文章编号　1004-924X(2026)11-1791-16

基座模型分割先验的列车故障轻量化检测

孙国栋[1,2*]，梁启航[1]，邹鹏坤[1]，李郭钰[1]，潘星宇[1]

(1. 湖北工业大学 机械工程学院，湖北 武汉 430068；
2. 湖北工业大学 现代制造质量工程湖北省重点实验室，湖北 武汉 430068)

摘要：货运列车的故障检测对于保障铁路运输安全和提升运营效率至关重要。然而，受限于高昂的计算成本，视觉基座模型难以直接部署于资源受限的轨旁检测设备。针对这一问题，提出一种融合特征知识迁移、注意力增强及逻辑感知蒸馏的轻量化故障检测框架，旨在将基座模型的先验知识高效地迁移至可部署模型。首先，利用多源预训练策略，结合FastSAM的语义表征优势与YOLOE的检测能力，构建了特征融合的学生网络模型。其次，在骨干网络中引入轻量化特征增强模块，以提升模型在复杂视觉工况下的特征提取与表征能力。最后，设计了一种逻辑感知多组件蒸馏策略，将教师模型的丰富知识压缩至学生模型，在保持低计算成本的同时显著提升检测精度。在自建货运列车故障数据集上的实验结果表明，该方法在大幅降低模型参数量与浮点运算量的同时，实现了具有竞争力的检测精度，展现了在边缘侧设备进行实时部署的巨大潜力。

关　键　词：计算机视觉；故障检测；知识蒸馏；知识迁移；实时检测

中图分类号：TP391.4　　**文献标识码**：A

doi：10.37188/OPE.20263411.1791　　**CSTR**：32169.14.OPE.20263411.1791

Prioritization of base model segmentation for light weight detection of train faults

SUN Guodong[1,2*], LIANG Qihang[1], ZOU Pengkun[1], LI Guoyu[1], PAN Xingyu[1]

(1. *School of Mechanical Engineering, Hubei University of Technology, Wuhan* 430068 *China*;
2. *Hubei Key Laboratory of Modern Manufacturing Quality Engineering, Hubei University of Technology, Wuhan* 430068, *China*)
* *Corresponding author, E-mail: sgdeagle*@163.*com*

Abstract: Fault detection in freight trains is essential for ensuring railway transportation safety and improving operational efficiency. However, the high computational cost of vision foundation models remains a major barrier to their direct deployment on resource-constrained trackside equipment. To address this challenge, a lightweight fault detection framework integrating feature knowledge transfer, attention enhancement, and logic-aware distillation is proposed to efficiently transfer prior knowledge from foundation models to deployable networks. Specifically, a fused student network is first constructed using a multi-source pretraining strategy that combines the semantic representation strengths of FastSAM with the robust detection capability of YOLOE. Subsequently, a lightweight feature enhancement module is embedded into the

收稿日期：2026-03-11；**修订日期**：2026-04-06.

基金项目：国家自然科学基金资助项目(No. 51775177)

backbone network to improve feature extraction and representation under complex visual conditions. Finally, a logic-aware multi-component distillation strategy is designed to compress the rich knowledge of the teacher model into the student network, thereby significantly improving detection accuracy while maintaining low computational cost. Experimental results on a self-constructed freight train fault dataset demonstrate that the proposed method achieves competitive detection accuracy with substantial reductions in both parameters and floating-point operations, indicating strong potential for real-time deployment on edge devices.

Key words: computer vision; fault detection; knowledge distillation; knowledge transfer; real-time detection

1 引　言

近年来,货运列车智能故障检测已成为保障现代铁路运营安全[1]、稳定及高效的关键环节。随着运输网络的快速扩张及物流自动化需求的日益增长,传统的人工列检方式已无法满足实时安全监测与维护的需求。因此,基于深度学习的视觉故障检测技术在货运列车列检系统中得到了广泛应用,实现了结构缺陷及组件异常的端到端自动识别。这些方法在精度与可扩展性方面均展现出良好的应用前景[2]。

尽管深度学习模型[3]取得了显著成效,但将其部署于轨旁或边缘设备等实际工业场景仍面临严峻挑战。近年来,得益于海量数据与深层网络架构,分割一切模型(Segment Anything Model,SAM)[4]等大规模模型及实时视觉感知模型(YOLOE)[5]等高容量目标检测器,在各类通用视觉任务中表现卓越。然而,将此类模型直接应用于工业环境仍存在诸多困难[6]。快速分割一切模型(Fast Segment Anything Model,FastSAM)[7]与 YOLOE 均为 YOLOv8(You Only Look Once version 8)[8]的变体,继承了大型 YOLOv8 模型计算量大、资源消耗高的缺点。极高的计算与内存需求使 FastSAM 难以部署在嵌入式或资源受限平台上。大型 YOLOE[8]变体(如 YOLOE-v8l)则存在推理延迟增加与硬件成本上升的问题。因此,如何在算力受限的边缘侧设备上,兼顾对大型基座模型丰富语义先验的高效融合,与低延迟、高精度的实例级故障定量检测,已成为当前亟待解决的关键难题。

由于现有方法缺乏对货运列车特定组件进行定量分析的能力(如评估闸瓦磨损程度等),故障诊断的全面性受到限制。为解决上述问题,本文提出了一种面向复杂视觉环境下货运列车故障诊断的轻量化检测框架(LAD-YOLO),利用逻辑感知蒸馏技术(Logic-Aware Distillation,LAD)融合基座模型的先验知识。该方法可将高容量模型的知识压缩至紧凑的可部署结构中。首先,通过整合 FastSAM 的骨干网络与 YOLOE-v8l 的检测头,构建了专为学生网络设计的预训练模型。这一融合预训练模型提供了稳定的蒸馏信号并加速了训练收敛,从而提升了轻量化模型的优化效率。然后,在学生模型中嵌入轻量化注意力增强模块,以提高其特征表征能力,特别是针对复杂光照、遮挡及背景干扰等场景。最后,采用基于二元交叉熵的蒸馏损失函数,将教师模型的语义与空间知识压缩至学生网络,在有限计算预算下实现高检测精度。

2 相关工作

2.1 货运列车图像故障检测

货运列车故障检测[9-11]是智能交通领域的关键课题。针对该任务,现有研究在提升特征表征与模型轻量化方面进行了大量探索。Li 等[12]结合一维卷积与生成对抗网络处理振动信号,缓解了小样本下的过拟合问题。Liu 等[13]则面向资源受限的边缘设备,提出融合多层特征的超微型分类网络以保障诊断精度。然而,现有方法主要存在两点局限:其一,过度追求极致轻量化,导致模型未能充分利用当前轨道交通日益升级的边缘计算硬件,造成算力浪费;其二,工程泛化能力不足,多数研究仅局限于特定故障的实验室场景,

缺乏在复杂工业环境及多类别真实数据集上的交叉验证。

随着视觉列检技术的发展，深度学习驱动的实例分割逐渐成为主流手段。早期如 Mask R-CNN[14]及结合 PointRend[15]的方法奠定了高精度掩码预测的基础，而 YOLACT[16]，SOLOv2[17]与 CondInst[18]等单阶段或无锚框（Anchor-free）设计则大幅降低了模型复杂度，为实时部署创造了条件。进入 Transformer 时代后，QueryInst[19]，SparseInst[20]及 Mask2Former[21]等模型通过引入查询机制与稀疏表征，进一步增强了复杂场景下的特征建模与任务泛化能力。此外，专为工业级实时部署设计的 RTMDet[22]在平衡高精度与低计算开销方面提供了极具参考价值的方案。

本文提出一种专为货运列车定制的轻量化视觉列检框架。该框架有效融合了高级语义知识与稳定的特征提取策略，在控制计算负担的前提下，不仅实现了复杂环境下多类故障的精准识别，还支持对关键部件磨损程度的像素级定量分析，如图 1 所示。

