

RT-DETR 引导 MobileSAM 的堆叠球团矿粒径测量^{*}

郭海杰 崔雪红 龚玉洁 王旭 谢永琪

(青岛科技大学信息科学技术学院 青岛 266061)

摘要: 针对炼铁过程中的原材料球团密集堆叠、边界模糊、光照不均匀的难题,提出了一种基于改进 RT-DETR 与 MobileSAM 的级联式粒径检测与分割框架。在检测器的骨干网络中嵌入坐标注意力机制(CA),在 AIFI 编码器中引入高效判别频域前馈网络(EDFFN),引入多尺度特征调制(MFM)融合模块替代传统融合方式,最后以检测框为提示引导 MobileSAM 生成初始掩码,并引入基于动态感受野与高效多尺度注意力机制(EMA)的细化器对掩码边缘进行多尺度自适应修复,并结合几何拟合实现精确的粒径测量。该方法在球团矿数据集上的召回率达到 95.7%,mIoU 为 90.1%,有效提高了分割精度,减少了过度分割。该研究为工业炼铁过程的在线粒径监测提供了一种高精度、高鲁棒性的解决方案。

关键词: 球团粒径检测;实例分割;RT-DETR;MobileSAM;特征融合

中图分类号: TP391; TN919.8 **文献标识码:** A **国家标准学科分类代码:** 520.60

RT-DETR-guided MobileSAM for densely stacked pellet sizing

Guo Haijie Cui Xuehong Gong Yujie Wang Xu Xie Yongqi

(College of Information Science and Technology, Qingdao University of Science and Technology, Qingdao 266061, China)

Abstract: To address the challenges of dense stacking, blurred boundaries and uneven illumination of raw material pellets in the ironmaking process, this paper proposes a cascaded particle size detection and segmentation framework based on an improved RT-DETR and MobileSAM. In the detection stage, a coordinate attention (CA) mechanism is embedded into the backbone network to enhance feature extraction capabilities. Furthermore, an efficient discriminative frequency domain-based feedforward network (EDFFN) is introduced into the AIFI encoder and a multi-scale feature modulation (MFM) fusion module is employed to replace the traditional feature fusion method. Subsequently, the generated detection boxes serve as prompts to guide MobileSAM in generating initial masks. To optimize segmentation quality, a refiner incorporating dynamic receptive fields and efficient multi-scale attention (EMA) is designed for multi-scale adaptive repair of mask edges, which is then combined with geometric fitting to achieve precise particle size measurement. Experimental results on the pellet dataset demonstrate that the proposed method achieves a recall rate of 95.7% and a mIoU of 90.1%, effectively improving segmentation accuracy and mitigating over-segmentation. This study provides a high-precision and robust solution for online particle size monitoring in the industrial ironmaking process.

Keywords: pellet particle size detection; instance segmentation; RT-DETR; MobileSAM; feature fusion

0 引言

球团矿是高炉炼铁的重要原料,其粒度大小是衡量球团质量的重要标准之一^[1-2]。近年来,随着智能感知技术的快速发展,基于视觉语义约束与多传感器融合的非接触式测量方法已在工业精密测量领域得到广泛应用^[3],为解决复杂环境下的目标重构与尺寸测量难题提供了重要参考。传统的人工取样筛检测量法具有耗时耗力、准确度低、实时

性低和危害健康等缺点。随着深度学习技术应用到计算机视觉,越来越多的深度学习算法应用到了球团矿粒径检测^[4]。为了满足工业产线对实时性的严苛要求,张上等^[5]提出了一种基于改进 RT-DETR 的工业缺陷检测算法,通过优化模型架构在保持高精度的同时显著提升了推理速度,证明了 RT-DETR 在工业场景中的巨大潜力。张学锋等^[6]提出了一种结合改进 CLAHE 图像增强与多尺度霍夫变换的球团检测方法,但此类基于几何先验的传统算法在

面对非标准球形及严重粘连遮挡的目标时适应性较差,难以获取精确的像素级实例掩码。王猛等^[7]提出了一种融合注意力机制的 Mask R-CNN 模型对球团进行分割与粒径测量。但此法对密集球团的分割召回率较低。针对微小目标与复杂背景的干扰问题,任杰等^[8]提出了一种结合多尺度特征提取与注意力机制的轻量化检测方法,有效增强了模型对关键特征的捕捉能力并抑制了背景噪声。马伟宁^[9]等提出了一种基于分割一切模型(segment anything model, SAM)^[10]的机器视觉粒度识别算法。该方法采用 OTSU 算法自动生成初始提示点,并通过寻优移动策略进行二次优化,最后利用 SAM 生成掩膜并自适应拟合粒径。此方法鲁棒性较差,面对复杂场景容易出现漏检误检。

近年来,随着基础大模型的兴起,采用检测器引导分割的级联式框架逐渐成为解决密集目标分割的主流思路。该范式利用检测器生成的边界框作为先验信息,引导分割模型聚焦于感兴趣区域,从而有效解决背景干扰与目标粘连问题。尽管针对球团矿的专用级联模型研究尚处于起步阶段,但相关策略已在多个具有相似视觉挑战的领域展现出卓越性能。龚玉洁等^[11]提出了一种结合改进 YOLO11 与微调 SAM2 的实例分割方法,有效解决了工业球团矿检测中因小目标密集、粘连及遮挡导致的分割精度低和鲁棒性差的问题。但是具有计算开销大、推理速度慢的缺陷。在智慧农业领域,Wang 等^[12]针对密集堆叠的果实,提出了基于 YOLO 与 SAM 的级联框架,利用检测框提示显著提升了严重遮挡下的目标检出率;在医学影像分析中,Mansoori 等^[13]构建了混合 YOLO-SAM 架构,成功利用边界框将伪装目标从复杂背景中剥离,验证了提示学习在处理边缘模糊目标时的鲁棒性;在遥感领域,RSPrompter^[14]亦采用提示学习策略处理海量密集小目标,证明了该范式优于传统端到端分割。上述跨领域研究表明,级联策略在处理密集、遮挡及纹理同质化目标上具有显著的通用优势,这为解决球团粒径检测难题提供了重要的跨领域启发。

然而,将上述通用级联策略直接迁移至球团矿检测任务仍面临三大瓶颈。首先是检测端的漏检痛点:现有主流方案多基于锚框架构,其依赖的非极大值抑制(non-maximum suppression, NMS)后处理在极度密集堆叠场景下,极易错误抑制代表真实目标的重叠框,导致下游分割因缺失提示而失效。其次是分割端的粘连难题:球团纹理高度同质化且边界模糊,通用模型(如原始 SAM)难以在单一提示框内区分前景与相似背景,频发欠分割或边缘漂移现象。最后是部署端的算力制约:针对特定领域优化的 SOTA 方法往往伴随巨大的参数量与计算开销,难以满足工业现场对实时在线监测的严苛时效要求。为此,本文构建了一种融合改进 RT-DETR^[15]与 MobileSAM^[16]的级联式实例分割框架。本文选择该组合的动机在于首先 RT-DETR 是基于 Transformer 架构的端到端检测器,其天然摒弃了 NMS 后处理,能够利用全局注意力机制在密集堆

