doi:10.6041/j.issn.1000-1298.2025.06.025

基于巡检机器人和改进 RT－DETR 的奶牛挑食行为识别方法

田富洋[1,2]　张立印[2,3]　张帅扬[4]　宋占华[1,3]　于镇伟[1,3]　张　姬[1,2]

（1. 山东农业大学机械与电子工程学院，泰安 271018；2. 山东省园艺机械与装备重点实验室，泰安 271018；
3. 山东省农业装备智能化工程实验室，泰安 271018；4. 中国农业机械化科学研究院集团有限公司，北京 100083）

摘要：针对目前复杂环境下奶牛在采食过程中挑食行为与采食行为差异不大、识别精度较低、人工识别劳动强度大等问题，本文提出了一种基于巡检机器人和改进 RT－DETR 模型的奶牛挑食行为识别方法。根据奶牛采食特性设计巡检机器人采集奶牛采食过程数据，分中午、下午和晚上 3 个时间段分别在 3 个牛棚进行采集，最终构建包含 3 个时间段共计 10 280 幅奶牛采食数据集。对 RT－DETR 模型进行改进，在 RT－DETR 模型浅层中引入 DAttention（DAT）模块和 Bi-Level Routing Attention（BRA）模块融合的 DBRA 结构，建立了新的图像特征提取结构，提升输入图像局部和全局特征深度融合能力；在 RT－DETR 模型编码器中融合 Efficient Multi－Scale Attention（EMA）模块，增强了提取高层次语义信息能力，更好地联系上下文信息。试验结果表明，改进后模型在奶牛采食视频数据集平均精度均值（mAP@0.5）为 99.1%，模型内存占用量为 39.6 MB，浮点计算量为 4.67×10^{10}，相较于原模型平均精度均值提高 7.4 个百分点，模型内存占用量降低 0.9 MB，浮点计算量减少 2%。与 Sparse R－CNN、YOLO v7－L、YOLO v8n、DINO、Swin Transformer 和 DETR 模型相比，平均精度均值（mAP@50）分别提高 8.5、9.8、7.8、6.6、11.4、9.5 个百分点。研究结果可以为实现畜牧养殖智能化提供技术支持。

关键词：奶牛挑食行为；巡检机器人；改进 RT－DETR；行为识别；注意力机制

中图分类号：TP391.4　　**文献标识码**：A　　**文章编号**：1000-1298(2025)06-0258-10　　**OSID**：

Selective Feeding Behavior Recognition Method for Dairy Cows Based on Inspection Robot and Improved RT－DETR

TIAN Fuyang[1,2]　ZHANG Liyin[2,3]　ZHANG Shuaiyang[4]　SONG Zhanhua[1,3]　YU Zhenwei[1,3]　ZHANG Ji[1,2]

（1. *College of Mechanical and Electronic Engineering*, *Shandong Agricultural University*, *Taian* 271018, *China*
2. *Shandong Provincial Key Laboratory of Horticultural Machinery and Equipment*, *Taian* 271018, *China*
3. *Shandong Provincial Agricultural Equipment Intelligent Engineering Laboratory*, *Taian* 271018, *China*
4. *Chinese Academy of Agricultural Mechanization Group Co.*, *Ltd.*, *Beijing* 100083, *China*）

Abstract: Aiming to address the challenges of low recognition accuracy, high labor intensity in manual identification, and minimal behavioral differences between selective feeding and normal feeding in dairy cows under complex environmental conditions, a method for identifying selective feeding behavior was proposed based on inspection robots and an improved RT－DETR model. An inspection robot was designed according to dairy cows' feeding characteristics to collect feeding process data. Data collection was conducted in three barns during three time periods (noon, afternoon, and night), ultimately establishing a dataset containing 10 280 feeding behavior images across these periods. The RT－DETR model was enhanced by integrating a DBRA structure, which combined the DAttention (DAT) module and Bi-Level Routing Attention (BRA) module into the shallow layers, creating a novel image feature extraction architecture to improve the deep fusion capability of local and global features. Additionally, the Efficient Multi－Scale Attention (EMA) module was incorporated into the model encoder to strengthen high-level semantic information extraction and contextual correlation. Experimental results demonstrated

收稿日期：2024－06－20　修回日期：2024－07－12

基金项目：国家重点研发计划项目（2023YFD2000704）和国家自然科学基金项目（32072786）

作者简介：田富洋（1979—），男，教授，博士，主要从事智能畜牧养殖技术与装备研究，E-mail：fytian@sdau.edu.cn

通信作者：张姬（1987—），女，副教授，博士，主要从事畜牧机械关键技术及智能化装备研究，E-mail：sdauzhangji@163.com

that the improved model achieved a mean average precision (mAP@0.5) of 99.1% on the dairy cow feeding video dataset, with a model memory occupancy of 39.6 MB and floating-point operations (FLOPs) of 4.67×10^{10}. Compared with the original model, the mAP@0.5 was increased by 7.4 percentage points, memory occupancy was reduced by 0.9 MB, and FLOPs was decreased by 2%. When compared with Sparse R－CNN, YOLO v7－L, YOLO v8n, DINO, Swin Transformer, and DETR models, the proposed model exhibited mAP@50 improvements of 8.5, 9.8, 7.8, 6.6, 11.4 and 9.5 percentage points, respectively. The findings enabled accurate differentiation between normal feeding and selective feeding behaviors, providing technical support for intelligent livestock farming.

Key words: cow selective feeding behavior; inspection robot; improved RT－DETR; behavior recognition; attention mechanism

0　引言

近年来,我国对畜牧业发展越来越重视,预计到2025 年,我国乳制品总产量将达到 4.1×10^7 t,同时,百头以上奶牛规模养殖占比也将提升至75%[1]。随着奶牛养殖业迅速发展和规模化程度提高,对奶牛健康管理和饲料加工面临着巨大挑战[2]。对奶牛采食行为实时识别检测,可以及时发现奶牛异常现象,对疾病防控、自身发育评估和饲料加工有重要意义[3-4]。