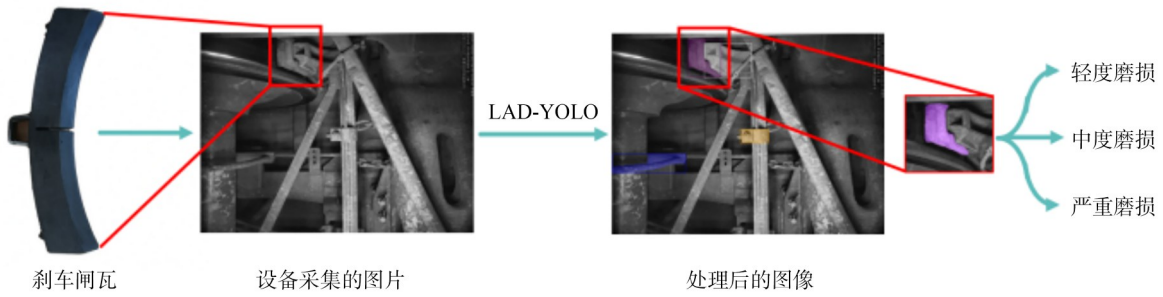


图 1　闸瓦磨损程度检测过程示意图

Fig. 1　Schematic diagram of detection process for brake shoe wear level

2.2　特征增强注意力机制

注意力机制赋予了模型动态聚焦显著特征的能力。经典的通道注意力（如 SENet[23]）和空间注意力（如 CBAM[24]）被广泛采用，但 CBAM 对人工设计池化操作和固定结构的依赖限制了其轻量化适应性。尽管 ECA-Net[25]和 SimAM[26]等方法简化了结构，但仍将通道与空间显式分离，未能充分挖掘两者的交互关系。基于 Transformer 的注意力机制（如 MobileViT[27]，TinyViT[28]）虽能有效建模长距离依赖，但其自注意力计算带来的高昂开销使其难以满足边缘部署等资源受限场景的实时性要求。

针对上述局限，这里提出了轻量化卷积注意力融合（Lightweight Channel-Aware Fusion network，LCAF）模块。LCAF 利用可学习的卷积操作（如 1×1 卷积与深度可分离卷积 DWConv）替代传统的全局池化，在极低参数量下消除了池化瓶颈。相比现有方法，LCAF 不仅高效实现了空间与通道跨维度的联合建模，还与轻量化检测及分割网络展现出高度的兼容性。

2.3　基座模型

基座模型（如 Vision Transformer[29]，CLIP[30]，Qwen-VL[31]，ChatGPT[32]及 Segment Anything Model）基于大规模数据预训练，在多项下游任务中展现出强大的泛化能力，深刻重塑了视觉理解领域的格局。其中，通用分割模型 SAM 备受关注[33]，其基于提示生成高质量分割结果的能力，以及在开放集分割和零样本泛化上的优势，使它作为强大的视觉先验广泛应用于医学、航空及自动驾驶等领域。

在货运列车故障检测中，SAM 等模型展现出巨大潜力：其稳定的特征表征能适应高噪声与复杂背景，对细粒度边界及上下文的捕捉亦有助于识别传统 CNN 难以察觉的微小异常。然而，基座模型的资源密集型特征限制了它在轨旁或车载等边缘端的部署；同时，面向通用任务的设计使其缺乏对工业特定场景（如小目标检测、抗遮挡干扰）的自适应能力。

为解决上述部署与适配难题，现有研究尝试通过微调、提示工程或特征融合[34]等策略将基座模型适配至下游任务。然而，全量微调成本高

昂，浅层适配又难以充分发挥其表征潜力。鉴于此，本文提出一种选择性知识迁移策略，融合 FastSAM 的预训练骨干网络与 YOLOE 的检测头权重。该方法在保留基座模型高级语义表征优势的同时，显著降低了计算负担，从而更好地满足实时货运列车检测系统的部署需求。

2.4　知识蒸馏

知识蒸馏[35]（Knowledge Distillation，KD）旨在将大型教师模型的知识迁移至轻量化的学生网络中，使其拟合教师的软输出或中间特征表征。该技术现已成为边缘设备和实时推理等资源受限场景下，轻量化模型设计的核心策略。

近年来，Logit 级蒸馏因其架构无关性及易于集成到单阶段检测流程中的特点，已成为轻量化检测器的重要解决方案。例如，定位蒸馏（Localization Distillation，LD）通过拟合 Logit 有效迁移定位知识，在 COCO[36]基准上展现出超越特征级方法的潜力。解耦知识蒸馏（Decoupled Knowledge Distillation，DKD）[37]将 KD 损失分解为目标与非目标类分量，提升了分类与检测的灵活性与性能。类别感知 Logit 知识蒸馏（Class-Aware Logit Knowledge Distillation，CLKD）[38]则通过整合类别相关性，进一步丰富了语义知识的迁移效果。

相比之下，基于特征的蒸馏通常需在师生网络间建立复杂的对齐或维度匹配模块[39]，这不仅增加了训练复杂度，还可能削弱轻量化带来的效率增益。此外，特征级信号的任务针对性较弱，易引入冗余。例如，尽管 SO-DETR[40]在 Transformer 检测器中应用 Logit 蒸馏取得了显著改进，但由于未能充分利用其他检测输出，仍存在知识冗余与浪费的局限性。

3　原　理

3.1　网络总体框架

本文提出的框架旨在通过统一架构实现兼具高精度与高效率的目标检测及实例分割多任务预测。如图 2 所示，该框架集成了跨模型预训练权重迁移策略、轻量化通道感知融合模块、逻辑感知多组件蒸馏机制，以及适配自 YOLOE 的分割头。输入图像首先经由使用 FastSAM 权重初始化的骨干网络处理，以提取语义丰富且具备边界感知能力的特征。随后，这些特征被送入源自 YOLOE 的检测头与分割头，各部分均采用其对应的预训练权重进行初始化。在特征传播过程中，轻量化通道感知融合模块在通道与空间维度上自适应地增强中间特征表征，从而提升学生模型的特征感知能力。

在训练阶段，学生网络通过逻辑感知多组件蒸馏机制，接收来自教师模型的额外多级监督信号。该监督涵盖多个预测分支，包括边界框回归、类别分类、目标置信度估计以及掩码生成。通过在不同粒度层级上将学生网络的输出与教

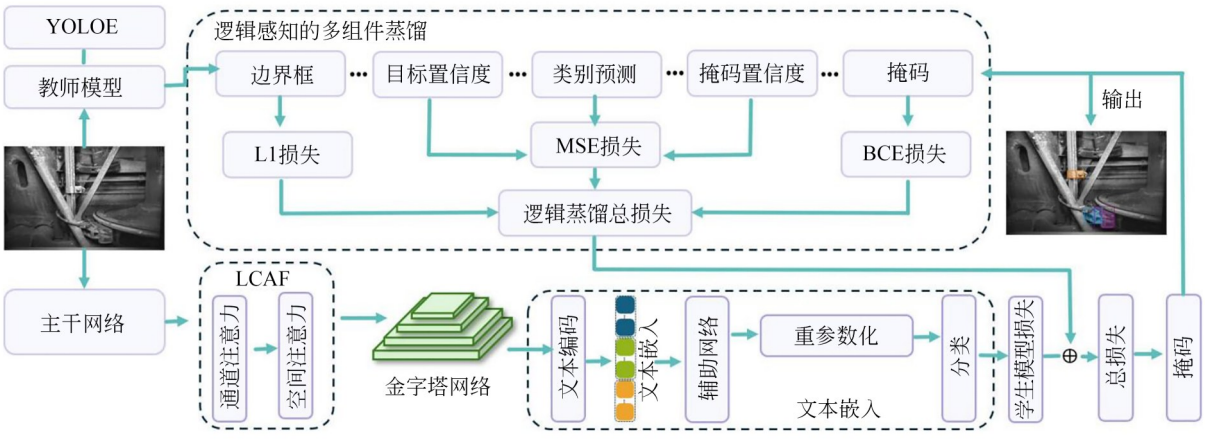


图 2　LAD-YOLO 整体架构

Fig. 2　Overall framework of LAD-YOLO

师模型的预测进行对齐，该框架不仅使学生网络能够复现教师的预测结果，更能捕捉其潜在的决策逻辑。这种集成化设计充分发挥了检测与分割任务的互补优势，在保持轻量化与高计算效率架构的同时，显著提升了多任务预测的精度与泛化能力。

3.2　跨模型预训练权重迁移

现有的先进检测与分割模型往往展现出互补的优势。YOLOE表现出强大的目标定位性能及高效的面向检测特征提取能力，而FastSAM凭借其专用的骨干网络设计，在边界保持与细粒度分割方面表现优异。然而，若从零开始训练此类模型或仅使用单一预训练源，可能会因某单一任务的归纳偏置占据主导地位，从而导致模型的泛化能力欠佳。