叠场景下精准区分重叠目标,从而为下游分割任务提供高召回率、高精度的完备边界框提示,从根本上避免了提示缺失;其次,选用 MobileSAM 是为了适配工业现场的算力限制,其通过轻量化蒸馏技术在大幅降低计算开销的同时保留了 SAM 的强分割能力,确保了在线监测的实时性。然而,尽管该基础级联框架具备了理论上的适用性,在面对球团矿这一特定复杂场景时仍显露出明显的性能短板:首先,原始 RT-DETR 的骨干网络缺乏对密集小目标的空间位置敏感度,且在处理球团模糊边界时易受背景噪声干扰,导致输出的提示框不够紧致,引入了额外的分割干扰;其次,传统的跨尺度特征融合方式难以有效应对球团尺寸的动态变化,易产生语义错配;最后,MobileSAM 为追求极致轻量化而牺牲了部分高频细节感知能力,导致其生成的掩膜在颗粒粘连处往往出现边缘平滑度不足或锯齿化现象,难以满足高精度粒径测量的要求。针对上述挑战,本文提出 4 项关键改进:首先,在骨干特征提取网络中嵌入坐标注意力机制(coordinate attention, CA)^[17]以增强对密集目标的定位感知;其次,在多尺度编码器中集成高效判别频域网络(efficient discriminative frequency domain-based feedforward network, EDFFN)^[18],在抑制噪声的同时重构模糊边缘;再次,引入多尺度特征调制(multi-scale feature modulation, MFM)融合模块^[19]以解决跨尺度特征融合时的语义错配问题。最后,基于 MobileSAM 设计了动态感受野模块(dynamic receptive field refinement module, DR)。该模块借鉴了 SKNet^[20]的思想,利用检测器输出的高精度边界框作为提示,通过多尺度感知与高效多尺度注意力机制(efficient multi-scale attention, EMA)^[21],对粗糙掩码进行像素级残差修复,从而实现高精度、高鲁棒性的在线粒径检测。

1 算 法

如图 1 所示,本文提出的球团矿粒径检测算法是一个端到端的级联架构,主要由 4 个核心组件构成:特征提取骨干网络、高效混合编码器、Transformer 解码器以及实例分割模块。

首先,图像输入至轻量化的 ResNet18 主干网络进行特征提取,在其最后两个阶段嵌入了坐标注意力模块,旨在特征提取过程中显式增强网络对密集球团目标的各个维度空间定位能力。随后,多尺度特征进入高效混合编码器;其中高层特征通过集成了高效判别频域网络的改进 AIFI 模块进行空频协同处理,以强化全局语义建模;而跨尺度特征则通过引入的调制融合模块,利用动态调制机制替代传统融合方式,实现了不同尺度特征间的自适应聚合。解码阶段沿用了 RT-DETR 的高效架构,通过不确定性最小化查询选择机制筛选出高质量的初始特征作为查询向量,并在 Transformer 解码器中进行端到端交互,从而摒弃繁琐的 NMS 后处理,直接输出高置信度的检测框。最后,该检测框作为位置提示引导 DR-MobileSAM 模块,利用动态感受

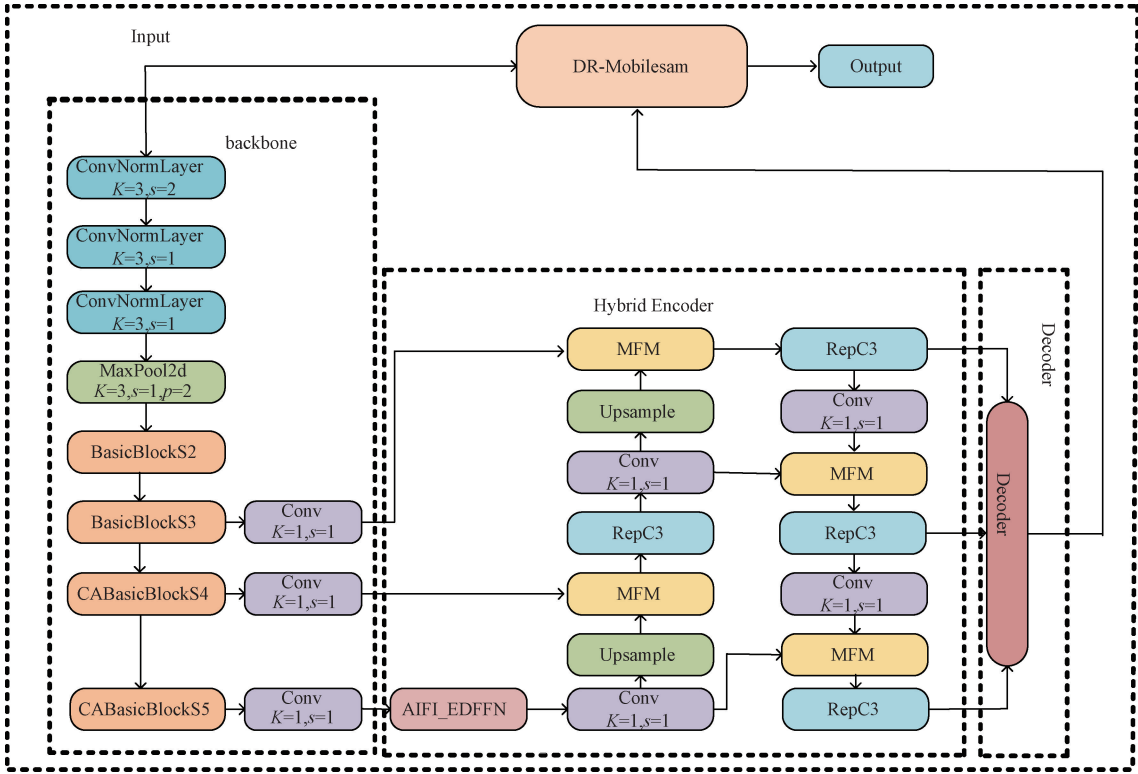


图 1 整体框架

Fig. 1 Overall framework

野机制精准捕获框内边界细节,生成高质量实例掩码以实现 对球团矿粒径的精确测量。

1.1 CEM-DETR 算法

1) 嵌入坐标注意力的 CABasicBlock

针对球团矿密集堆叠且相互遮挡的特性,传统的全局池化操作往往会导致空间位置信息的丢失,难以捕捉局部细节。为此,在骨干网络的 BasicBlock 中嵌入了坐标注意力机制,构建了 CABasicBlock 模块,如图 2 所示。该模块采用了一种分解式的空间编码策略,将传统的二维全局平均池化分解为两个并行的一维特征编码过程。具体而言,模块分别沿水平坐标和垂直坐标对输入特征图 X 进行聚合,生成一对方向感知的特征图。这种设计允许网络在捕捉沿一个空间方向的长程依赖关系的同时,保留另一个方向的精确位置信息。随后,这两个方向的特征图被拼接并在共享的 1×1 卷积层中进行变换与降维,以捕捉通道间的交互信息。生成的中间特征经过非线性激活后,再次沿空间维度被分裂为两个独立的张量,分别通过卷积变换生成对应方向的注意力权重向量。最终,通过将这两个互补的注意力向量与原始输入特征进行逐元素相乘,实现了对特征图在宽度和高度方向上的重加权。这一过程不仅强化了模型对密集目标的定位敏感度,还有效抑制了背景噪声的干扰。整个坐标注意力机制的计算过程可概括为:

$$Y = X \odot \sigma(F_h(f_h)) \odot \sigma(F_w(f_w)) \quad (1)$$

其中, X 为输入特征, Y 为加权后的输出特征, $\odot$ 表示

逐元素乘法, σ 为 Sigmoid 激活函数。 f_h 和 f_w 分别表示经过中间层编码与分裂后的水平和垂直方向特征张量, F_h 和 F_w 则代表将这些特征变换为最终注意力权重的卷积操作。通过这种“编码-分裂-重权”的机制, CABasicBlock 能够在不引入显著计算开销的前提下,显著提升检测器对密集球团的空间感知能力。

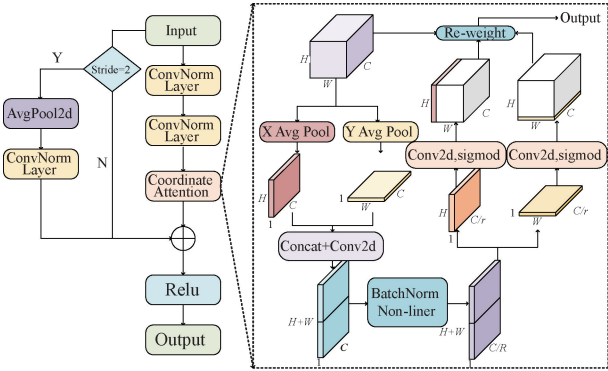


图 2 引入 CA 注意力机制的 Basicblock

Fig. 2 BasicBlock incorporating CA attention

2) AIFI-EDFFN: 高效判别频域前馈网络

为了解决传统 AIFI 模块在序列化过程中丢失空间结构信息的缺陷,本文设计了改进的 AIFI-EDFFN 编码器,如图 3 所示。该模块采用了序列-空间交替处理的策略。首先,输入特征 X 被展平为序列,并叠加二维位置编码以

保留空间坐标信息。经过多头自注意力(multi-head self-attention, MHSA)聚合全局上下文后,执行空间重构:

$$\mathbf{X}_{\text{spatial}} = \text{Reshape}(\text{LN}(\mathbf{X}_{\text{seq}} + \text{MHSA}(\mathbf{X}_{\text{seq}}))) \quad (2)$$

其中, $\mathbf{X}_{\text{spatial}} \in \mathbf{R}^{C \times H \times W}$, $\text{LN}(\cdot)$ 为层归一化。这一逆变换使得后续模块能够直接在二维特征图上进行操作,为引入空频协同机制奠定了基础。

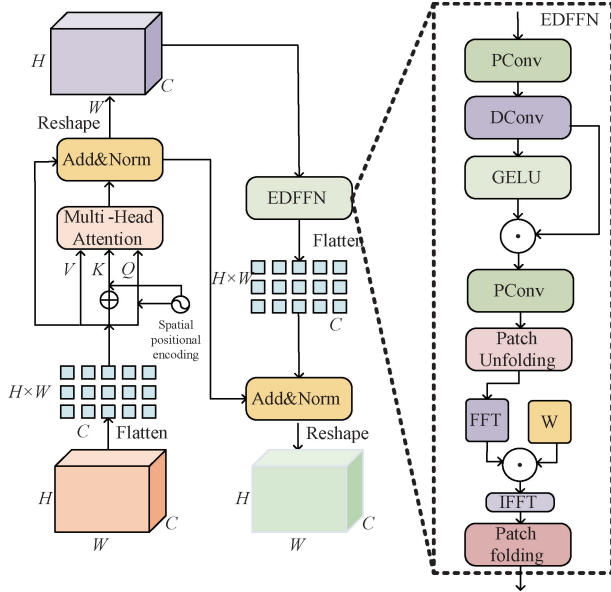


图 3 引入 EDFFN 的 AIFI

Fig. 3 AIFI incorporating EDFFN

重构后的空间特征进入高效判别频域前馈网络 EDFFN,该模块串联了空域局部增强与频域全局滤波两个阶段。空域门控增强的实现过程为:输入特征首先通过点卷积 PConv 升维,随后利用 3×3 深度卷积 DConv 提取局部几何特征。为了增强非线性表达并筛选有效信息,引入 GELU 门控机制。DConv 显式地编码了像素间的相对位置关系,强化了球团的边缘轮廓。实现过程如式(3)所示。

$$\mathbf{X}_{\text{gate}} = \text{GELU}(\text{DConv}(\mathbf{X}_{\text{proj}})) \odot \mathbf{X}_{\text{proj}} \quad (3)$$

受 FFTformer^[22] 频谱门控机制的启发,在频域全局滤波阶段旨在保留细节的同时滤除噪声。特征图首先经过 PConv 调整,通过分块与快速傅里叶变换 FFT 映射至频域。在频域中,利用可学习的权重矩阵 $\mathbf{W}$ 对频谱进行全局调制,如式(4)所示。

$$\mathbf{X}_{\text{seq}} = \text{F}^{-1}(\text{F}(\mathbf{X}_{\text{patch}}) \odot \mathbf{W}) \quad (4)$$

其中, F 和 F^{-1} 分别表示傅里叶变换和逆傅里叶变换。这种频域门控机制相当于在空域施加了全局感受野的滤波器,有效实现了去噪与语义增强。

经过空频协同处理后的特征图通过 Patch Folding 还原,并再次展平为序列格式。该序列与 MHSA 的输出进行残差连接,经过层归一化后输出,具体实现过程如式(5)所示。

$$\mathbf{Y}_{\text{out}} = \text{LN}(\mathbf{X}_{\text{seq}} + \text{Flatten}(\mathbf{X}_{\text{freq}})) \quad (5)$$

3) MFM 多尺度特征调制融合模块

针对传统特征融合方式平等对待不同尺度特征而忽略了特征在不同通道上表达能力差异的问题,本文引入了 MFM 融合模块,如图 4 所示。MFM 利用全局上下文信息来挖掘通道间的依赖关系,动态生成反映特征重要性的权重系数。通过这些系数对深层语义特征和浅层细节特征进行非线性调制,使模型能够根据输入图像的具体内容,自主抑制无效背景噪声,增强对球团边缘与纹理恢复至关重要的特征分量,从而实现跨尺度信息的高效互补与融合。初始聚合为了获取全局视野并整合跨尺度信息,首先对输入特征进行逐元素求和,生成融合特征 $\mathbf{U}$ 。为了捕获通道维度的全局依赖性,利用全局平均池化将空间信息压缩为通道描述符 $\mathbf{S}$,即:

$$\mathbf{S} = \text{GAP}(\mathbf{U}) = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W \mathbf{U}_{i,j} \quad (6)$$