目前,常见的奶牛行为识别方法主要分为基于佩戴式传感器的识别方法[5]和基于计算机视觉的识别方法[6]。基于佩戴式传感器的识别方法是在奶牛不同部位佩戴传感器,可采集奶牛站立、行走、反刍以及甩尾等活动数据[7],判断奶牛行为。但这种方法可能会导致奶牛产生应激反应,并且传感器容易受损、老化以及能量消耗,进而影响对奶牛行为的准确识别。基于计算机视觉的识别方法大多利用监控等设备或者人工拍摄视频或图像,再利用人工智能技术对视频或图像进行处理,自动识别和分析奶牛各种行为。段青玲等[8]提出了一种基于 SNSS－YOLO v7 的检测模型,用于实现复杂环境下奶牛 7 种常见行为识别,识别平均精度均值达到 95.20%。程菲[9]提出了基于三轴加速度的奶牛采食行为检测方法,实现了奶牛采食行为识别,识别准确率达到92.8%。宋怀波等[10]通过比较压缩跟踪算法和核相关滤波算法在反刍检测中的效果,证明了 KCF 算法可行性。姬江涛等[11]利用改进 FlowNet2.0 光流算法实现了奶牛反刍咀嚼频次检测,识别准确率达到99.39%。王克俭等[12]提出一种基于时空特征的奶牛行为识别模型,实现了奶牛进食、行走、躺卧、站立行为识别,平均识别准确率达到 82.6%。WANG 等[13]利用改进的三维卷积模型实现了奶牛行走、喝水、进食等行为识别,识别准确率达到 98.17%。MA 等[14]通过改进 Rexnet 3D 网络实现了单头奶牛基本运动行为识别,识别准确率达到 95%。YU 等[15]利用改进 YOLO v4 模型实现了对奶牛采食行为检测,识别准确率达到 97.16%。FENG 等[16]使用一种可穿戴鼻环,通过收集数据,最终实现了奶牛采食、反刍行为识别,平均识别准确率达到 85%。

基于计算机视觉的识别方法可以有效降低对奶牛的影响,但数据采集时会存在劳动强度大、监控设备长期使用维修成本高且不会选择性拍摄采食视频等问题,并且奶牛挑食行为与采食行为动作相似度高,现有模型存在对小动作特征提取能力低、识别类似动作精度低等缺点,不适用挑食行为识别。针对这些问题,本文提出基于巡检机器人和改进 RT－DETR－DMSE 的奶牛采食过程中挑食行为识别方法,在实际养殖环境下,利用巡检机器人采集奶牛采食视频,构建奶牛采食行为和挑食行为数据集;选择 RT－DETR 的基础模型,在其浅层替换 DBRA 模块,替换后模型在提高对特征区域提取的同时减少计算量;在原 RT－DETR 模型编码器中融入 EMA 模块,改进模型在增强上下文的语义识别的同时进一步减少计算量。

1　材料与预处理

1.1　数据采集

利用巡检机器人与深度学习技术实现奶牛挑食行为识别,选取山东省泰安市金兰奶牛养殖中心 3 个牛棚作为试验场地,共有状态良好的泌乳奶牛 45 头。利用巡检机器人(图 1)采集奶牛采食数据。试验选用 STEREOLLABS 公司 ZED2 型双目深度相机,其最大拍摄角度为 110°(H)×70°(V)×120°(D),单眼分辨率为 1 280 像素×720 像素。为确保数据稳定性和准确性,传输帧速率设定为 30 f/s。在试验中,为确保对奶牛行为的有效捕捉,巡检机器人距离围栏 2 m,相机高度为 0.8 m(图 2)。视频采集时段为 12:00—21:00,由于光照不同会影响识别准确率,将其分为 12:00—14:00、16:00—18:00 和 19:00—21:00 3 个时间段进行采集,每个时间段分别在 3 个牛棚寻找正在采食的奶牛,使用巡检机器

人(图 2)固定拍摄 1 段或 2 段采食过程视频,奶牛采食持续时间大约为 20 min。

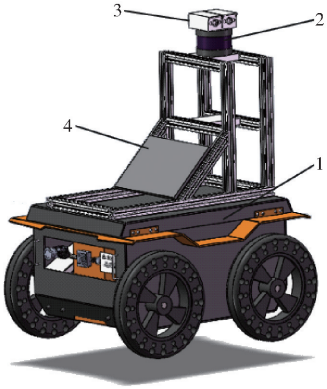

图 1 巡检机器人

Fig. 1 Inspection robots

1. 巡检车 2. 导航雷达 3. ZED2 型双目深度相机 4. 服务器

1.2 数据预处理

通过采集,共收集奶牛采食视频片段 13 条,视频长度为 20 ~ 25 min,每段视频奶牛数量为 3 ~ 7 头,选取有效视频时长共计 171.3 min。为保证挑食行为的连贯性,对奶牛采食视频数据按照每秒抽取

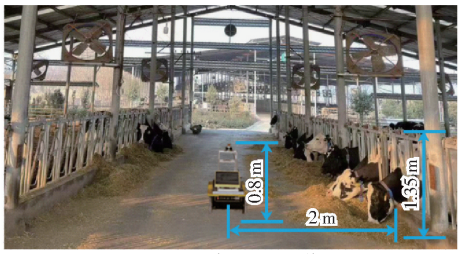

图 2 巡检机器人位置

Fig. 2 Inspecting location of robot

1 帧的方式截取图像,共获得图像 10 280 幅,使用 Labelimg 对奶牛行为进行标注。将 1 段连续帧中进行左右摆头拱食动作定义为挑食行为,根据观察和统计,1 次挑食行为持续时间为 3 ~ 20 s,将这一段连续帧定义为 1 次挑食动作,并对每幅图像中每头奶牛进行标注,奶牛挑食行为与正常采食行为判别标准如图 3 所示,奶牛 1 在这段连续帧中左右摆头为挑食行为;奶牛 2 在这段连续帧中无摆动为正常采食行为。奶牛行为数据集信息如表 1 所示。数据集按照比例 7∶2∶1 划分为训练集、验证集和测试集,最终获得训练集、验证集、测试集图像数量为 7 196、2 056、1 028 幅。

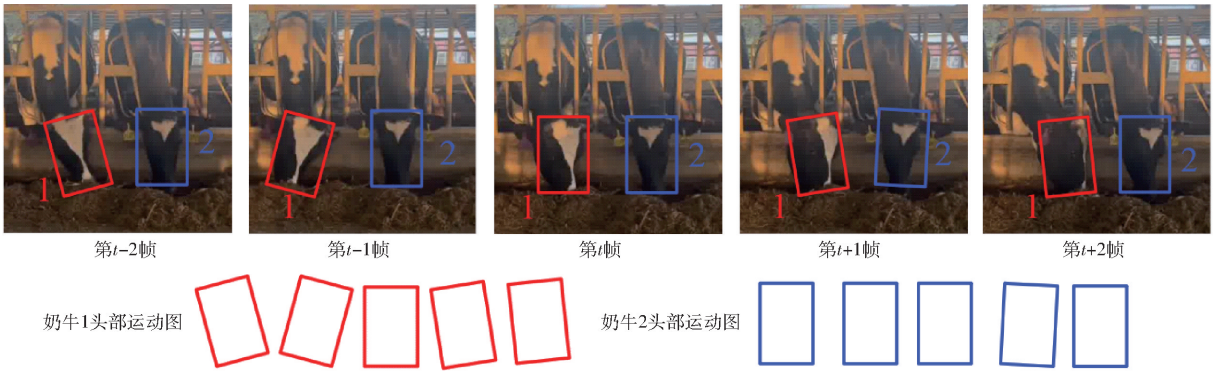

图 3 奶牛采食行为判别标准

Fig. 3 Criteria for judging feeding behavior of dairy cow

表 1 奶牛行为数据集信息

Tab. 1 Cow behavior dataset information

行为类别	行为描述	标签类别	标签数量
正常采食行为	正常低头吃食	normal	24 600
挑食行为	一段连续帧为左右摆头拱食动作	Selective	25 800
抬头咀嚼	抬头咀嚼	TT	16 796