针对这一问题，本文提出一种跨模型预训练权重迁移学习方法，在架构组件层级融合YOLOE与FastSAM的预训练权重，如图3所示。鉴于两者均源于YOLOv8架构，其在层级化特征提取逻辑与特征图尺度上具备天然的结构兼容性，但在颈部聚合逻辑上存在功能性差异：FastSAM侧重于全场景语义聚合，而YOLOE则通过重参数化实时聚合器（Re-parameterizable Region-Text Alignment，RepRTA）模块强化了重参数化的任务相关特征表征。具体而言，保留了FastSAM的骨干网络权重，以维持其语义丰富性与轮廓敏感度，这对于生成高质量掩码至关重要。同时，保留YOLOE中经RepRTA增强的检测头及其分割头权重，以确保稳定的目标定位能力与检测驱动的特征表征。

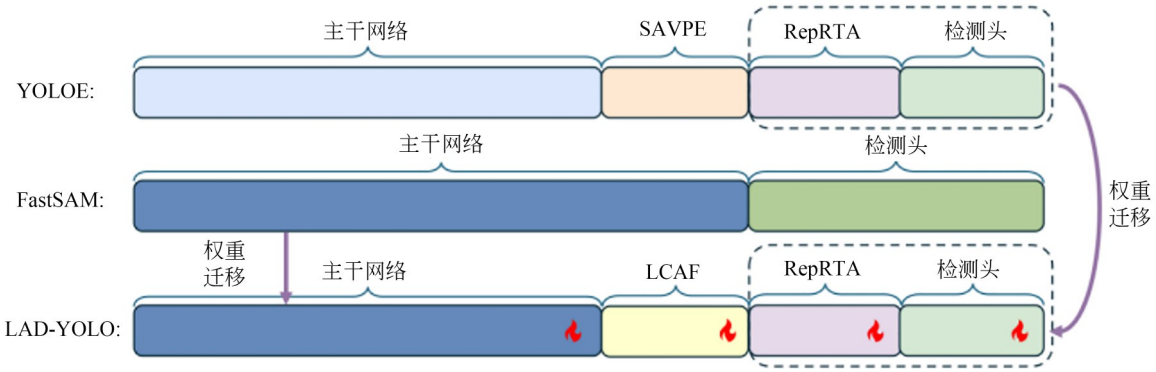


图3　混合预训练权重融合原理

Fig. 3　Fusion principle of hybrid pretrained weights

该融合策略在参数初始化阶段实施。骨干网络参数加载自FastSAM，而头部参数则使用YOLOE的预训练权重进行初始化。对于两种架构中并存但功能各异的组件，则进行选择性替换或合并，以确保兼容性。在微调阶段，骨干网络与头部权重进行联合优化，使网络能够在避免灾难性遗忘的前提下，协调检测与分割知识。由此产生的初始化方式提供了平衡的归纳偏置，通过结合FastSAM的边界精度与YOLOE的定位稳定性，显著提升了模型在检测与分割任务中的收敛速度及性能。

3.3　轻量化通道感知融合

给定中间特征图$F\in\mathrm{R}^{C\times H\times W}$，该模块首先利用通道注意力捕捉“什么”是重要的特征内容，随后通过空间注意力强调包含丰富信息的区域位于“何处”。最后，引入一种可学习的残差融合策略，以自适应地将原始特征与经过注意力增强后的特征进行结合。

轻量化通道感知特征融合模块如图4所示。该模块结合了通道注意力、空间注意力及残差聚合机制，以增强中间特征的表征能力。与传统的串行注意力流程不同，LCAF采用一种可学习的多分支融合策略，将原始特征与提炼后的特征进行整合，从而实现更优的表征学习。

为了建模通道间的相互依赖关系，通过平均池化和最大池化来聚合全局空间上下文信息，这两种操作已被证明能够捕捉互补的统计线索。这两个描述符随后通过一个共享的多层感知机

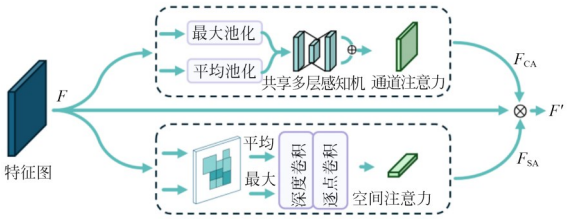


图 4　LCAF 网络架构

Fig. 4　Architecture of LCAF network

(MLP)进行处理,该 MLP 由两层带有 ReLU 激活函数的 1×1 卷积层实现:

$$\begin{cases} F_{\text{avg}}^{c}=\text{AvgPool}(F) \\ F_{\text{max}}^{c}=\text{MaxPool}(F) \\ M_c(F)=\sigma\left(\text{MLP}\left(F_{\text{avg}}^{c}\right)\right)+MLP\left(F_{\text{max}}^{c}\right) \\ F_{\text{ca}}=M_c(F)\cdot F \end{cases}, \tag{1}$$

其中:$\sigma(\cdot)$表示 Sigmoid 函数,MLP 由具有中间通道缩减比的 1×1 卷积构成;AvgPool(•)和 MaxPool(•)分别代表全局平均池化和全局最大池化操作;$M_c(F)$为形状是 $C\times1\times1$ 的通道注意力图。

为了定位特征图中的显著区域,在经过通道注意力增强的特征 F_{ca} 的通道维度上应用平均池化和最大池化操作。由此生成两个空间特征图:

$$\begin{cases} F_{\text{avg}}^{s}=\text{Mean}(F_{\text{ca}},dim=C) \\ F_{\text{max}}^{s}=\text{Max}(F_{\text{ca}},dim=C) \end{cases}, \tag{2}$$

其中:$\text{Mean}(\cdot,dim=C)$ 和 $\text{Max}(\cdot,dim=C)$ 分别表示沿通道维度进行的平均和最大化操作。

池化后的特征图在拼接后输入到一个两阶段卷积模块中,该模块包含一个深度可分离卷积(Depthwise Convolution)和一个逐点卷积(Pointwise Convolution):

$$M_s(F_c)=\sigma(\text{PWConv}_{1\times1}(\text{DWConv}_{k\times k}([F_{\text{avg}}^{s};F_{\text{max}}^{s}])))$$

$$F_{\text{sa}}=M_s(F_{\text{ca}})\cdot F_{\text{ca}}, \tag{3}$$

其中:$[\cdot;\cdot]$表示通道维度的拼接操作,DWConv(•)代表卷积核尺寸为 k 的深度卷积,PWConv(•)表示逐点卷积,$M_s(F)$为空间注意力图,该图与输入特征图 F_{ca} 进行广播相乘。该空间注意力图在保持特征分辨率的同时,选择性地增强了具有丰富空间信息的区域。

不同于仅依赖最终的注意力细化输出,本文提出了一种可学习的残差聚合方案,以整合来自不同阶段的特征。具体而言,原始输入、通道注意力特征及空间注意力特征通过可学习的标量权重进行组合:

$$F_{\text{out}}=\alpha F+\beta F_{\text{ca}}+\gamma F_{\text{sa}}, \tag{4}$$

其中:α,β,γ 为 3 个可学习的标量权重,用于调节各分支对最终输出的贡献度,F_{out} 表示最终增强后的特征图。

该公式允许网络在训练过程中动态调整各分支的贡献,从而实现更优的梯度流动与特征复用。

3.4　逻辑感知多组件蒸馏

为了提升轻量化学生模型在检测与分割任务中的性能,本文提出一种逻辑感知多组件知识蒸馏(Logic-Aware Multi-Component Distillation, LAMCD)策略。不同于单任务或单输出的蒸馏方法,该框架联合蒸馏来自教师模型的多级监督信号,涵盖目标定位、类别分类、置信度估计及掩码生成。这种全面的策略使学生模型不仅能学习教师模型的预测结果,还能领悟其做出该预测的决策逻辑,从而显著提升模型的泛化能力。

在该框架中,学生与教师网络处理同一张输入图像。推理结束后,教师模型输出的边界框、类别得分、目标置信度及分割掩码,将通过一种精心设计的匹配机制用于指导学生网络的训练。针对每一个匹配的目标对象,计算跨不同预测组件的蒸馏损失。

在本文提出的 LAMCD 框架中,学生模型通过多个预测组件接受教师模型的监督,每个组件均配置了专用的损失函数。对于边界框回归,由于边界框坐标属于连续数值回归,在蒸馏训练初期,学生与教师的预测可能存在较大偏差,相较于 L2 损失,L1 损失对异常值具有更强的稳定性,能够提供更稳定的梯度分布,从而实现空间位置的稳健对齐。因此,采用 L1 损失[41]来最小化预测框与目标框坐标之间的距离:

$$L_{L1}^{\text{box}}=\frac{1}{N}\sum_{i=1}^{N}\left\|\hat{b}_i^{(s)}-b_i^{(t)}\right\|_1, \tag{5}$$