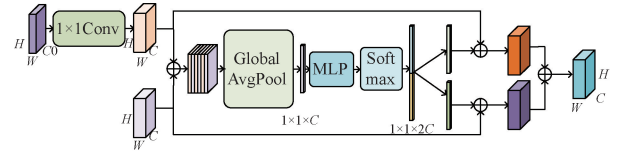


图 4 MFM 融合模块

Fig. 4 MFM fusion module

随后,利用由两个卷积层和 ReLU 激活函数组成的 MLP 网络对通道特征进行编码与映射,并使用 Softmax 函数在分支维度上进行归一化,生成针对每个输入的自适应调制权重 $[a_0, a_1]$ 。

$$[a_0, a_1] = \text{Softmax}(\mathbf{W}_2(\delta(\mathbf{W}_1 \mathbf{s}))) \quad (7)$$

其中, $\mathbf{W}_1$ 和 $\mathbf{W}_2$ 为可学习的权重参数, δ 表示 ReLU 激活函数。 a_0 和 a_1 别代表了 $\mathbf{X}_0$ 和 $\mathbf{X}_1$ 在当前通道下的重要性系数。最后,利用生成的权重对原始输入特征进行逐通道调制,并通过加权求和得到最终的融合输出 $\mathbf{Y}$ 。实现过程如式(8)所示。

$$\mathbf{Y} = \mathbf{X}_0 \odot a_0 + \mathbf{X}_1 \odot a_1 \quad (8)$$

1.2 引入动态感受野的 MobileSAM

为了解决 MobileSAM 在复杂边缘分割中细节丢失的问题,本文提出了 DR 模块。如图 5 所示,该模块通过多尺度空洞卷积捕获上下文信息,并结合 EMA 自适应地聚焦于高频边界区域。

原 MobileSAM 输出的初始掩码与经过上采样的编码器 stage1 的特征图进行拼接。随后送至细化模块。输入特征首先通过三路并行的空洞卷积分支进行多尺度特征提取。为了兼顾局部细节与长距离上下文,本文设置空洞率分别为 $r \in \{1, 2, 4\}$ 。不同感受野下的特征融合过程表示为:

$$\mathbf{U}_k = \text{ReLU}(\text{BN}(\text{Conv}_{r_k}(\mathbf{X}))), k \in \{1, 2, 3\}, r_k \in \{1, 2, 4\} \quad (9)$$

随后,为了生成能够精准定位边缘的注意力图,设计

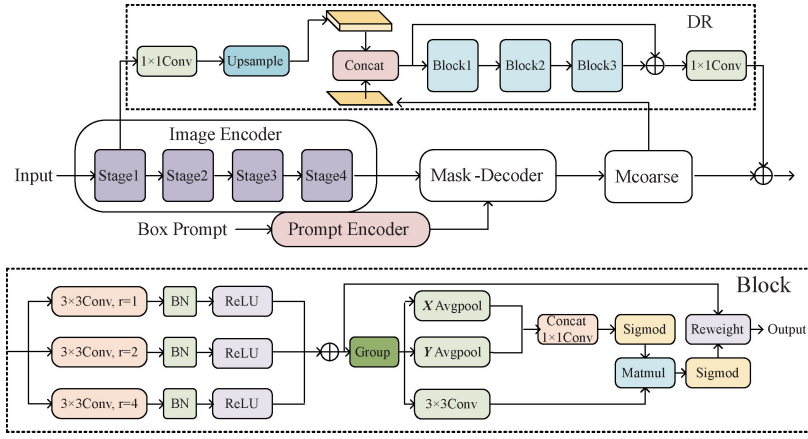


图 5 DR-MobileSAM 整体网络架构

Fig. 5 Overall network architecture of DR-MobileSAM

了双分支交互结构。如图 5 Block 所示,上下文分支利用水平与垂直方向的平均池化捕获长距离的空间坐标依赖;特征分支则通过 3×3 卷积提取局部纹理细节。两者的交互过程为:

$$\mathbf{W}_{\text{attn}} = \sigma(\sigma(\text{Conv}_{1 \times 1}[\text{Pool}_h(\mathbf{F}), \text{Pool}_w(\mathbf{F})]) \otimes \text{Conv}_{3 \times 3}(\mathbf{F})) \quad (10)$$

其中,Conv 代表卷积操作,[.]代表拼接操作。该公式表明,网络首先生成包含全局坐标信息的描述符,随后通过矩阵乘法 matmul 将其注入到局部特征空间中。这种设计利用全局坐标上下文动态地“引导”局部特征的响应。最后,生成的注意力图 $\mathbf{W}_{\text{attn}}$ 对 $\mathbf{F}_{\text{multi}}$ 进行 Reweight 重加权,并通过残差连接输出最终的修正偏差:

$$\mathbf{M}_{\text{final}} = \mathbf{M}_{\text{coarse}} + \text{Conv}_{1 \times 1}(\mathbf{F}_{\text{multi}} \odot \mathbf{W}_{\text{attn}}) \quad (11)$$

该模块遵循 EMA 的轻量化范式,实现了性能与计算成本的平衡。其有效性源于两点:一是分组特征解耦(group-wise decoupling),如图 5 所示,Group 操作将局部纹理提取与全局坐标编码隔离在独立子空间建模。这种设计有效降低了通道间的计算冗余,在大幅压缩参数数量的同时,避免了不同尺度特征的相互干扰。跨维度协同重构:基于低秩重建思想,模块利用 Matmul 实现了空间坐标与局部特征的高阶交互。水平与垂直方向的压缩向量被用于动态校准局部特征响应,从而在保留长距离依赖的同时实现像素级的边缘定位。综上所述,该设计通过少量参数即可实现像素级的特征重加权,确立了网络在复杂场景下对细微边缘的敏锐感知能力,从而高效地完成了对粗糙掩码的精细化修复。

2 实验与结果分析

2.1 数据集

为了验证所提出本框架在真实工业场景中的有效性,本文采集了某大型钢铁企业球团车间的原始生产图像。采集系统安装于球盘上方,使用工业相机在不同光照、堆积状态下共获得 304 张原始图像,每组分辨率 1024×768 。并通过翻转、旋转、噪声扰动等方案扩充至 1520 张。使用

ISAT 标注软件进行多边形标注,并按 8:1:1 的比例划分训练集、验证集与测试集,其中训练集 1215 张、验证集 150 张、测试集 155 张。

2.2 实验参数及评价指标

本文的实验平台为 Ubuntu20.04。硬件配置包括 RTX3090 显卡和 Intel(R) Xeon(R) Gold6326CPU @ 2.90 GHz 处理器。实验使用 Python3.8。基于 PyTorch 深度学习框架,torch 和 CUDA 版本分别是 2.4.1 和 11.8。输入图像为 $[640, 640]$,Epoch 设置为 300,BatchSize 设置为 4,学习率为 0.001,Optimizer 为 AdamW。