2 奶牛挑食行为识别

2.1 RT-DETR 模型

与当前广泛应用的 YOLO 系列相比,DETR 模型舍弃锚点 anchors 生成过程,在推理阶段取消了非极大值抑制 NMS 来剔除多余的低质量结果,极大地简化了模型流程,实现了端到端检测模型[17]。RT-DETR 是首个实时端到端目标检测模型,融合了主干网络、混合编码器以及配备辅助预测头的 Transformer 解码器,形成了一个完整检测框架。RT-DETR 利用主干网络最后 3 个阶段输出特性作为编码器输入。混合编码器则通过尺度内的交互和跨尺度融合,将多尺度特征转换为一系列图像特征。随后,采用 IoU 感知的查询选择方法,从编码器输出序列中精选出固定数量图像特征,作为解码器初始对象查询。最后,配备辅助预测头的解码器通过迭代优化这些对象查询,最终生成精准的预测框和置信度分数[18]。

2.2 改进 RT-DETR-DMSE

为了解决奶牛正常采食行为和挑食行为特征相似度高、容易丢失小动作特征和动作重复性高得问题,本文提出 RT-DETR-DMSE 模型,模型结构如图 4 所示。为了解决行为特征相似度高的问题,在

骨干浅层引入一个深度融合模块(DBRA)代替原模型下采样层 DWConv,DBRA 模块包括 DAT 和 BRA 模块,DAT 模块可有效解决原始网络卷积和池化操作导致小动作特征丢失的问题,获得更详细的图像信息;BRA 模块则可有效捕获全局信息。利用融合 DBRA 模块,可将图像空间特征转换为深度特征,实现图像局部特征和全局特征之间的深度交互作用,解决了正常采食行为和挑食行为特征相似度高的问题,同时也提高了计算效率。此外,因为奶牛挑食行为动作重复性较高,辨识度较低,所以在编码器中融合了 EMA 模块,可以进一步增强高级语义特征的子注意力学习,充分联系上下文实现深度特征映射表示。

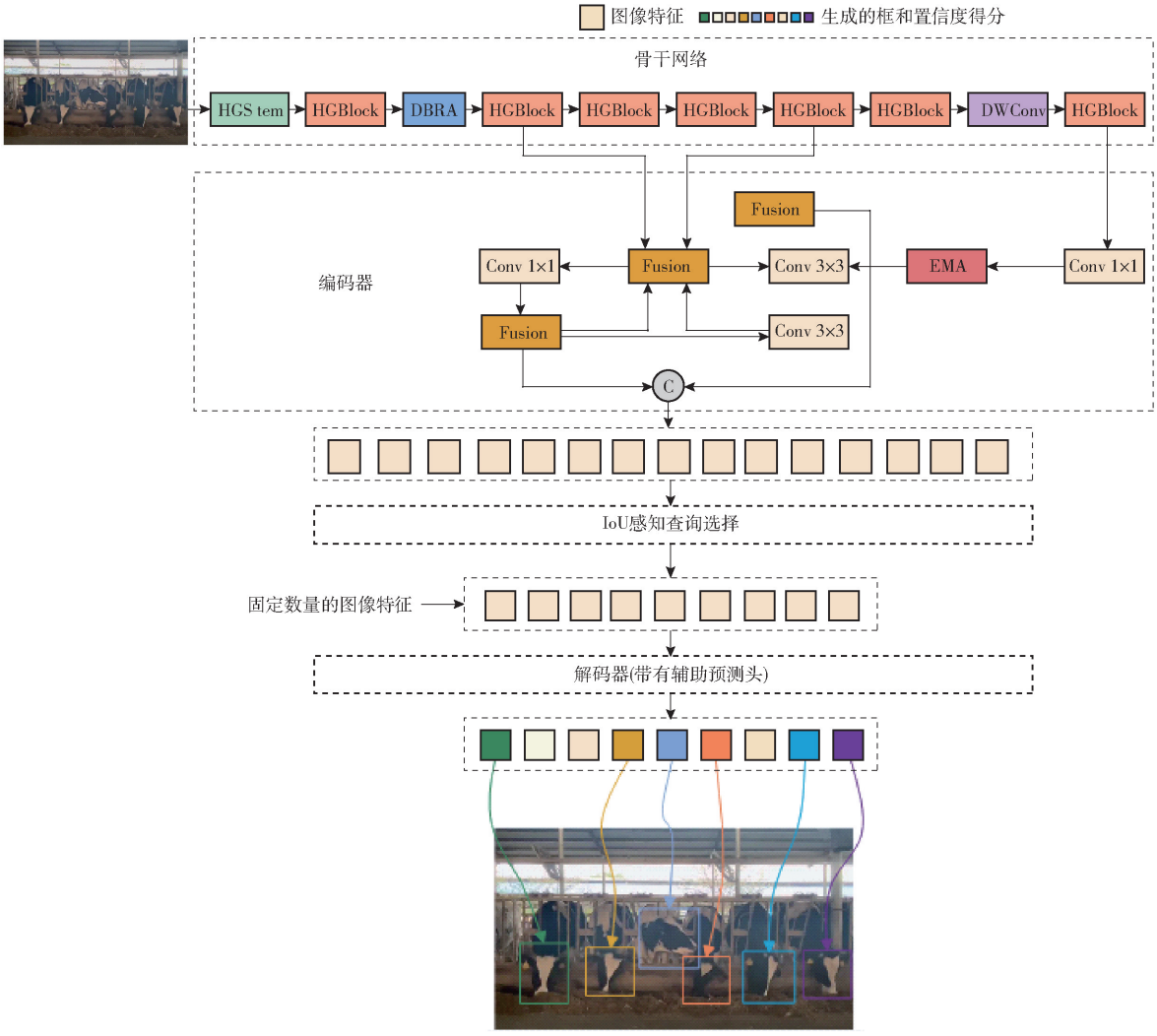

图 4　改进 RT－DETR－DMSE 模型结构

Fig. 4　Structure of improved RT－DETR－DMSE model

(1)DAT 模块

RT－DETR 网络模型使用传统的注意力机制,包含链卷积和池操作[19],计算量较大,在处理图像时会关注所有像素。这些操作可能会导致信息非歧视性损失,主要是因为在输入时低分辨率和小动作特征丢失,无法联系上下文相关信息导致。为了解决这一问题,引入 Dattention(DAT)模块。DAT 模块[20]是一种引入可变形注意力机制的视觉 Transformer。它只关注图像中的关键区域,在减少计算量的同时保持良好性能。模型结构有如图 5 所示,处理图像时分为 2 组:①参考点均匀放在特征图上,通过偏移学习可以查询偏移量。②从采样特征根据变形点投影变形计算特征值,可更好地得出动作特征。