其中:$\hat{b}_i^{(s)}$ 和 $b_i^{(t)}$ 分别表示第 i 个目标对象的学生模型预测边界框和对应的教师模型输出。N 表示在当前处理的图像中,学生网络与教师网络之间成功匹配的目标对象的总数量。

对于分类得分、目标置信度及掩码置信度,

均采用均方误差（Mean Square Error，MSE）损失[42]进行蒸馏。MSE损失能够严格地拉近师生网络在Logit层级的连续概率分布距离，有效迫使学生网络继承教师模型的潜在决策逻辑。MSE损失可表示为：

$$L_{\mathrm{MSE}}=\frac{1}{N}\sum_{i=1}^{N}\left\|\hat{b}_i^{(s)}-b_i^{(t)}\right\|_2^2, \tag{6}$$

其中：$l_i^{(s)}$和$l_i^{(t)}$分别代表学生和教师模型的预测值。在本文方法中，$l_i^{(s)}$和$l_i^{(t)}$具体指代分类得分、目标置信度及掩码置信度。

对于分割掩码，在由掩码系数重构的预测掩码与对应的教师掩码之间应用二元交叉熵（BCE）损失[43]，其表达式为：

$$L_{\mathrm{mask}}=-\frac{1}{HW}\sum_{x=1}^{H}\sum_{y=1}^{W}\left[M_{xy}^{(t)}\log\tau\left(M_{xy}^{(s)}\right)+\left(1-M_{xy}^{(s)}\right)\log\left(1-\sigma\left(M_{xy}^{(s)}\right)\right)\right], \tag{7}$$

其中：$M_{xy}^{(t)}$和$M_{xy}^{(s)}$分别为教师和学生模型解码后的掩码预测值，$\tau(\cdot)$为Sigmoid激活函数。HW是这个掩码特征图上的总像素点数量（即$H\times W$）。此处采用BCE损失是因为实例分割本质上是像素级别的独立二分类任务。BCE损失通过独立评估每个像素点的概率散度，能够精准衡量学生掩码与教师掩码在空间拓扑结构上的差异，从而指导学生模型学习更精细的边界轮廓特征。

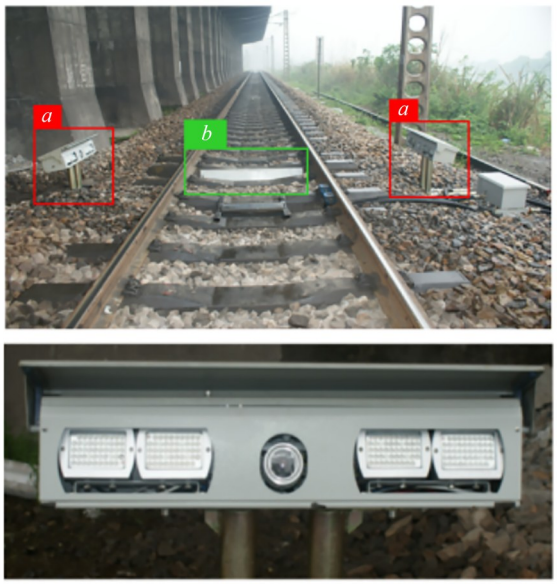
这种多分支监督机制使得学生模型能够同时学习粗粒度的逻辑信息，如目标存在性与类别，以及细粒度的空间信息，如掩码质量。

4　实验与结果分析

4.1　实验设置

4.1.1　数据集

为了验证模型在实例分割应用中的有效性，采用一个综合性的货运列车数据集。该数据集包含4 410张分辨率为700×512的图像，涵盖了6种不同的运行场景和15个类别。数据采集自多个货运列车检测站，包含侧视和底视两个视角，以捕捉多样化的故障模式。成像系统采用工业相机，在自然光或夜间LED辅助照明条件下完成拍摄，图像硬件设备如图5所示。图像中自然包含了粉尘、油污和遮挡等环境干扰因素。例如，闸瓦插销常被闸瓦本体或防脱铁丝环遮挡，导致目标特征不完整。其次，由于采集设备分布于野外轨旁或轨底，图像受自然光和补光灯交替影响，存在多变的环境光照与显著的亮度差异。数据集按照随机分配原则划分为训练集和测试集，其中70%用于模型训练，其余30%用于评估。数据集的详细信息如表1和图6所示。



(a) 轨旁视角
(a) Side view

(b) 轨底视角
(b) Bottom view

图5　野外环境下的图像采集

Fig. 5　Image acquisition in wild environment

数据标注工作在铁路列检员指导下使用LabelMe软件手动完成，生成实例级的多边形掩码，以确保标注的一致性和准确性。对于闸瓦等磨损类部件，磨损严重程度根据磨损限量剩余比例进行定义：轻微磨损（70%～100%）、中度磨损（30%～70%）以及严重磨损（<30%）。对于结构性部件（如转向架止退键、闸瓦插销等），故障被分类为“正常”或“损坏”[44]。

表 1　货运列车故障检测数据集

Tab. 1　Freight train fault detection dataset

部　件	类型	训练集	验证集
转向架挡键	T	725	244
	F	201	101
闸瓦插销	T	442	302
	F	363	53
轴承鞍	T	720	244
	F	201	101
制动拉环	T	443	247
	F	86	98
截断塞门	T	665	286
	F	382	371
夹板螺栓	T	646	224
	F	157	124
集尘器	T	443	247
	F	86	98
刹车闸瓦	T	436	261
	—	—	—

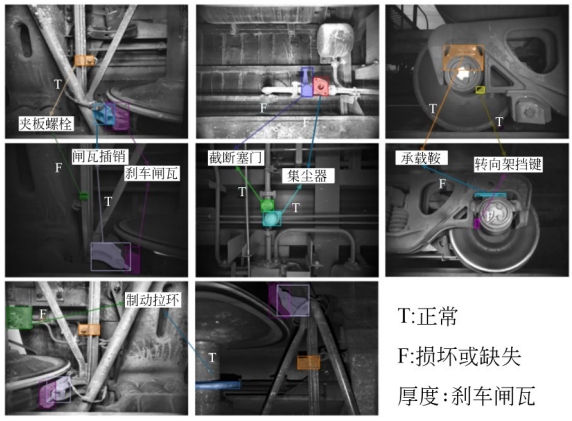


图 6　数据集可视化样本

Fig. 6　Visualization samples of dataset

4.1.2　评价指标

为了全面评估模型性能，构建一套涵盖精度与计算效率的多维评估指标体系。在精度评估方面，选取 AP^{box} 与 AP^{mask} 作为主要指标，分别用于量化目标检测与实例分割任务在 0.50 至 0.95 交并比（IoU）阈值范围内的平均精度（mAP）。此外，引入 AP^{box}_{50} 和 AP^{mask}_{50} 以衡量模型在 0.50 单一 IoU 阈值下的精度表现，从而反映模型在较宽松评估标准下的检测能力。召回率性能则通过 AR^{box}_{50-90}，AR^{mask}_{50-95}，AR^{box}_{50} 及 AR^{mask}_{50} 进行表征，旨在评估模型在不同 IoU 阈值下的检测完整性。为评估实际部署的可行性，本文还考察了计算开销，以分析精度与效率之间的平衡关系。

4.1.3　实施细节

本实验基于前述数据集，遵循标准化的训练流程进行评估。图像预处理阶段将所有输入样本的分辨率统一调整为 1 024×1 024 像素。为增加数据多样性并增强模型的稳定性，引入了常规的数据增强策略，具体包括水平翻转与大尺度抖动技术。

实验硬件环境采用 2 张 NVIDIA GeForce RTX 4090 GPU 配置。模型参数优化采用随机梯度下降（SGD）算法，初始学习率设定为 0.01，批处理大小（Batch Size）设为 10。训练周期设定为 250 个 Epoch，并采用结合线性预热阶段的余弦退火学习率调度策略。算法基于 PyTorch 深度学习框架实现，组件中仅 LCAF 从零开始初始化与训练。测试环境采用单张 NVIDIA GeForce RTX 4090 GPU，批处理大小设定为 10。

4.2　与现有先进方法的比较

为确保对比的公平性，所有基线模型均在相同的货运列车数据集（统一输入分辨率 1 024×1 024 及基础数据增强）下进行评估。各基线模型的超参数（如优化器、批大小和训练轮数）均采用最优配置，以展示其最佳性能界限。本节对比了三类主流框架：传统的 CNN 与 Transformer 两阶段方法、基于 SAM 的方法以及基于 YOLO 的框架。各方法的定性分割效果见图 7，定量精度与模型复杂度详见表 2，模型参数量、精度与推理速度的综合对比见图 8。