本文的指标包括精准度(precision,P)、召回率(recall,R)、平均精度均值(mean average precision,mAP)、以及基于分割实验的平均交并比(mean intersection over union,mIoU)。计算公式如式(12)所示。

$$\begin{cases} P = \frac{TP}{TP + FP} \times 100\% \\ R = \frac{TP}{TP + FN} \times 100\% \\ AP = \int_0^1 P(R) dR \times 100\% \\ mAP = \frac{1}{N} \sum_{i=1}^N AP_i \times 100\% \\ mIoU = \frac{1}{N} \sum_{i=1}^N \frac{TP}{TP + FP + FN} \end{cases} \quad (12)$$

2.3 检测对比实验

为了验证改进版 CEM-DETR 在密集球团检测任务中的性能优势,本文将其与当前主流的目标检测算法进行了对比,包括 YOLO 系列以及基于 Transformer 的 RT-DETR 和 DEIM^[23]。实验结果如表 1 所示。从表 1 可以看出,本文提出的改进算法在各项核心指标上均取得了最优性能。改进后的模型 mAP@0.5:0.95 达到 91.8%,P 为 96.5%,R 为 93.1%,相较于基线 RT-DETR 分别提升了 2.7%、1.6%、2.5%,且显著优于 YOLO11n 和

YOLOv10n。CEM-DETR 架构凭借端到端的预测范式，展现了更强的抗遮挡能力，证明了其作为球团检测基准模型的合理性。

2.4 分割对比实验

对比经典实例分割算法及最新级联框架的实验结果如表 2 所示,本文方法的 mAP@0.5 达到 97.4%,mIoU 达到 90.1%。R 高达 95.7%。为了全面验证本文提出框架在球团实例分割任务中的性能,本文将其与 Mask R-CNN、YOLACT、Mask2Former 以及文献[9]进行了对比实验。可视化轮廓效果对比如图 6 所示。

Mask R-CNN 展现了较高的分割精度,P 为 94.0%,但受限于非极大值抑制后处理,在高重叠区域容易抑制

表 1 检测对比实验

Table 1 Comparative experiments on detection performance (%)

模型	P	R	mAP@0.5	mAP@0.5:0.95
FastRCNN	87.4	82.3	86.3	80.1
YOLOv8n	90.7	86.5	88.6	83.4
YOLOv10n	94.4	90.9	92.6	88.1
YOLO11n	94.1	91.3	93.6	88.3
DEIM	93.6	90.8	91.7	88.7
RT-DETR	94.9	90.6	93.8	89.1
本文	96.5	93.1	96.4	91.8

表 2 分割对比实验

Table 2 Comparative experiments on segmentation performance

模型	mIoU	P/%	R/%	mAP@0.5/%	FPS/fps	Param/($\times 10^6$)
Mask R-CNN	94.0	99.0	84.6	84.0	17.8	42.00
文献[7]	94.8	93.6	85.9	93.1	21.3	43.60
YOLACT	58.7	93.2	82.3	86.6	19.2	47.45
Mask2Former	59.1	32.0	32.9	18.7	16.5	127.30
文献[9]	83.5	52.9	58.4	58.0	13.2	89.42
YOLOv11seg	86.2	95.1	93.7	95.7	28.3	2.90
文献[11]	90.3	94.6	87.7	90.6	14.7	92.17
YOLO11n+SAM2	91.5	93.7	91.2	94.6	1.8	244.10
YOLO11n+SAM3	91.7	93.9	91.7	95.1	1.3	850.20
RT-DETR+SAM2	91.2	94.5	94.0	96.2	1.2	265.50
RT-DETR+SAM3	92.7	94.3	92.6	96.0	0.3	885.50
RT-DETR+MobileSAM	73.6	83.5	81.1	88.7	13.7	30.15
本文	90.1	97.7	95.7	97.4	14.6	29.14

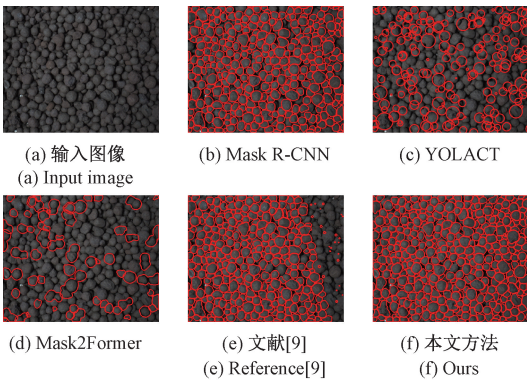


图 6 不同模型的分割效果

Fig. 6 Segmentation results of different models

有效目标框,导致召回率仅为 84.6%,存在明显漏检现象,无法准确统计球团总数。文献[7]虽然 P 较高,mAP@0.5 比原始 Mask R-CNN 有显著提升,但其 R 仅为 85.9%,仍

然无法突破密集检测的瓶颈。YOLACT 生成的掩码质量较低,mIoU 仅为 58.7%。由于其采用原型掩码线性组合的方式生成实例,对于球团边缘的细微几何特征捕捉不足。Mask2Former 在本任务中表现最差,mAP@0.5 仅为 18.7%,R 为 32.9%,几乎不能完成复杂场景的粒径检测任务。

除了上述经典实例分割模型,本文还重点评估了基于检测器引导分割范式的最新大模型级联方案,并特别对比了针对球团场景优化的文献[11]方法,其可视化检测与分割效果如图 7 所示。从表 2 数据可以看出,得益于 SAM2 以及 SAM3 强大的零样本泛化能力,这类组合模型在 mIoU 上达到了 91.5%、91.7%、91.2%、92.7% 的极高精度,略高于本文方法。然而,这种精度的微弱优势是以巨大的计算代价为交换的:即便是文献[11]这类改进方案,也难以摆脱大模型底座的高算力束缚。以 YOLO11n+SAM2 为例,其参数量高达 244.1×10^6 ,是本文方法的