给定一个输入特征图,生成网格作为参考,参考点值为线性间隔的 2D 坐标,在变形点位置进行特征采样,作为 key 和 value,然后投影矩阵分别表示 key 嵌入和 value 嵌入,投影矩阵为

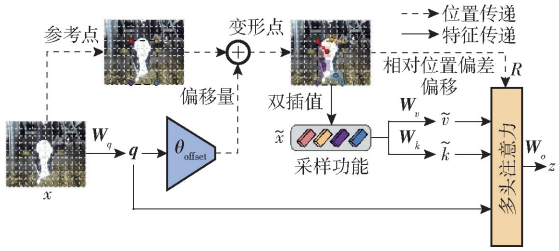

图 5　DAT 模块结构

Fig. 5　DAT module structure

$$\begin{cases} \boldsymbol{q} = x\boldsymbol{W}_q \\ \tilde{k} = \hat{x}\boldsymbol{W}_k \\ \tilde{v} = \hat{x}\boldsymbol{W}_v \end{cases} \tag{1}$$

$$\begin{cases} \Delta p = \theta_{\text{offset}}(\boldsymbol{q}) \\ \tilde{x} = \phi(x; p + \Delta p) \end{cases} \tag{2}$$

式中 $\boldsymbol{q}$——向量 x——标量

$\tilde{k}$、$\tilde{v}$——键和值嵌入

$\tilde{x}$——特定线性变换

$\boldsymbol{W}_k$——键输出向量

$\boldsymbol{W}_v$——值输出向量

$\boldsymbol{W}_q$——查询权重矩阵

p——参数量 Δp——参数增量

$\theta_{\text{offset}}(\boldsymbol{q})$——接受一个输入 $\boldsymbol{q}$ 并输出 Δp 的函数

$\phi(\cdot;\cdot)$——双线性插值,使其可微

可以解释为将采样函数 ϕ 设置为双线性插值,使其可微,即

$$\phi(z;(p_x,p_y)) = \sum_{(r_x,r_y)} g(p_x,r_x)g(p_y,r_y)z[r_y,r_x,:] \tag{3}$$

式中 z——输入数据

r——坐标

最终进行多头注意力,并采用相对位置偏移,输出表示为

$$z_m = \sigma(\boldsymbol{q}_m\tilde{\boldsymbol{k}}_m^{\mathrm{T}}/\sqrt{d} + \phi(\hat{B};R))\tilde{v}_m \tag{4}$$

式中 $\boldsymbol{q}_m$——查询向量

$\tilde{\boldsymbol{k}}_m$——键向量矩阵

d——键向量维度

B——位置信息

R——控制位置编码特征

使用 DAT 模块有效解决了模型计算量大的问题,在减少计算量同时,还可以增强提取能力,提高精度。

(2)BRA 模块

为了进一步降低模型计算量和提高模型准确度,引入 BiLevelRouting Attention(BRA)模块,BRA 模块[21]将一个特征图分成若干的小特征图区域,即首先将输入的特征图转换为特征向量,通过 reshape 变形将特征向量分为小特征图区域,再将这些特征图区域通过线性变化得到 3 块小特征向量,结构如图 6 所示。

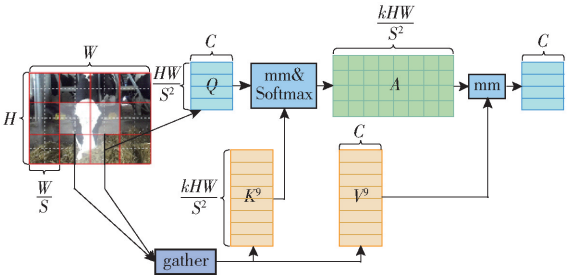

图 6 BRA 模块结构

Fig. 6 BRA module structure

BRA 模块可以过滤掉不相关区域,只关注框内特征区域,进一步减少计算量,并且提取关键特征,进一步提高精度。

(3)DBRA 模块

为了在性能和计算效率之间取得平衡,提出一种 Transformer 模型,即 DBRA 模块。DBRA 模块充分利用 CNN 和 Transformer 的优势,可以有效提高捕捉全局信息和局部信息的能力,同时减少计算速度。DBRA 模块结构如图 7 所示,包括卷积(DWConv)、DAT 层、BRA 和融合模块。

DBRA 模块由 2 个分支组成:①使用 DWConv,生成特征映射 X_1。②DBRA 旨在通过 DWConv 和 DAT 模块降低特征图,提高计算效率,同时保持特征细节,从而得到 X_2,然后通过 DAT 将空间信息进一步转换为通道信息,方便空间和深度特征的融合。利用 BRA 来识别关键特征区域和补充全局信息,得到 X_3。最后,DBRA 融合组件将 X_1、X_2 和 X_3 使用连接组件(CAT)沿着通道尺寸进行特征映射生成特征

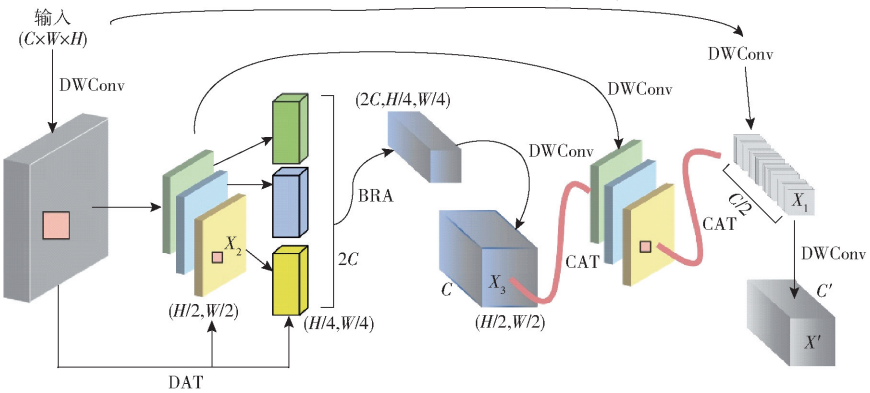

图 7 DBRA 模块结构

Fig. 7 DBRA module structure

映射 X'。DBRA 模块充分利用 DAT 将空间特征转换通道特征和 BRA 识别关键区域特征，实现了空间信息和深度信息的融合和关键特征提取。有效提高了精度，同时还提高了模型计算效率。