实验表明，基于 Transformer 的方法凭借全局注意力机制，在长距离依赖建模与复杂背景下的定位能力优于传统 CNN。同时，得益于大规模预训练，SAM 模型展现出卓越的边界建模与零样本泛化能力。FastSAM-x 的参数量庞大，AP^{mask} 为 76.8 和 AR^{mask}_{50-95} 为 82.7。然而，综合考量精度与轻量化设计，融合了多模态信息的 YOLOE-v8-l 展现出最佳的整体性能，故本文选其作为 LAD-YOLO 的教师模型。

在此基础上，LAD-YOLO 通过跨模块知识迁移、轻量化特征增强及 Logit 级多组件知识蒸馏三大核心策略，实现了显著的性能优化。在计算开销（参数量、GFLOPs 和 FPS）与 YOLOv8-s

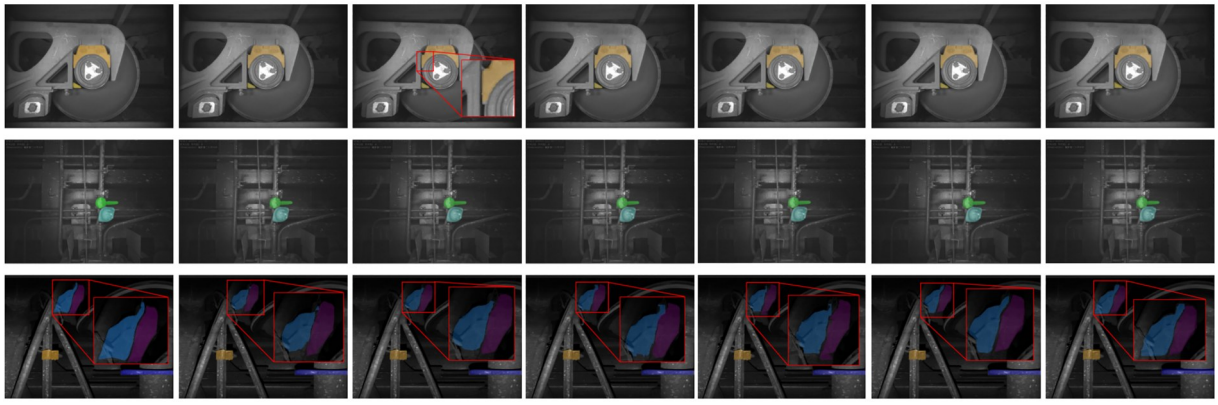


图 7　典型图像样本分割结果的对比可视化（其中，“*”表示采用 Swin-T 架构的模型，放大位置为瑕疵比较）

Fig. 7　Comparative visualization of segmentation results for image samples (The “ * ” denotes the Swin-T architecture, and the magnified regions show the comparison of defects)

表 2　货运列车数据集上与现有主流先进方法的精度对比

Tab. 2　Comparison of accuracy with state-of-the-art methods on freight trains dataset

方　法	主干网络	AP^{box}	AP^{box}_{50}	AR^{box}_{50-95}	AP^{mask}	AP^{mask}_{50}	AR^{mask}_{50-95}	Model Size	GFLOPs	FPS	Params/M
Mask R-CNN[14]	Resnet50	70.1	94.8	78.2	70.7	93.7	77.9	336.4	234	44.6	44.2
PointRend[15]	Resnet50	71.3	94.2	78.4	71.3	94.2	78.4	431.6	186	38.1	56.4
YOLACT[16]	Resnet50	63.6	93	71.2	68.4	93	76	272	337	42.8	31.4
SOLOv2[17]	Resnet50	—	—	—	69.8	93.8	76.8	353.9	283	33.7	46.3
CondInst[18]	Resnet50	69.4	93.2	74.7	67.3	92.8	73.3	272	250	27.1	48.2
SparseInst[20]	Resnet50	—	—	—	66.5	87.9	72.2	399.8	221	76.3	47.9
RTMDet[22]	CSPNeXt[45]	71.3	94.2	61.2	68.8	94	74.8	436	120	23.1	32.5
QueryInst[19]	Resnet50	73	77.4	82.6	66.8	94.5	77.5	1 986.6	263	16.0	62.5
Mask2Former[21]	Resnet50	74.2	92.7	80.1	72.6	93.3	78.9	665.2	245	13.0	46.3
Mask R-CNN[14]	Swin-T[46]	72.6	95.4	78.9	70.5	95.3	76.8	549	241	36.4	47.5
PointRend[15]	Swin-T[46]	69	95.1	76.6	71.8	95	79.1	688.2	198	30.1	65.3
YOLACT[16]	Swin-T[46]	59.9	82.2	67.8	66.6	92.7	74.3	275.4	282	31.6	50.2
SOLOv2[17]	Swin-T[46]	—	—	—	67	94.4	75.8	380.6	312	28.3	49.7
CondInst[18]	Swin-T[46]	65.6	94.5	75.2	72.6	94.9	78.8	293.4	258	30.6	68.3
RTMDet[22]	Swin-T[46]	73.5	94.4	79.3	70.3	94.4	75.3	1 790.6	323	25.6	86.6
Mask2Former[21]	Swin-T[46]	74.3	93	81	73.8	93	79.3	739.5	252	12.8	49.5
FastSAM-x[7]	-	**80.2**	**95.6**	86.1	**76.8**	95.7	**82.7**	141.6	443	9.1	72.0
SAM-seg[34]	SAM-B[4]	72.4	94.6	78.9	71.9	94.7	77.4	450.2	411	8.6	97.2
SAM-det[34]	SAM-B[4]	72	93.2	79	58.9	84.5	67.9	528	233	10.2	106.7
Rsprompter-query[34]	SAM-B[4]	72.7	92.7	79.7	71.9	93.2	77.8	303.3	425	7.1	131.6
YOLOv8-s	—	75.8	**95.6**	82.3	72.4	95.2	78.6	22.9	42.5	**121.6**	11.2
YOLOv8-l	—	78.9	**95.6**	86.8	74.9	95.4	81.7	88.1	220	65.4	45.9
YOLO11-s	—	75	95.4	81.2	72.3	95.5	78.8	39	35.6	120.8	10.6
YOLO12-s	—	74.9	94.7	81.5	71.6	94.6	77.7	35.6	**22.8**	116.5	**9.2**
YOLO26-s		76.7	95.2	84.6	73.1	92.3	77.6	23.5	34.1	132.0	10.3
LAD-YOLO	—	79.7	95.3	**86.7**	76.2	**95.8**	**82.7**	**24**	42.7	118.8	11.7
YOLOE-v8-l(Teacher)	—	**81.8**	**95.9**	**89.9**	**77.5**	**96.1**	**84.4**	88.1	220	49.8	45.9

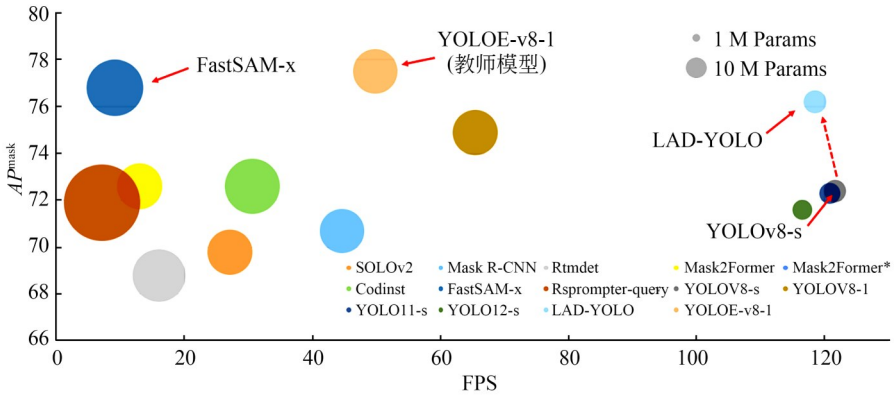


图 8　主流先进(SOTA)方法对比:模型精度与模型参数量的关系(图中横轴表示推理速度(FPS),纵轴表示理论模型精度(AP^{mask}),各模型圆点的大小与其参数量成正比;"*"表示采用 Swin-T 架构的模型)

Fig. 8　Comparison of SOTA methods: illustrating the relationship between model accuracy and model parameters(In the figure, the horizontal axis represents inference speed, the vertical axisrepresents theoretical model accuracy, and the graphical size of each model is proportional to itsnumber of parameters. The " * " denotes the Swin-T architecture)