8 倍以上,而推理速度仅为 1.8 fps,远低于工业现场实时监测通常要求的 10 fps。这意味着此类重型 SOTA 方案

虽然效果惊艳,但极高的算力门槛使其难以在资源受限的工业边缘设备上实际部署。

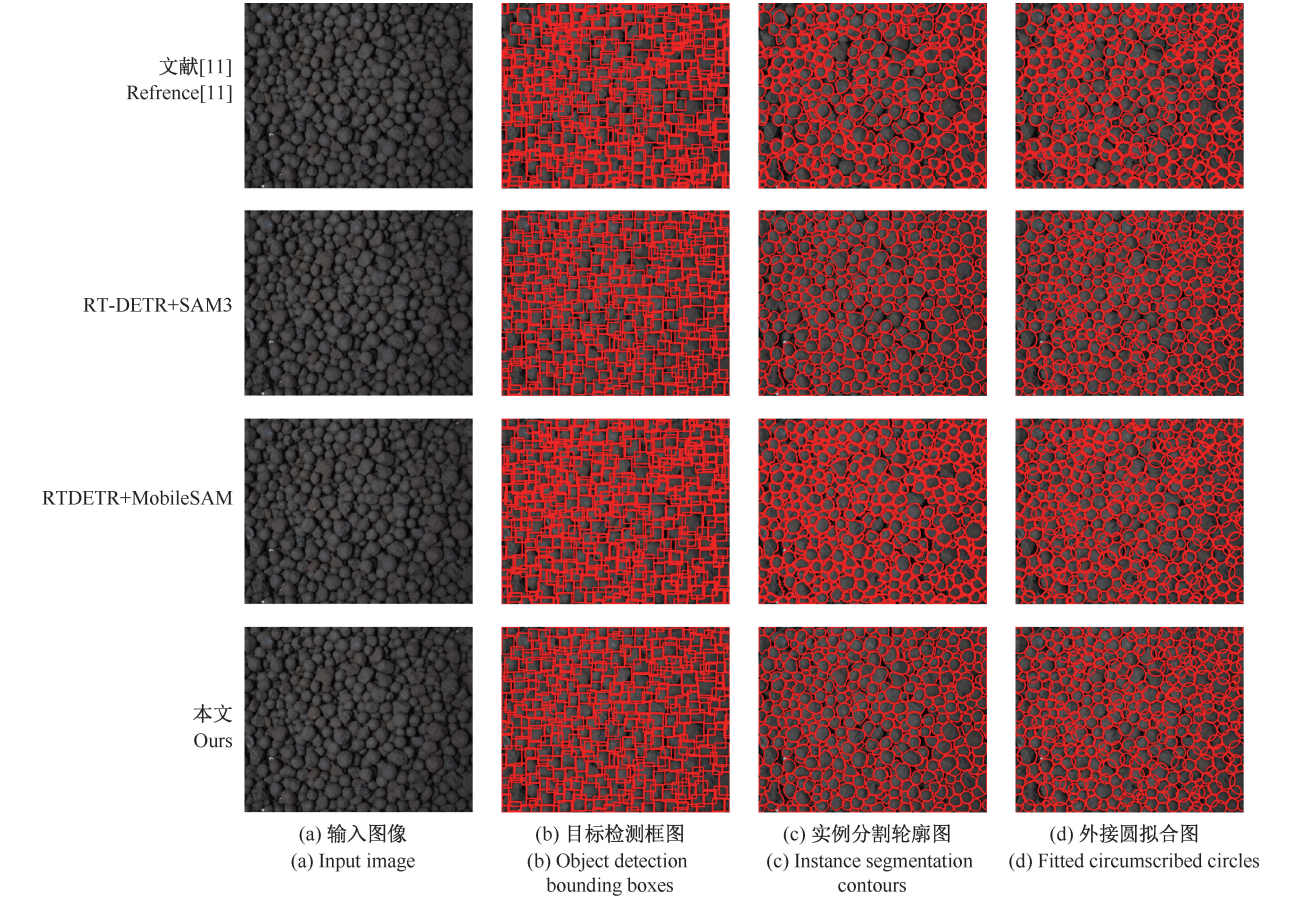


图 7 级联算法检测及分割效果

Fig. 7 Detection and segmentation results of the cascade algorithms

相比之下,本文提出的级联框架在高精度与高效率取得了最佳平衡,mAP@0.5 达到 97.4%,mIoU 达到 90.1%。改进后的检测器通过在骨干网络嵌入坐标注意力、在编码器引入高效判别频域前馈网络以及采用多尺度特征调制模块,大幅增强了对密集堆叠球团的感知能力,有效解决了漏检痛点,R 高达 95.7%。针对分割边界模糊问题,本文引入的 DR 细化器对 MobileSAM 生成的初始掩码进行自适应边缘修复。这一创新设计有效捕捉并恢复了球团边缘的高频几何特征,生成的掩码边缘平滑、轮廓清晰。与原 RT-DETR 引导 MobileSAM 的基准方案相比,本文方法在保持推理速度基本持平的前提下,将 mIoU 提升了 16.5%,完美解决了原版模型因提示框松散和细节丢失导致的欠分割难题。综上所述,本文方法不仅在统计指标上显著优于现有 SOTA 模型,在可视化效果上也更贴近球团的真实物理轮廓,为后续的粒径精确测量奠定了坚实基础。

2.5 消融实验

如表 3 所示,基于自建球团数据集的消融实验验证了

改进模块的有效性。原 RT-DETR 的 mAP@0.5:0.95 为 89.1%,参数量为 19.87×10^6 。首先引入 CA 坐标注意力后,mAP 提升至 90.6%,R 升至 92.8%,证明位置编码增强了对密集目标的定位能力。随后加入 AIFI-EDFFN,R 进一步增至 93.0%,表明空频协同机制有效改善了对模糊边缘的捕捉。最后引入 MFM 模块,模型整体 mAP 最终跃升至 91.8%,相比基线提升了 2.7%,P 与 R 分别达到 96.5%和 93.1%。关键在于,模型的总参数量降至 19.66×10^6 ,低于原始模型,证明该方案在显著提升检测精度的同时有效降低了模型复杂度,具有极高的工业应用性价比。

为了验证 CA 在球团矿密集检测任务中的优越性,并在骨干网络中寻找最佳的注意力嵌入方案,本文在保证其他模块 EDFFN、MFM 不变的前提下,分别将 ResNet18 中的 CABasicBlock 替换为原 BasicBlock 以及嵌入了 SE^[24]、ECA^[25]、CBAM^[26]的变体模块进行对比实验。实验结果如表 4 所示。

为了验证改进模块对级联框架的贡献,本文以原 RT-

表 3 CEM-DETR 消融实验

Table 3 Ablation study on CEM-DETR

RT-DETR	CA	EDFFN	MFM	$P/\%$	$R/\%$	mAP@0.5;0.95/%	Para/($\times 10^6$)
✓				94.9	90.6	89.1	19.87
✓	✓			95.6	92.8	90.6	19.94
✓	✓	✓		95.1	93.0	90.6	20.03
✓	✓	✓	✓	96.5	93.1	91.8	19.66

表 4 不同注意力机制对检测指标的影响

Table 4 Impact of different attention mechanisms on detection metrics (%)

注意力	P	R	mAP@0.5	mAP@0.5;0.95
None	93.5	89.2	92.1	88.7
SE	94.1	90.1	92.8	89.2
ECA	94.0	89.8	92.6	87.5
CBAM	94.8	91.5	93.4	90.6
CA	96.5	93.1	96.4	91.8