在 RT－DETR 基础网络中，浅层主要捕捉局部详细特征，因此，将原模型浅层 DWConv 模块替换为 DBRA 模块（图 4），提高全局特征和捕捉局部细节特征。

（4）EMA 模块

为了解决奶牛挑食行为和正常采食行为重复性动作、难识别等问题，在编码器融入一个 Efficient Multi－Scale Attention（EMA）模块。EMA 模块[22]是一种高效多尺度注意力模块，可以保留每个通道信息的同时减少计算量，充分联系上下文信息，使空间语义特征在每个特征组中分布良好，这样可以更好地关注挑食动作，减少其他干扰，使挑食行为动作与正常采食行为动作精准区分。此外，EMA 模块还通过跨维度交互来聚合并行分支的输出特征，融合不同尺度上下文信息，进而捕获像素级的成对关系。结构如图 8 所示。

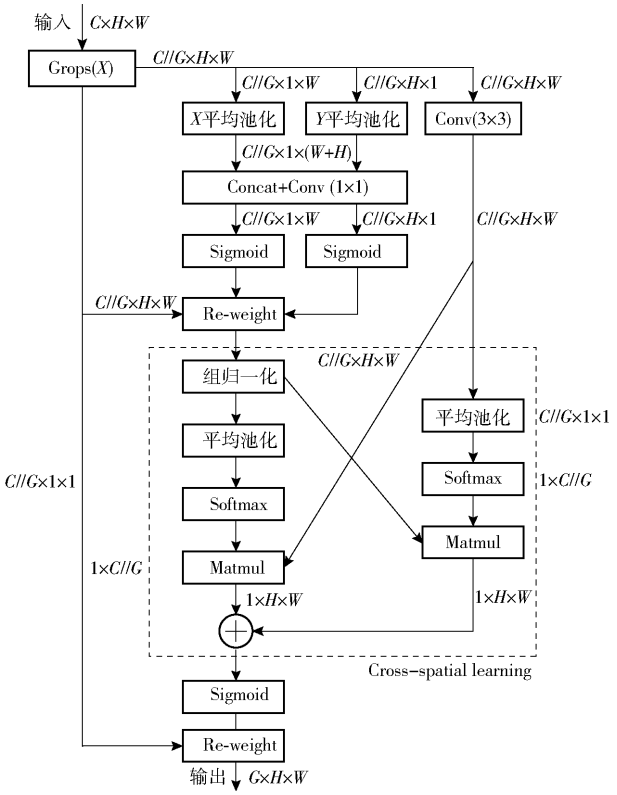

图 8　EMA 模块结构

Fig. 8　EMA module structure

EMA 模块主要包括一个 1×1 分支、一个 3×3 分支和一个跨维度学习的模块组成。1×1 分支将每组内通道注意力图像聚合，进而实现不同跨通道交互特征。跨维度学习模块使用二维全局平均池化对 1×1 分支输出进行全局信息编码，通过自然函数 Softmax 激活函数来拟合以上的线性变化，再通过并行处理的输出和运算，得到第 1 幅空间注意图。在 3×3 分支中，同样使用二维全局平均池化对 3×3 分支的输出经过 Softmax 激活函数后与 1×1 分支的输出结果点积，获得第 2 幅保留精准信息的空间注意力图。最后再与第 1 幅空间注意力图聚合通过 Softmax 激活函数和加权计算获取联系上下文的全局信息，可以有效解决背景干扰问题。

2.3　训练环境与参数设置

用于训练和测试的硬件平台配置为 Intel ® Xeon（R）W－2135CPU@3.70 GHz 处理器、32 GB 运行内存、11 GB GeForce RTX 2080 Ti 显卡和 500 G 固态硬盘，平台操作系统是一个 64 位的 Ubuntu 18.04 操作系统。模型构建使用 Python 3.8、CUDA 10.2 和 Pytorch 1.10.1。为了保证计算机运算速度并且不影响识别精度，在配置文件中设置输入图像尺寸为 640 像素×640 像素，关于超参数设置如表 2 所示。

表 2　超参数设置

Tab. 2　Hyperparameter settings

参数	数值
输入图像尺寸/（像素×像素）	640×640
迭代次数	200
耐心值	100
批量大小	4
初始学习率	0.000 1
最终学习率	1.0
权重衰减	0.000 1

2.4　评价指标

采用平均精度均值（mean Average Precision，mAP@0.5）、准确率（Precision，P）、召回率（Recall，R）、模型内存占用量、参数量（Params）、每秒帧数（FPS）和浮点计算量（Floating point operations，FLOPs）作为评价指标。平均精度均值（mAP@0.5）为目标检测中性能的关键指标[23]，mAP@50 表示 IoU 阈值为 0.5 的平均精度。

3　试验结果与分析

3.1　RT－DETR 基础模型选择

LV 等[18]设计出 5 种不同结构和规模的 RT－DETR 基础模型，包括 RT－DETR－R50、RT－DETR－R101、RT－DETR－R18、RT－DETR－L 和 RT－DETR－X。本文数据集检测效果如表 3 所示。

针对于本文的动作识别检测，RT－DETR－R18 模型在检测平均精度均值（mAP@50）方面表现优异，达到 91.7%。相比于 RT－DETR－R50、RT－

表 3 不同 RT - DETR 模型检测结果

Tab. 3 Test results of different RT - DETR models

模型	主干	mAP@50/%	模型内存占用量/MB	浮点计算量
RT - DETR - R50	ResNet50	91.5	86.1	1.348×10^{11}
RT - DETR - R101	ResNet101	90.5	131.3	2.476×10^{11}
RT - DETR - R18	ResNet18	91.7	40.5	58.60×10^{10}
RT - DETR - L	HGNetv2 - L	89.6	66.2	1.083×10^{11}
RT - DETR - X	HGNetv2 - X	90.1	128.6	2.323×10^{11}

DETR - R101、RT - DETR - L 和 RT - DETR - X 等模型,RT - DETR - R18 在该任务中的表现优势并不明显,但是其模型规模更小,具有更高的运行效率。RT - DETR - R18 精度比 RT - DETR - R50 只高 0.2 个百分点,这说明,较小的模型可以取得与较大模型相当的性能表现。另外,相比较其他 3 个较大的模型,即 RT - DETR - R101、RT - DETR - L 和 RT - DETR - X,RT - DETR - R18 在本文动作识别任务上的平均精度均值(mAP@50)分别提升 1.2、2.1、1.6 个百分点。因此,选择 RT - DETR - R18 作为改进的基础模型。