相当的前提下,LAD-YOLO 的精度超越了大型模型 YOLOv8-l,并逼近教师模型。为排除随机种子干扰,多次重复实验亦验证了该模型稳定且优异的性能表现,具体指标为 $AP^{box}=79.7\pm0.12$,$AP^{mask}=76.2\pm0.15$。

尽管 LAD-YOLO 的综合检测与分割精度大幅提升,但其 AP^{box}_{50} 指标 95.3 较教师模型 YOLOE-v8-l 的 95.9 及基线 YOLOv8-s 的 95.6 略有下降。该现象主要源于轻量化网络在多任务蒸馏中的容量限制与优化权衡:一方面,LAD-YOLO(参数量仅 11.7 M)在联合蒸馏多组件特征时,有限的表征容量会优先向难度更高的掩码预测及严格阈值定位倾斜,导致在宽松阈值下产生微弱的任务竞争损耗;另一方面,逻辑感知蒸馏强化的"软标签"拟合带来了较强的正则化效应,虽大幅提升了模型在复杂场景下的泛化能力,但在极低 IoU 阈值的粗粒度召回上做出了微弱让步。

为全面评估 LAD-YOLO 的工程适用性,对部分典型的失败案例进行了定性分析。研究发现,模型的检测与分割失效主要集中在列车高速运行导致的严重运动模糊,以及不良天气(如阴天、雾霾等)引起的日间低照度场景中。在这些复杂的视觉条件下,自然光照的衰减与大气散射效应使得目标部件与背景的对比度急剧下降,图像的高频纹理细节严重丢失,导致轻量化网络难以提取有效的边缘特征。这使得模型在进行实例分割时,往往无法精准预测出目标(如闸瓦、挡键等)的正确边界,偶有发生掩码溢出或边缘残缺的现象。这些失败案例客观地表明,尽管当前架构已具备良好的全局稳定性,但在应对真实铁路环境中由恶劣天气和动态模糊带来的边界特征衰减问题时,仍具有进一步优化的空间。

综上所述,LAD-YOLO 的核心贡献在于利用"蒸馏+特征增强"的师生架构,有效克服了货运列车场景中背景复杂、遮挡严重等工业视觉难题。该框架不仅加速了模型收敛,更在整体检测性能、分割精细度与边缘端实时推理效率之间达成了卓越的平衡,为实际工业环境的边缘侧部署提供了极具竞争力的轻量化方案。

4.3　消融实验

4.3.1　预训练策略消融

为了验证所提混合预训练权重的有效性,在不同预训练策略下,对 YOLOv8-s 和 LAD-YOLO 两个代表性模型进行了对比实验,结果如表 3 所示。引入预训练权重能够显著提升模型精度,从而突显了预训练在目标检测和实例分割任务中的关键作用。

当使用 SA-1B 作为预训练数据集时,分割性能得到了大幅提升。在 SA-1B 上预训练的 LAD-YOLO 实现了 $AP^{mask}_{50}=96.3$,这表明大规模分割数据使模型能够学习到更具泛化性和判别性的掩码表征。相比之下,在 Objects365 上进行预训练则对检测性能的提升更为显著,LAD-YOLO

表 3　货运列车数据集上预训练模型的对比

Tab. 3　Comparison pretrained models on freight train dataset

方　法	预训练	AP^{box}	AP^{box}_{50}	AR^{box}_{50}	AP^{mask}	AP^{mask}_{50}	AR^{mask}_{50}
yolov8-s	—（无）	75.8	95.6	95.4	72.4	95.2	95.4
	COCO2017	76.5	95.9	95.8	73.2	95.6	95.5
	SA-1B	76.9	96	95.9	73.5	95.7	95.6
	object365	77.1	96.1	96	73.1	94.9	95.3
	混合预训练权重	77.4	96.2	96.1	74	95.3	95.8
LAD-YOLO	—（无）	78.9	95.7	96.2	75.3	95.9	95.7
	COCO2017	79.2	96	**96.4**	75.6	96	95.9
	SA-1B	79.3	95.7	95.6	75.8	**96.3**	**96.1**
	object365	79.5	**96.1**	96.3	75.6	95.7	95.7
	混合预训练权重	**79.7**	95.3	95.5	**76.2**	95.8	95.6

取得了 $AP^{box}_{50}=96.1$。这一观察结果表明，预训练数据集与下游任务之间的对齐程度在很大程度上决定了哪一方面的性能获益最大。

本文提出的混合预训练权重融合了多种预训练源的互补优势，使模型能够在检测和分割任务中获得均衡的增益。在此设置下，LAD-YOLO取得了最佳结果，即 $AP^{box}=79.7$ 和 $AP^{mask}=76.2$。这些发现证实，利用异构预训练数据不仅巩固了不同数据集的优势，更为下游任务提供了更强且更稳定的初始化参数。

4.3.2　LCAF 模块的有效性

为了评估所提轻量化通道感知融合（LCAF）模块的有效性，在 YOLOv8-s 和 LAD-YOLO 上进行了消融实验，并将其与常规注意力模块 CBAM 及具备大感受野性能的分离核注意力模块[47]（Large Separable Kernel Attention，LSKA）进行了对比。如表 4 所示，当集成到 LAD-YOLO 中时，LCAF 将检测精度提升至 $AP^{box}=79.7$，分割精度提升至 $AP^{mask}=76.2$，均优于基线模型及 CBAM，LSKA 增强变体。

表 4　不同特征增强方法的影响

Tab. 4　Impact of different feature enhancement methods

方　法	特征增强	AP^{box}	AP^{box}_{50}	AR^{box}_{50}	AP^{mask}	AP^{mask}_{50}	AR^{mask}_{50}
yolov8-s	—（无）	75.8	95.6	95.4	72.4	95.2	95.4
	CBAM	76.0	95.5	95.3	72.6	95.3	95.2
	LSKA	76.3	95.4	95.4	72.7	95.3	95.4
	LCAF	76.5	95.4	95.5	73.0	95.6	95.5
LAD-YOLO	—（无）	79.3	95.5	95.8	75.8	95.7	95.3
	CBAM	79.5	95.5	95.6	76.0	95.7	95.4
	LSKA	79.5	95.4	95.4	76.0	95.8	95.4
	LCAF	79.7	95.3	95.5	76.2	95.8	95.6

这种提升不仅归功于注意力的引入，更在于LCAF的可学习多分支通道融合机制。与应用固定通道和空间注意力的CBAM不同，LCAF通过可学习的标量权重 α,β,γ 动态地调节原始特征、通道增强特征与空间增强特征的融合比例。为了直观展示各分支的贡献，提取了训练收敛后的权重参数，代表性取值如下：$\alpha\approx0.507,\beta\approx0.531,\gamma\approx0.789$。其中，$\gamma$ 显著高于其他分支，表明在背景复杂且故障目标微小的铁路列检场景中，空间维度的特征增强对精准定位故障区域起到了关键的驱动作用。这种设计减轻了由单一注意力路径导致的过度抑制或冗余风险。

此外，LCAF 在 YOLOv8-s 和 LAD-YOLO

上均稳定提升。这种设计更好地契合了工业场景下细粒度特征提取与多尺度上下文建模的需求。综上所述,LCAF模块在有效抑制无关背景噪声的同时,增强了任务相关的判别性表征,提升了检测和分割性能。

4.3.3　逻辑感知多组件蒸馏消融

为了验证所提逻辑感知多组件蒸馏(LAMCD)的有效性,在LAD-YOLO框架上进行了对比实验,结果如表5所示。配备LAMCD的LAD-YOLO实现了$AP^{box}=79.7$和$AP^{mask}=76.2$,表现出优于其他蒸馏方法的显著提升。相比之下,SO-DETR通过目标级响应引导增强了定位一致性,但未能充分利用跨多任务的互补信息;而MGD虽然在特征表征方面表现出优势,但其监督主要局限于中间特征,限制了其捕捉检测与分割之间逻辑关联的能力。

LAMCD的优越性在于其逻辑感知机制,该机制使学生模型能够学习框、掩码及分类输出之间的内在依赖关系,而多组件融合则为更精确的定位与分割提供了更丰富的监督信号。由此表明,LAMCD能够充分发挥LAD-YOLO的结构优势,在保持模型紧凑性的同时,有效地提升小目标检测与实例分割的稳定性和准确性。