DETR 结合 MobileSAM 为基准进行了逐级消融实验,如

表 5 所示。在检测器改进阶段(引入 CA、EDFFN 及 MFM),模型 mIoU 从 73.6%稳步提升至 75.2%, R 增长至 85.3%。这一增益证明了特征增强策略能有效生成更紧致的边界框提示,确立了检测器作为精准引导者减少背景噪声的核心价值。而在引入 DR 动态感受野模块后,模型性能实现了质的飞跃:mIoU 大幅跃升至 90.1%, P 与 R 分别提升至 97.7%和 95.7%,较基线提升 14.6%。这一显著的数据跳变表明,DR 模块成功在粗引导的基础上完成了对球团粘连边缘的像素级精修。综上所述,该架构凭借 NMS-free 无漏检引导与动态细化器保障边缘精度的深层协同,证明了其作为球团粒径测量基准方案的优越性。

表 5 级联消融实验

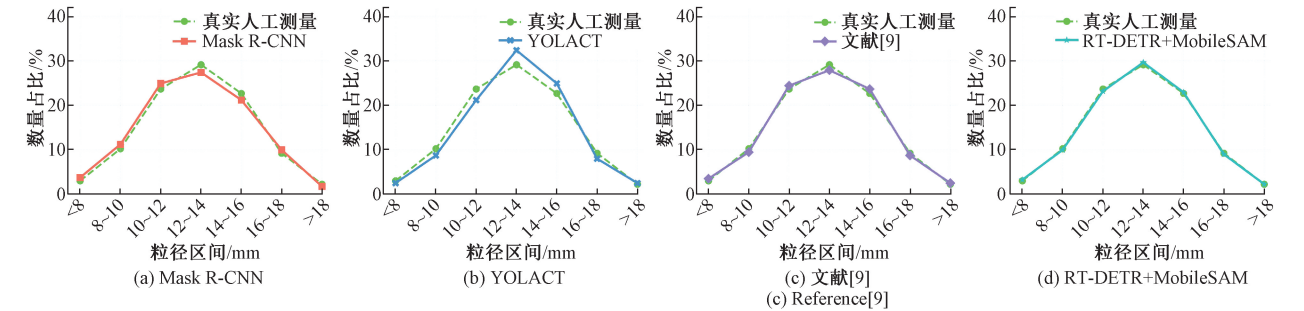
Table 5 Ablation study on the overall cascaded framework

RT-DETR+MbileSAM	CA	EDFFN	MFM	DR	mIoU	$P/\%$	$R/\%$	mAP@0.5/%
✓					73.6	83.5	81.1	80.5
✓	✓				74.1	86.8	82.9	82.6
✓	✓	✓			74.1	90.4	83.1	83.9
✓	✓	✓	✓		75.2	89.6	85.3	84.1
✓	✓	✓	✓	✓	90.1	97.7	95.7	97.4

2.6 检测准确率

为了进一步验证模型在实际工业生产中的计量精度,本文选取包含 402 个球团的典型现场图像进行粒度分布统计实验。实验结果如图 8 所示,将人工测量结果作为基准,计算各算法在不同粒径区间的分布偏差。从图 8(a)~(c)可以看出,Mask R-CNN、YOLACT 以及文献[9]等传统方法在峰值区间(12~14 mm)均存在显著的统

计偏差,平均相对误差分别为 6.89%、7.24%和 4.86%,难以准确反映真实的粒度分布。相比之下,基于先进级联范式的 SOTA 方案,YOLOv11+SAM3、RT-DETR+SAM3,以及针对球团优化的文献[11],由图 8(e)~(g)可以看出,其预测曲线与真值高度吻合,证明了检测引导分割策略在计数任务上的优越性。值得注意的是,本文提出的级联框架,即图 8(h)在保持轻量级部署优势的同时,也达到了与



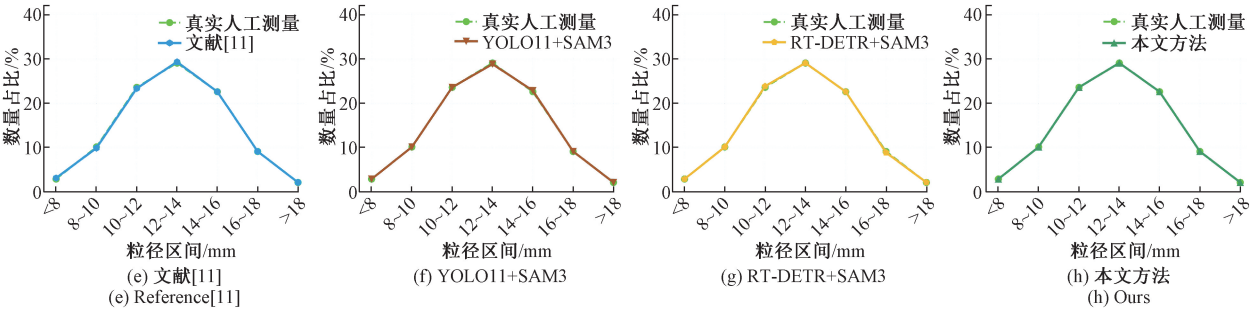


图 8 粒径检测比例
Fig. 8 Line chart of particle size detection ratio

上述重型 SOTA 模型同等的统计准确性。实验表明,本文方法的平均相对误差仅为 1.06%,且在样本总数预测上仅偏差 2 个实例。特别是在 8~18 mm 的合格粒径区间内,本文方法有效克服了密集堆叠带来的计数干扰,证明了其作为一种兼顾“高精度”与“高效率”的方案,完全满足工业现场对球团粒度在线高精度监测的需求。

3 结 论

本文提出了一种融合 CEM-DETR 检测器与 MobileSAM 的级联实例分割框架,为球团质量的数字化监控提供了高精度与高效率并重的解决方案。针对小目标定位困难与特征混叠问题,本文通过在 RT-DETR 骨干网络中嵌入坐标注意力机制以增强空间感知能力,其次,在 AIFI 编码器中构建 EDFFN,最后,引入 MFM 替代传统融合方式,通过自适应加权有效解决了跨尺度特征的语义错配问题。该框架利用基于动态感受野以及 EMA 的细化器的 MobileSAM 进行并行掩码生成,实现了从定位到分割的端到端性能跃升。实验结果表明,该方法在分割精度(mAP@0.5 达到 97.4%)与召回率(95.7%)上显著超越了经典主流模型以及级联模型,有效克服了传统算法在密集粘连区域普遍存在的漏检严重与边缘锯齿化瓶颈。更重要的是,与当前基于 SAM3 的最新大模型级联方案相比,本文方法在保持相当分割精度的前提下,推理速度提升了近 8 倍,参数量仅为其 1/8,实现了检测精度与工业实时性的最佳平衡。此外,基于该框架的粒径统计平均相对误差仅为 1.06%,验证了其在复杂工业现场极高的可靠性与应用价值。未来工作将致力于模型的轻量化剪枝与量化策略研究,旨在进一步降低计算开销,在嵌入式边缘设备上实现更高效的在线部署与应用。