3.2 DBRA 模块消融试验

3.2.1 DBRA 模块位置比较

RT - DETR 的原始主干分别在第 2 层、第 4 层和第 8 层通过 DWConv 模块集成了一个简单的双重降采样操作。然而,这种操作会导致小动作特征信息的丢失。本文提出 DBRA 模块取代 DWConv,得到 7 种不同的结果如表 4 所示。

表 4 使用 DBRA 模块代替不同层的检测结果

Tab. 4 Using DBRA module instead of detection results of different layers

DBRA 数量	替换层	mAP@50/%	模型内存占用量/MB	浮点计算量
0		91.7	40.5	5.860×10^{11}
1	2	95.9	41.2	6.060×10^{10}
1	4	91.7	47.1	1.156×10^{11}
1	8	89.6	59.6	1.563×10^{11}
2	2,4	96.0	47.5	1.097×10^{11}
2	2,8	90.9	53.9	1.235×10^{11}
2	4,8	90.6	56.4	1.326×10^{11}
3	2,4,8	91.7	60.7	1.809×10^{11}

由表 4 可得,将 DBRA 模块放置在浅层中展现出了显著的优势。DBRA 模块不仅能够捕获浅层的详细特征,还能为整个网络提供全局信息,从而提升了模型性能。然而,当将 DBRA 模块替换第 8 层后,mAP@50 出现了下降。原因可能是将该模块放置在更深的层中导致了梯度爆炸问题,从而对模型泛化能力产生了负面影响。在当前试验中,DBRA 模块替换第 2 层时 mAP@50 为 95.9%,模型内存占用量为 41.2 MB;替换第 2 层和第 4 层时 mAP@50 为 96.0%,模型内存占用量为 47.5 MB。可以看出,相较于替换第 2 层,同时替换第 2 层和第 4 层可以使 mAP@50 提高 0.1 个百分点,但模型内存占用量增加 15.29%,在精度相近的情况下,替换第 2 层可以大大节省计算资源,提高效率。

综上分析,选择将原始模型中位于 DWConv 第 2 层的卷积层替换为 DBRA 模块。

3.2.2 DBRA 模块有效性

为了证明 DBRA 模块构建的有效性,进行了 DBRA 模块上的消融试验,结果如表 5 所示。

表 5 DBRA 模块的消融试验

Tab. 5 Ablation experiments on DBRA module

模型	mAP@50/%	模型内存占用量/MB	浮点计算量	帧率/(f·s^{-1})
RT - DETR - R18 + DBRA	95.9	41.2	6.06×10^{10}	38.7
+ DBRA 无 DAT	94.3	39.4	7.16×10^{10}	23.6
+ DBRA 无 BRA	94.9	39.6	8.93×10^{10}	40.3

由表 5 可知,在 DBRA 模块去除 DAT 模块导致 mAP@50 降低 1.6 个百分点,FPS 减少 15.1 f/s,去除 BRA 模块导致 mAP@50 降低 1 个百分点。虽然去除 DAT 和 BRA 模块模型内存占用量有所降低,但是 FLOPs 却有所增加,说明了运算速度并没有因为模型内存占用量降低而提高,BRA 模块组件对模型泛化能力有利,提高了检测位置场景能力;DAT 模块在增加通道数量的同时减少了特征图大小,进而减少注意周期,加速模型计算效率,而且,在没有 DAT 模块时,缺少了对细节特征的捕捉,导致平均精度均值(mAP@50)降低较多。综上所述,完整的 DBRA 模块有效平衡了模型识别精度和计算效率,很好地促进了目标局部和全局特征深度交互。

3.3 EMA 模块优越性

RT - DETR - R18 模型融入 DBRA 模块后,在挑食行为数据集中平均精度均值(mAP@50)和计算效率都有所提高。但原始 RT - DETR 模型在编码器中使用的是传统的自注意力机制,在重复性高的动作任务,会限制特征表达映射。为了解除这一限制,并进一步提高泛化能力,还需要保持原模型识别预测能力,改进模型在编码器中融合可以联系上下文信息的 EMA 模块。

不同注意力模块识别结果如表 6 所示,当编码器融入 ParNetAttentio 模块平均精度均值(mAP@50)降低 0.3 个百分点,并且模型内存占用量比原改

进模型高 0.7 MB，编码器融入 PSA 模块和 EMA 模块的模型平均精度均值（mAP@50）有所提升，融入 EMA 模块提高提升更为显著，为 3.2 个百分点，并且模型内存占用量比原改进模型降低 1.6 MB。EMA 模块可以进一步增加改进模型泛化能力，从而提高精度和计算效率。

表 6　融入不同注意力模块编码器检测结果

Tab.6　Encoder incorporated detection results of different attention modules

模型	mAP@50/%	模型内存占用量/MB	浮点计算量
RT－DETR－R18＋DBRA	95.9	41.2	6.060×10^{10}
＋EMA	99.1	39.6	4.670×10^{10}
＋PSA[24]	97.2	40.8	8.940×10^{10}
＋ParNetAttention[25]	95.6	41.9	1.009×10^{11}

3.4　不同模型试验结果对比

选择检测模型与 RT－DETR－DMSE 模型进行对比。比较模型包括 Sparse R－CNN[26]、YOLO v7[27]、YOLO v8、DINO[28]、Swin Transformer[29] 和 DETR，其中模型类型为 CNNs 有 Sparse R－CNN、YOLO v7 和 YOLO v8，模型类型为 Transformer 有 DINO 和 DETR。模型检测识别效果对比如表 7 所示。

表 7　不同模型识别效果对比

Tab.7　Comparison results of different models

模型	模型类型	mAP@50/%	模型内存占用量/MB	浮点计算量
Sparse R－CNN	CNNs	90.6	78.4	2.136×10^{11}
YOLO v7－L	CNNs	89.3	38.6	4.590×10^{10}
YOLO v8n	CNNs	91.3	6.2	1.120×10^{10}
DINO	Transformer	92.5	51.7	2.259×10^{11}
Swin Transformer	Transformer	87.7	46.5	2.543×10^{11}
DETR	Transformer	89.6	41.9	1.004×10^{11}
RT－DETR－R18	Transformer	91.7	40.5	5.860×10^{10}
RT－DETR－DMSE	Transformer	99.1	39.6	4.670×10^{10}