表5　不同蒸馏方法对模型的影响

Tab. 5　Impact of distillation methods on different models

模　型	蒸馏方法	AP^{box}	AP^{box}_{50}	AR^{box}_{50}	AP^{mask}	AP^{mask}_{50}	AP^{mask}_{50}
yolov8-s	—(无)	75.8	**95.6**	95.4	72.4	95.2	95.4
yolov8-s	SO-DETR	76.8	95.5	95.2	72.8	95.7	95.3
yolov8-s	MGD	76.5	95.5	95.3	72.7	95.5	95.4
LAD-YOLO	SO-DETR	77	95.5	95.4	73.5	95.8	**95.8**
LAD-YOLO	LAMCD	**79.7**	95.3	**95.5**	**76.2**	**95.8**	95.6
yoloe-v8-l	教师模型	81.8	95.9	95.6	77.5	96.1	95.5

4.3.4　组件集成分析

本节进行了详细的消融研究,以评估所提组件的有效性。结果汇总于表6。在未引入任何额外模块的情况下,基线模型表现出相对有限的性能($AP^{box}=75.8$,$AP^{mask}=72.4$)。

单独引入LCAF模块带来了检测与分割性能的一致提升,证明了其增强特征表征并有益于下游任务的能力。当进一步集成混合权重(Hybrid Weight)机制时,性能显著增加,AP^{box}升至77.8,AP^{mask}升至74.3,表明该机制有效地促进了知识迁移与优化。

最后,结合完整的LAMCD框架取得了最佳结果,将AP^{box}提升至79.7,AP^{mask}提升至76.2,分别较基线提升了3.9%和3.8%。同时,召回率指标也保持在较高水平。这些发现证实了所提多组件蒸馏策略的高度有效性,各模块之间相辅相成,共同推动了性能的大幅提升。

表6　组件集成消融实验

Tab. 6　Ablation study on component integration

LCAF	混合权重	LAMCD	AP^{box}	AP^{box}_{50}	AR^{box}_{50}	AP^{mask}	AP^{mask}_{50}	AR^{mask}_{50}
			75.8	95.6	95.4	72.4	95.2	95.4
√			76.4	95.2	94.7	72.9	95.8	95.4
√	√		77.8	95.4	95.3	74.3	95.8	95.5
√	√	√	79.7	95.3	95.5	76.2	95.8	95.6

4.3.5　闸瓦磨损测量

应用LAD-YOLO方法评估闸瓦磨损状况,并将其与两种传统方法进行了比较:一种是基于目标检测获取边界框后的投影计算方法,另

一种是基于边缘检测后的几何分析方法。在评价体系方面，本文采用分类准确率（Accuracy）衡量磨损状态判别的可靠性。其中，磨损程度的分类通过计算预测掩码的像素数量，或提取目标检测框的几何特征，与其标准完整闸瓦状态下相应值的比值，并按4.1.1中磨损程度的定义判别。

对比结果如表7所示。实验结果表明，在磨损检测任务中，基于实例分割的方法显著优于传统的投影与几何分析方法[48]，高精度的分割模型带来了更优越的检测性能，避免了传统机器视觉对环境变化敏感的局限。这种改进主要归因于集成在LAD-YOLO框架内的SAM模块所提供的高质量分割结果。

表7　方法磨损程度检测精度对比

Tab. 7　Comparison of detection accuracy across different wear levels using various methods　（%）

方　法	轻微磨损	中度磨损	严重磨损
LAD-YOLO	96.7	94.8	97.5
Mask R-CNN	91.6	91.2	93.9
Bounding Box	82.3	84.1	88.7
Edge detection[48]	75.6	78.4	81.2

5　讨　论

LAD-YOLO在货运列车故障检测任务中实现了精度与效率的有效平衡。针对实际轨旁列检系统至少30 FPS的实时性要求，LAD-YOLO（参数量为11.7 M，计算量为42.7 GFLOPs）实现了118.8 FPS的推理速度。相比于教师模型YOLOE-v8-l（参数量为45.9 M，计算量220为GFLOPs），本方法在维持高精度的同时，推理速度相较与教师模型提升了近2.4倍，计算开销显著降低。这表明该模型完全具备下沉部署至算力与功耗受限的边缘侧平台（如NVIDIA Jetson系列）的可行性，能够稳定满足工业现场的实时处理要求。

这种优势源于跨模型预训练权重迁移、LCAF特征增强及LAMCD蒸馏策略的协同作用，使轻量化模型有效继承了基座模型的语义与定位先验。此外，在闸瓦磨损定量分析中，基于实例分割的LAD-YOLO优于传统投影与几何分析方法，验证了其在不同磨损等级评估中的稳定性。尽管如此，当前框架仍存在局限：一是依赖固定教师模型和纯监督蒸馏，未来可探索半监督蒸馏以适应标注稀缺场景；二是针对极端天气或运动模糊等复杂铁路环境，需进一步开展域自适应研究以提升模型泛化能力。

6　结　论

本文提出了一种面向货运列车列检的轻量化故障检测框架LAD-YOLO，该框架融合了基座模型分割先验与多组件知识蒸馏策略。通过结合FastSAM的语义建模能力与YOLOE的检测稳定性，引入LCAF模块进行特征增强，并设计逻辑感知多组件蒸馏机制，LAD-YOLO在保持低计算开销的同时实现了高精度的检测与分割。在自建货运列车数据集上的广泛实验验证表明，LAD-YOLO在精度和稳定性方面均优于现有的轻量化检测器，为实时部署提供了可靠支持。在闸瓦磨损测量中应用进一步表明，该框架具备高精度的定量故障分析能力，对铁路安全监测与预测性维护具有重要意义。

未来的工作将致力于增强模型的自适应性与可扩展性，引入半监督学习以降低标注成本、融合硬件感知优化以支持在边缘设备上的高效部署，以及扩展至更复杂的铁路场景。综上所述，本研究为智能铁路列检提供了一种实用且高效的解决方案，并为将基座模型知识迁移至轻量化可部署系统提供了新的思路。

作者贡献声明：

孙国栋：方法提出，论文审核与编辑写作；
梁启航：实验设计，数据分析与论文撰写；
邹鹏坤：数据整理与分析；
李郭钰、潘星宇：实验设计与样本采集。

参考文献：

[1] WEI X K, JIANG S Y, LI Y, *et al*. Defect detection of pantograph slide based on deep learning and image processing technology[J]. *IEEE Transactions on Intelligent Transportation Systems*, 2020, 21(3): 947-958.

[2] ZHANG Y, LIU M Y, YANG Y, *et al*. A unified light framework for real-time fault detection of freight train images[J]. *IEEE Transactions on Industrial Informatics*, 2021, 17(11): 7423-7432.

[3] PAN X Y, LIANG Q H, SUN G D, *et al*. Feature interconnection for fault detection with inaccurate bounding boxes[J]. *Cluster Computing*, 2025, 28(12): 804.

[4] KIRILLOV A, MINTUN E, RAVI N, *et al*. Segment anything[C]. 2023 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). *October* 1-6, 2023, *Paris*, *France*. IEEE, 2024: 3992-4003.

[5] WANG A, LIU L H, CHEN H, *et al*. YOLOE: real-time seeing anything[EB/OL]. 2025: *arXiv*: 2503.07465. https://arxiv.org/abs/2503.07465.

[6] 朱广, 顾晨, 徐立云, 等. 改进YOLOv8的风机叶片多尺度缺陷检测[J]. 光学 精密工程, 2025, 33(9): 1496-1514.
ZHU G, GU CH, XU L Y, *et al*. Improvement of YOLOv8 for multi-scale defect detection in wind turbine blades[J]. *Optics and Precision Engineering*, 2025, 33(9): 1496-1514. (in Chinese)

[7] ZHAO X, DING W C, AN Y Q, *et al*. Fast segment anything [EB/OL]. 2023: *arXiv*: 2306.12156. https://arxiv.org/abs/2306.12156.

[8] YASEEN M. What is YOLOv8: an in-depth exploration of the internal features of the next-generation object detector [EB/OL]. 2024: *arXiv*: 2408.15857. https://arxiv.org/abs/2408.15857.

[9] LIU L, ZHOU F Q, HE Y Z. Automated visual inspection system for bogie block key under complex freight train environment[J]. *IEEE Transactions on Instrumentation and Measurement*, 2016, 65(1): 2-14.