参考文献

[1] 林双,陆伟文,胡守景,等. 结合 Mask-RCNN 和最小二乘法的球团粒度识别模型[J]. 钢铁研究学报,2024, 36(7):855-865.
Lin Shuang, Lu Weiwen, Hu Shoujing, et al. Pellet granularity recognition model combining Mask-RCNN and least squares method[J]. Journal of Iron and Steel

Research, 2024, 36(7): 855-865.
[2] 陈哲,周琳,李强. 集装箱载铁矿石智能机械取样设备系统探究[J]. 质量与认证,2024(6):42-44.
Chen Zhe, Zhou Lin, Li Qiang. Research on intelligent mechanical sampling equipment system for containerized iron ore[J]. Quality and Certification, 2024(6): 42-44.
[3] 何赞泽,郭猜,郭隆强,等. 基于视觉语义约束的激光雷达大目标尺寸测量方法[J]. 仪器仪表学报,2025, 46(8):244-254.
He Yunze, Guo Cai, Guo Longqiang, et al. Large target size measurement method of LiDAR based on visual semantic constraints[J]. Chinese Journal of Scientific Instrument, 2025, 46(8): 244-254.
[4] 吴鑫,李靓杰,魏建好. 融合双重注意力与多尺度级联网络的密集球团粒度测量方法[J]. 控制与决策,2025, 40(10):3155-3166.
Wu Xin, Li Liangjie, Wei Jianhao. Dense pellet granularity measurement method fusing dual attention and multi-scale cascade network [J]. Control and Decision, 2025, 40(10): 3155-3166.
[5] 张上,朱帅,张岳. 基于改进 RT-DETR 的转向节表面缺陷检测算法[J]. 电子测量技术,2026, 49(2): 230-241.
Zhang Shang, Zhu Shuai, Zhang Yue. Surface defect detection algorithm for steering knuckle based on improved RT-DETR [J]. Electronic Measurement Technology,2026,49(2):230-241.
[6] 张学锋,秦继洋,黄刘松,等. 基于图像增强及改进霍夫检测算法的球团粒度检测[J]. 中国冶金,2024,34(8): 59-68.
Zhang Xuefeng, Qin Jiyang, Huang Liusong, et al. Pellet granularity detection based on image enhancement and improved Hough detection algorithm [J]. China Metallurgy, 2024, 34(8): 59-68.
[7] 王猛,刘卫星,李喆,等. 融合 CBAM 的 Mask R-CNN 模型在球团识别与粒径测量中的应用[J]. 烧结球团, 2025,50(1):85-94, 125.
Wang Meng, Liu Weixing, Li Zhe, et al. Application of Mask R-CNN model fusing CBAM in pellet

- recognition and particle size measurement [J]. Sintering and Pelletizing, 2025, 50(1): 85-94, 125.
- [8] 任杰,迟荣华,李红旭. 基于多尺度特征提取和注意力机制的轻量化晶圆缺陷检测方法[J]. 电子测量与仪器学报, 2025, 39(8): 13-21.
- Ren Jie, Chi Ronghua, Li Hongxu. Lightweight wafer defect detection method based on multi-scale feature extraction and attention mechanism[J]. Journal of Electronic Measurement and Instrumentation, 2025, 39(8): 13-21.
- [9] 马伟宁,杨磊,李杰,等. 基于 SAM 的球团图像粒度识别算法[J]. 烧结球团, 2025, 50(2): 54-62, 86.
- Ma Weining, Yang Lei, Li Jie, et al. Pellet image granularity recognition algorithm based on SAM[J]. Sintering and Pelletizing, 2025, 50(2): 54-62, 86.
- [10] Kirillov A, Mintun E, Ravi N, et al. Segment anything[C]//IEEE/CVF International Conference on Computer Vision, 2023: 4015-4026.
- [11] 龚玉洁,王旭,郭海杰,等. 基于视觉联合检测的球团矿粒径检测方法[J]. 电子测量技术, 2026, 49(2): 203-211.
- Gong Yujie, Wang Xu, Guo Haijie, et al. Pellet particle size detection method based on visual joint detection[J]. Electronic Measurement Technology, 2026, 49(2): 203-211.
- [12] Wang Y N, Fei Z H, Li R C, et al. Learn from foundation model: Fruit detection model without manual annotation[J]. Pattern Recognition, 2025, 174: 112799.
- [13] Mansoori M, Shahabodini S, Abouei J, et al. Self-prompting polyp segmentation in colonoscopy using hybrid yolo-sam2 model[C]//2025 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), 2025: 1-5.
- [14] Chen K Y, Liu C Y, Chen H, et al. RSPrompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 1-17.
- [15] Zhao Y A, Lyu W Y, Xu S, et al. DETRS beat YOLOs on real-time object detection[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 16965-16974.
- [16] Zhang C N, Han D S, Qiao Y, et al. Faster segment anything: Towards lightweight sam for mobile applications[PP/OL]. V2. [2026-06-30]. <https://arxiv.org/abs/2306.14289>.
- [17] Hou Q B, Zhou D Q, Feng J S. Coordinate attention for efficient mobile network design[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 13713-13722.
- [18] Kong L S, Dong J X, Tang J C, et al. Efficient visual state space model for image deblurring[C]//Computer Vision and Pattern Recognition Conference, 2025: 12710-12719.
- [19] Zhang Y F, Zhou S, Li H F. Depth information assisted collaborative mutual promotion network for single image dehazing[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 2846-2855.
- [20] Li X, Wang W H, Hu X L, et al. Selective kernel networks[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 510-519.
- [21] Ouyang D Z, He S H, Zhang G Z, et al. Efficient multi-scale attention module with cross-spatial learning[C]//2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023: 1-5.
- [22] Kong L S, Dong J X, Ge J J, et al. Efficient frequency domain-based transformers for high-quality image deblurring [C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 5886-5895.
- [23] Huang S H, Lu Z C, Cun X D, et al. DEIM: DETR with improved matching for fast convergence[C]//Computer Vision and Pattern Recognition Conference, 2025: 15162-15171.
- [24] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//IEEE Conference on Computer Vision and Pattern Recognition, 2018: 7132-7141.
- [25] Wang Q L, Wu B G, Zhu P F, et al. ECA-Net: Efficient channel attention for deep convolutional neural networks [C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 11534-11542.
- [26] Woo S, Park J, Lee J Y, et al. CBAM: Convolutional block attention module[C]//European Conference on Computer Vision(ECCV), 2018: 3-19.

作者简介

郭海杰, 硕士研究生, 主要研究方向为目标检测、图像分割。

E-mail: 15890354319@163.com

崔雪红(通信作者), 博士, 副教授, 主要研究方向为人工智能、计算机视觉。

E-mail: 19853226869@163.com

龚玉洁, 硕士研究生, 主要研究方向为目标检测、图像分割。

E-mail: gongyujie0303@163.com

王旭, 硕士研究生, 主要研究方向为目标检测。

E-mail: 17860273050@163.com

谢永琪, 硕士研究生, 主要研究方向为图像分割。

E-mail: 15628913880@163.com