由表 7 可知，在基于 CNNs 的模型中，YOLO 系列虽然在模型内存占用量比 RT－DETR－DMSE 模型略小，但在检测精度和泛化能力方面显著低于 RT－DETR－DMSE 模型，YOLO v7－L 和 YOLO v8n 模型平均精度均值（mAP@50）相较于 RT－DETR－DMSE 模型分别低 9.8、7.8 个百分点，并且 Sparse R－CNN 模型相较于 RT－DETR－DMSE 模型平均精度均值（mAP@50）低 8.5 个百分点，模型内存占用量增加 38.8 MB。提出的 RT－DETR－DMSE 模型在检测能力、鲁棒性、泛化能力和 GFLOPs 方面具有优势。

基于 Transformer 的模型中，DINO 模型与 RT－DETR－R18 模型平均精度均值（mAP@50）都超过了 Swin Transformer 模型和 DETR 模型，然而，DINO 模型相较于提出的 RT－DETR－DMSE 模型，平均精度均值（mAP@50）低 6.6 个百分点，并且模型内存占用量高 12.1 MB，FLOPs 需要 RT－DETR－DMSE 模型两倍以上。Swin Transformer 模型平均精度均值（mAP@50）比 RT－DETR－DMSE 模型低 11.4 个百分点，并且表现出较慢的收敛速度，这是因为窗口位移操作过程中区域特征分散到非重叠窗口中，这对学习的完整性提出挑战[30]。

为了验证模型检测效果，选取 1 028 幅连续图像构成测试集，具体检测结果如图 9 所示，选取 3 个时间段图像进行对比。根据试验结果，各个模型在中午和下午的数据集中，中午数据集检测结果平均高 2.2 个百分点；在中午和晚上数据集中，中午数据集检测结果平均高 5.6 个百分点，可以看出每个时



图 9　不同模型置信度检测结果对比

Fig. 9　Comparison of detection results of different models

间段因为光照影响有差异,尤其是晚上有较大差异。RT－DETR－DMSE 相比于其他模型,在中午时间段,置信度平均提高 3.3 百分点;在傍晚时间段,置信度平均提高 2.91 个百分点;在晚上时间段,置信度平均提高 7.32 百分点,可以看出,这些模型在检测挑食行为和正常采食行为的动作时,本文提出的 RT－DETR－DMSE 网络模型受光照影响较小,且在各个时间段具有显著优势,更可以说明,挑食行为和正常采食行为的区分需要充分联系上下文的特征,学习更为细节的小动作特征。

不同模型平均精度均值对比如图 10 所示,可以看出 Sparse R－CNN、DETR、Swin Transformer 模型比 RT－DETR－DMSE 模型收敛速度较慢,YOLO v7－L、YOLO v8n、DINO 模型与 RT－DETR－DMSE 模型收敛速度相似,然而收敛后平均精度均值(mAP@50)均低于 RT－DETR－DMSE 模型。

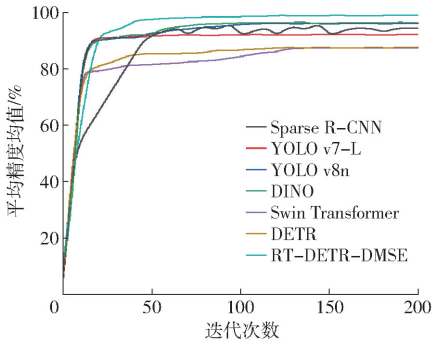

图 10 不同模型平均精度均值曲线对比

Fig. 10 Comparison of mAP@50 curves of different models

为了进一步验证 RT－DETR－DMSE 模型优越性,对 RT－DETR 与 RT－DETR－DMSE 两个模型在不同时间段输出特征进行了可视化分析。从图 11 可以看出,RT－DETR 模型在特征提取上存在不足,表现为特征响应图对目标奶牛头部的覆盖不够全面,且背景区域的激活度较高,导致目标与背景的界限模糊。相比之下,RT－DETR－DMSE 模型在 3 个时间段内均呈现出更加聚焦且精确的特征提取能力,特征响应图清晰地集中在奶牛头部的关键区域,显著提高了目标与背景之间的区分度,证明了其改进的有效性。

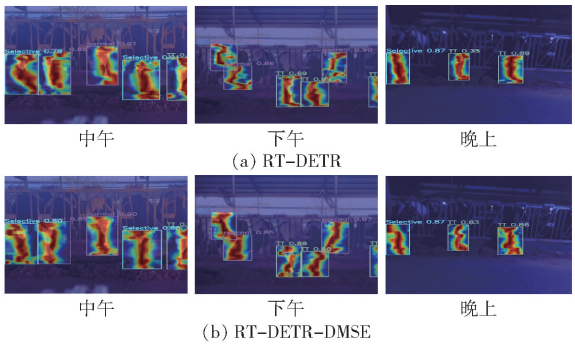

图 11 RT－DETR 与 RT－DETR－DMSE 模型热力图

Fig. 11 Heat maps of RT－DETR and RT－DETR－DMSE

综上分析,RT－DETR－DMSE 模型更适合本文采食数据集。

4 结论

(1)提出了基于巡检机器人和 RT－DETR－DMSE 模型的奶牛挑食行为识别方法。根据奶牛采食的特性,设计巡检机器人来采集奶牛采食过程的数据,构建采食数据集。在 RT－DETR－R18 网络主干的第 4 层引入一个深度融合模块(DBRA)代替原模型的下采样层 DWConv,解决了挑食行为和正常采食行为特征性相似度高的问题;此外,在编码器中融合 EMA 模块,充分联系了上下文实现深度特征映射的表示,提高了挑食行为平均精度均值。本文提出的改进模型可有效识别区分正常采食行为和挑食行为,平均精度均值为 99.1%,模型内存占用量为 39.6 MB,浮点计算量为 4.67×10^{10}。

(2)RT－DETR－DMSE 模型在本文数据集中平均精度均值分别高于 Sparse R－CNN、YOLO v7－L、YOLO v8n、DINO、Swin Transformer 和 DETR 模型 8.5、9.8、7.8、6.6、11.4、9.5 个百分点,优于其他网络模型。

参 考 文 献

[1] 农业农村部:“十四五”奶业竞争力提升行动方案[J]. 北方牧业, 2022(5): 14－15.

[2] 张星. 奶牛养殖产业发展现状及对策分析[J]. 现代畜牧科技, 2023(1): 25－28.
ZHANG Xing. Analysis of the development status and countermeasures of dairy farming industry[J]. Modern Animal Husbandry Science and Technology, 2023(1): 25－28. (in Chinese)

[3] DENIBEN J, BEINTMANN S, HOPPE S, et al. Influence of the addition of water to total mixed rations on the feeding behaviour, feed intake and milk performance of high-yielding dairy cows[J]. Livestock Science, 2021, 254: 104743.

