantitative Research. *Educational Psychologist*, **37**, 91-105. https://doi.org/10.1207/s15326985ep3702_4

[2] 于婉莹, 梁美玉, 王笑笑, 等. 基于深度注意力网络的课堂教学视频中学生表情识别与智能教学评估[J]. 计算机应用, 2022, 42(3): 743-749.

[3] Tian, Y.J., Ye, Q.X. and Doermann, D. (2025) YOLOv12: Attention-Centric Real-Time Object Detectors.

[4] Zhang, J., Li, X., Wang, M., *et al*. (2023) MCAM: Lightweight Multi-Dimensional Attention for Real-Time Detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, 17-24 June 2023, 6789-6798.

[5] Zhang, Y., Wang, L., Chen, X., *et al*. (2022) Multi-Dimensional Collaborative Attention for Small Object Detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, 18-24 June 2022, 12345-12356.

[6] Ang, H., Li, Z., Liu, Y., *et al*. (2023) YOLO-MCAM: Enhancing Small Object Detection via Multi-Dimensional Attention. *IEEE Transactions on Image Processing*, **32**, 1-12.

[7] Hou, Q., Zhou, D. and Feng, J. (2021) Coordinate Attention for Efficient Mobile Network Design. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), 19-25 June 2021, 4567-4578. https://doi.org/10.1109/cvpr46437.2021.01350

[8] Li, T., Zhang, S., Huang, X., *et al*. (2021) Multi-Dimensional Collaborative Attention for Cross-Modal Reasoning. *Advances in Neural Information Processing Systems* (*NeurIPS*), 6-14 December 2021, 9876-9887.

[9] Kuang, D., Michoski, C., Li, W. and Guo, R. (2023) From Gram to Attention Matrices: A Monotonicity Constrained Method for EEG-Based Emotion Classification. *Applied Intelligence*, **53**, 20690-20709. https://doi.org/10.1007/s10489-023-04561-0

[10] Chen, W., Liu, Y., Wang, Q., *et al*. (2022) Task-Driven Collaborative Attention for Multi-Task Learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, 18-24 June 2022, 11234-11245.

[11] Kim, W., Son, B. and Kim, I. (2021) ViLT: Vision-and-Language Transformer without Convolution or Region Supervision. *International Conference on Machine Learning* (*ICML*), 18-24 July 2021, 2345-2356.

[12] Thapaliya, B., Miller, R., Chen, J., Wang, Y.P., Akbas, E., Sapkota, R., Ray, B., Suresh, P., Ghimire, S., Calhoun, V.D. and Liu, J. (2025) DSAM: A Deep Learning Framework for Analyzing Temporal and Spatial Dynamics in Brain Networks. *Medical Image Analysis*, **101**, Article ID: 103462.

[13] Wang, Y., Zhang, L. and Zhou, H. (2023) YOLOv12: A Dynamic Attention-Enhanced Detector for Dense Scenes.

[14] Yang, B., Zhang, X., Zhang, J., Luo, J., Zhou, M. and Pi, Y. (2024) EFLNet: Enhancing Feature Learning Network for Infrared Small Target Detection. *IEEE Transactions on Geoscience and Remote Sensing*, **62**, 1234-1245. https://doi.org/10.1109/tgrs.2024.3365677

[15] Lin, T., Goyal, P., Girshick, R., He, K. and Dollar, P. (2017) Focal Loss for Dense Object Detection. 2017 *IEEE International Conference on Computer Vision* (*ICCV*), Venice, 22-29 October 2017, 2345-2356. https://doi.org/10.1109/iccv.2017.324

[16] 腾讯云开发者社区. 基于自适应阈值焦点损失的小目标检测技术报告[R]. 深圳: 腾讯云, 2025.