[10] 赵进, 郭寅, 尹仕斌, 等. 强环境噪声下的双目视觉受电弓轨旁异常检测[J]. 光学 精密工程, 2025, 33(3): 438-451.
ZHAO J, GUO Y, YIN SH B, *et al*. Binocular vision-based trackside pantograph anomaly detection under strong environmental noise[J]. *Optics and Precision Engineering*, 2025, 33(3): 438-451. (in Chinese)

[11] 刘彦磊, 李孟喆, 王宣宣. 轻量型YOLOv5s车载红外图像目标检测[J]. 中国光学(中英文), 2023, 16(5): 1045-1055.
LIU Y L, LI M ZH, WANG X X. Lightweight YOLOv5s vehicle infrared image target detection [J]. *Chinese Optics*, 2023, 16(5): 1045-1055. (in Chinese)

[12] LI C Z, YANG K H, TANG H C, *et al*. Fault diagnosis for rolling bearings of a freight train under limited fault data: few-shot learning method[J]. *Journal of Transportation Engineering*, Part A: Systems, 2021, 147(8): 04021041.

[13] LIU Y, SHI H M, QIU J, *et al*. Tiny network for faults recognition in freight cars[J]. *Engineering Applications of Artificial Intelligence*, 2025, 144: 110018.

[14] HE K M, GKIOXARI G, DOLLAR P, *et al*. Mask R-CNN[C]. 2017 *IEEE International Conference on Computer Vision* (*ICCV*). *October* 22-29, 2017. *Venice*. IEEE, 2017: 2980-2988.

[15] KIRILLOV A, WU Y X, HE K M, *et al*. PointRend: image segmentation as rendering[C]. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 13-19, 2020. *Seattle*, *WA*, *USA*. IEEE, 2020: 9796-9805.

[16] BOLYA D, ZHOU C, XIAO F Y, *et al*. YOLACT: real-time instance segmentation[C]. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). *October* 27-*November* 2, 2019. *Seoul*, *Korea*. IEEE, 2019: 9156-9165.

[17] WANG X L, ZHANG R F, KONG T, *et al*. SOLOv2: dynamic and fast instance segmentation [EB/OL]. 2020: *arXiv*: 2003.10152. https://arxiv.org/abs/2003.10152.

[18] TIAN Z, SHEN C H, CHEN H. Conditional convolutions for instance segmentation[C]. *Computer Vision-ECCV* 2020. Cham: Springer, 2020: 282-298.

[19] FANG Y X, YANG S S, WANG X G, *et al*. Instances as queries[C]. 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). *October* 10-17, 2021. *Montreal*, *QC*, *Canada*. IEEE, 2021: 6890-6899.

[20] CHENG T H, WANG X G, CHEN S Y, *et al*. Sparse instance activation for real-time instance seg-

mentation[C]. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June* 18-24, 2022, *New Orleans, LA, USA*. IEEE, 2022: 4423-4432.

[21] CHENG B W, MISRA I, SCHWING A G, *et al*. Masked-attention mask transformer for universal image segmentation[C]. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June* 18-24, 2022, *New Orleans, LA, USA*. IEEE, 2022: 1280-1289.

[22] LYU C Q, ZHANG W W, HUANG H A, *et al*. RTMDet: an empirical study of designing real-time object detectors[EB/OL]. 2022: *arXiv*: 2212.07784. https://arxiv.org/abs/2212.07784.

[23] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition. June* 18-23, 2018, *Salt Lake City, UT, USA*. IEEE, 2018: 7132-7141.

[24] WOO S, PARK J, LEE J Y, *et al*. *CBAM: Convolutional Block Attention Module*[M]. Computer Vision-ECCV 2018. Cham: Springer International Publishing, 2018: 3-19.

[25] WANG Q L, WU B G, ZHU P F, *et al*. ECA-net: efficient channel attention for deep convolutional neural networks[C]. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, *Seattle, WA, USA*. IEEE, 2020: 11531-11539.

[26] YANG L X, ZHANG R Y, LI L D, *et al*. SimAM: a simple, parameter-free attention module for convolutional neural networks[C]. *International Conference on Machine Learning*, 2021.

[27] MEHTA S, RASTEGARI M. MobileViT: light-weight, general-purpose, and mobile-friendly vision transformer [EB/OL]. 2021: *arXiv*: 2110.02178. https://arxiv.org/abs/2110.02178.

[28] WU K, ZHANG J N, PENG H W, *et al*. TinyViT: fast pretraining distillation forSmall vision transformers[C]. *Computer Vision-ECCV* 2022. Cham: Springer, 2022: 68-85.

[29] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, *et al*. An image is worth 16x16 words: transformers for image recognition at scale[EB/OL]. 2020: *arXiv*: 2010.11929. https://arxiv.org/abs/2010.11929.

[30] RADFORD A, KIM J W, HALLACY C, *et al*. Learning transferable visual models from natural language supervision[EB/OL]. 2021: *arXiv*: 2103.00020. https://arxiv.org/abs/2103.00020.

[31] BAI J Z, BAI S, CHU Y F, *et al*. Qwen technical report[EB/OL]. 2023: *arXiv*: 2309.16609. https://arxiv.org/abs/2309.16609.

[32] BROWN T B, MANN B, RYDER N, *et al*. Language models are few-shot learners[EB/OL]. 2020: *arXiv*: 2005.14165. https://arxiv.org/abs/2005.14165.

[33] LIU X X, ZHAO Y, WANG S G, *et al*. G-SAM: GMM-based segment anything model for medical image classification and segmentation[J]. *Cluster Computing*, 2024, 27(10): 14231-14245.

[34] 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: 4701117.

[35] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[EB/OL]. 2015: *arXiv*: 1503.02531. https://arxiv.org/abs/1503.02531.

[36] ZHENG Z H, YE R G, WANG P, *et al*. Localization distillation for dense object detection[C]. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June* 18-24, 2022, *New Orleans, LA, USA*. IEEE, 2022: 9397-9406.

[37] ZHAO B R, CUI Q, SONG R J, *et al*. Decoupled knowledge distillation[C]. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June* 18-24, 2022, *New Orleans, LA, USA*. IEEE, 2022: 11943-11952.

[38] ZHANG S X, LIU H P, HOPCROFT J E, *et al*. Class-aware information for logit-based knowledge distillation[EB/OL]. 2022: *arXiv*: 2211.14773. https://arxiv.org/abs/2211.14773.

[39] YANG Z D, LI Z, SHAO M Q, *et al*. Masked generative distillation[C]. *Computer Vision-ECCV* 2022. Cham: Springer, 2022: 53-69.

[40] ZHANG H X, ZHANG H, MEI A R, *et al*. SO-DETR: leveraging dual-domain features and knowledge distillation for small object detection [EB/OL]. 2025: *arXiv*: 2504.11470. https://arxiv.org/abs/2504.11470.

[41] XUE Y, XU T, ZHANG H, *et al*. SegAN: ad-

versarial network with multi-scale L1 loss for medical image segmentation [J]. *Neuroinformatics*, 2018, 16(3): 383-392.

[42] LECUN Y, BOTTOU L, BENGIO Y, *et al*. Gradient-based learning applied to document recognition [J]. *Proceedings of the IEEE*, 1998, 86(11): 2278-2324.

[43] SU J, LIU Z, ZHANG J, *et al*. DV-Net: Accurate liver vessel segmentation *via* dense connection model with D-BCE loss function [J]. *Knowledge-Based Systems*, 2021, 232: 107471.

[44] 中国国家铁路集团有限公司. 铁路货车运用维修规程: TG/CL 113-2018 [S]. 北京: 中国铁道出版社, 2018.
China State Railway Group Co., Ltd. Railway freight car operation regulations and mainterance: TG/CL 113-2018 [S]. Beijing: China Railway Publishing House, 2018. (in Chinese)

[45] CHEN X Q, YANG C Z, MO J, *et al*. CSPNeXt: a new efficient token hybrid backbone [J]. *Engineering Applications of Artificial Intelligence*, 2024, 132: 107886.

[46] LIU Z, LIN Y T, CAO Y, *et al*. Swin transformer: hierarchical vision transformer using shifted windows[C]. 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). *October* 10-17, 2021, *Montreal*, *QC*, *Canada*. IEEE, 2022: 9992-10002.

[47] LAU K W, PO L M, REHMAN Y A U. Large separable kernel attention: rethinking the large kernel attention design in CNN [J]. *Expert Systems with Applications*, 2024, 236: 121352.

[48] KIM H, KIM W Y. Automated inspection system for rolling stock brake shoes [J]. *IEEE Transactions on Instrumentation and Measurement*, 2011, 60(8): 2835-2847.

作者简介：

孙国栋(1981—),男,湖北天门人,博士,教授,2002年、2008年于华中科技大学分别获得学士和博士学位,主要从事机器视觉和机器学习等方面的研究。E-mail: sgdeagle@163.com