[4] DEVRIES T J, BEAUCHEMIN K A, DOHME F, et al. Repeated ruminal acidosis challenges in lactating dairy cows at high and low risk for developing acidosis: feeding, ruminating, and lying behavior[J]. Journal of Dairy Science, 2009, 92(10): 5067－5078.

[5] LIU N, QI J, AN X, et al. A review on information technologies applicable to precision dairy farming: focus on behavior, health monitoring, and the precise feeding of dairy cows[J]. Agriculture, 2023, 13(10): 1858.

[6] 王政, 宋怀波, 王云飞, 等. 奶牛运动行为智能监测研究进展与技术趋势[J]. 智慧农业(中英文), 2022, 4(2): 36－52.
WANG Zheng, SONG Huaibo, WANG Yunfei, et al. Research progress and technology trend of inteligent morning of dairy cow motion behavior[J]. Smart Agriculture, 2022, 4(2): 36－52. (in Chinese)

[7] 熊强强, 齐志艺, 曾美琳. 基于 LoRa 无线传感网络的群养奶牛个体信息监测系统的设计与实现[J]. 电子设计工程, 2023, 31(23): 63－67, 71.
XIONG Qiangqiang, QI Zhiyi, ZENG Meilin. Design and implementation of individual information monitoring system for herd cows based on LoRa wireless sensor network[J]. Electronic Design Engineering, 2023, 31(23): 63－67, 71. (in Chinese)
[8] 段青玲, 赵芷青, 蒋涛, 等. 基于 SNSS－YOLO v7 的肉牛行为识别方法[J]. 农业机械学报, 2023, 54(10): 266－274, 347.
DUAN Qingling, ZHAO Zhiqing, JIANG Tao, et al. Behavior recognition method of beef cattle based on SNSS－YOLO v7[J]. Transactions of the Chinese Society for Agricultural Machinery, 2023, 54(10): 266－274, 347. (in Chinese)
[9] 程菲. 基于三轴加速度的奶牛采食行为监测与分析方法研究[D]. 哈尔滨: 东北农业大学, 2022.
CHENG Fei. Study on the monitoring and analysis method of ingestive-related behaviors of dairy cows based on triaxial acceleration[D]. Harbin: Northeast Agricultural University, 2022. (in Chinese)
[10] 宋怀波, 牛满堂, 姬存慧, 等. 基于视频分析的多目标奶牛反刍行为监测[J]. 农业工程学报, 2018, 34(18): 211－218.
SONG Huaibo, NIU Mantang, JI Cunhui, et al. Monitoring of multi-target cow ruminant behavior based on video analysis technology[J]. Transactions of the CSAE, 2018, 34(18): 211－218. (in Chinese)
[11] 姬江涛, 刘启航, 高荣华, 等. 基于改进 FlowNet 2.0 光流算法的奶牛反刍行为分析方法[J]. 农业机械学报, 2023, 54(1): 235－242.
JI Jiangtao, LIU Qihang, GAO Ronghua, et al. Ruminant behavior analysis method of dairy cows with improved FlowNet 2.0 optical flow algorithm[J]. Transactions of the Chinese Society for Agricultural Machinery, 2023, 54(1): 235－242. (in Chinese)
[12] 王克俭, 孙奕飞, 司永胜, 等. 基于时空特征的奶牛视频行为识别[J]. 农业机械学报, 2023, 54(5): 261－267, 358.
WANG Kejian, SUN Yifei, SI Yongsheng, et al. Video behavior recognition of dairy cows based on spatio-temporal features [J]. Transactions of the Chinese Society for Agricultural Machinery, 2023, 54(5): 261－267, 358. (in Chinese)
[13] WANG Y, LI R, WANG Z, et al. E3D: an efficient 3D CNN for the recognition of dairy cow's basic motion behavior[J]. Computers and Electronics in Agriculture, 2023, 205: 107607.
[14] MA S, ZHANG Q, LI T, et al. Basic motion behavior recognition of single dairy cow based on improved Rexnet 3D network [J]. Computers and Electronics in Agriculture, 2022, 194: 106772.
[15] YU Z, LIU Y, YU S, et al. Automatic detection method of dairy cow feeding behaviour based on YOLO improved model and edge computing[J]. Sensors, 2022, 22(9): 3271.
[16] FENG W, FAN D, WU H, et al. Cow behavior recognition based on wearable nose rings[J]. Animals, 2024, 14(8): 1187.
[17] SUN H, ZHANG S, TIAN X, et al. Pruning DETR: efficient end-to-end object detection with sparse structured pruning[J]. Signal, Image and Video Processing, 2023, 18(1): 129－135.
[18] LV W, XU S, ZHAO Y, et al. DETRs beat YOLOs on real-time object detection[J]. ArXiv, abs/2304.08069, 2023.
[19] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84－90.
[20] XIA Z, PAN X, SONG S, et al. DAT++: spatially dynamic vision transformer with deformable attention[J]. ArXiv, abs/2309.01430, 2023.
[21] ZHU L, WANG X, KE Z, et al. BiFormer: vision transformer with Bi-Level routing attention[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023: 10323－10333.
[22] OUYANG D, HE S, ZHAN J, et al. Efficient multi-scale attention module with cross-spatial learning[C]//ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023: 1－5.
[23] LIANG J, CHEN X, LIANG C, et al. A detection approach for late-autumn shoots of litchi based on unmanned aerial vehicle (UAV) remote sensing[J]. Computers and Electronics in Agriculture, 2023, 204: 107535.
[24] ZHANG H, ZU K, LU J, et al. EPSANet: an efficient pyramid squeeze attention block on convolutional neural network[C]// Proceedings of the Asian Conference on Computer Vision, 2021.
[25] GOYAL A, BOCHKOVSKIY A, DENG J, et al. Non-deep networks[J]. ArXiv, abs/2110.07641, 2021.
[26] SUN P, ZHANG R, JIANG Y, et al. Sparse R－CNN: End-to-End object detection with learnable proposals[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021: 14449－14458.
[27] WANG C Y, BOCHKOVSKIY A, LIAO H Y M. YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[J]. arXiv e-prints, 2022.
[28] ZHANG H, LI F, LIU S, et al. DINO: DETR with improved denoising anchor boxes for End-to-End object detection[J]. ArXiv, abs/2203.03605, 2022.
[29] LIU Z, LIN Y, CAO Y, et al. Swin transformer: hierarchical vision transformer using shifted windows[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021: 9992－10002.
[30] GAO L, ZHANG J, YANG C, et al. Cas-VSwin transformer: a variant swin transformer for surface-defect detection[J]. Computers in Industry, 2022, 140: 103689.