文章编号:1671-4598(2026)01-0033-09 DOI:10.16526/j.cnki.11-4762/tp.2026.01.005 中图分类号:TP391.4 文献标识码:A

基于全局语义与局部特征融合的铁路异物侵限检测

袁花明[1]，薛云龙[1]，许　剑[1]，虞浩凡[2]
(1. 陕西靖神铁路有限责任公司，陕西 榆林　719054；
2. 西安交通大学 信息与通信工程学院，西安　710049)

摘要：针对铁路异物侵限检测中传统方法泛化能力差以及基于深度学习的检测模型存在漏检率和误检率较高的问题，提出了一种全局语义与局部特征融合的铁路异物检测方法；通过解耦处理不同光照条件（白天/夜晚）与摄像头模态（可见光/红外）下的检测任务，结合 YOLOv7 检测模型与 BLIP 多模态大模型的语义理解能力，构建了双阈值动态判定策略；采用 YOLOv8 分割模型精准提取铁轨区域以减少背景干扰；训练适用于不同模态和光照条件的 YOLOv7 检测模型，并引入低光增强与噪声抑制技术优化夜间检测性能；利用 BLIP 模型对图像进行语义分析，根据其输出动态调整 YOLOv7 的检测阈值以平衡漏检率与误检率；经实验测试，在自建铁路异物检测数据集上该方法的 *mAP* 达到 88.9%，相比基线模型提升 0.5%，在真实场景的测试集上误检率和漏检率分别低至 1.09%和 0.22%；该方法具备良好的实时性与鲁棒性，满足复杂环境下的工程应用需求。

关键词：铁轨异物检测；通专结合；YOLOv7；BLIP 模型；双阈值策略

Railway Foreign Object Detection Based on the Fusion of Global Semantics and Local Features

YUAN Huaming[1], XUE Yunlong[1], XU Jian[1], YU Haofan[2]
(1. Shaanxi Jingshen Railway Co., Ltd., Yulin　719054, China;
2. School of Information and Communication Engineering, Xi'an Jiaotong University, Xi'an　710049, China)

Abstract: To address the issues of poor generalization ability of traditional methods and high false detection and missed detection rates of deep learning-based detection models in railway foreign object intrusion detection, a railway foreign object detection method based on the the Fusion of Global Semantics and Local Features is proposed. By decoupling the detection tasks under different lighting conditions (day/night) and camera modalities (visible light/infrared), and combining the YOLOv7 detection model with the semantic understanding ability of the BLIP multimodal large model, a dual-threshold dynamic determination strategy is constructed. The YOLOv8 segmentation model is used to precisely extract the railway track area to reduce background interference. YOLOv7 detection models suitable for different modalities and lighting conditions are trained, and low-light enhancement and noise suppression techniques are introduced to optimize the detection performance at night. The BLIP model is used to perform semantic analysis on the images, and the detection threshold of YOLOv7 is dynamically adjusted based on its output to balance the missed detection rate and false detection rate. Experimental tests show that the *mAP* of this method on the self-built railway foreign object detection dataset reaches 88.9%, an improvement of 0.5% compared to the baseline model. On the real scene test set, the false detection rate and missed detection rate are as low as 1.09% and 0.22%, respectively. This method has good real-time performance and robustness, meeting the engineering application requirements in complex environments.

Keywords: railway foreign object detection; general-specific combination; YOLOv7; BLIP model; dual-threshold strategy

收稿日期:2025-07-17； **修回日期**:2025-09-03。

作者简介:袁花明(1977-),男,大学本科,高级工程师。

引用格式:袁花明,薛云龙,许　剑,等.基于全局语义与局部特征融合的铁路异物侵限检测[J].计算机测量与控制,2026,34(1):33-41.

0 引言

一般情况下，异物侵限检测可以分为接触式和非接触式两种[1]，由于接触式检测高度依赖系统布置，存在维护成本高、灵活性差，检测性能较低等问题。基于这些问题，近年来众多学者针对非接触式异物侵限检测进行了研究。早期研究人员多采用传统机器学习和图像处理方法实现异物检测。文献［2］采用背景差分和支持向量机（SVM，support vector machine）检测并分类异物，并采用卡尔曼滤波器（Kalman Filter）滤波器进行目标跟踪，滤除非侵限干扰。文献［3］在 SVM 的基础上引入词频－逆文档频率（TF-IDF，term frequency-inverse document frequency）变换，通过利用超像素内的特征提升分类性能。而文献［4］针对背景差分中抗干扰能力差的问题，提出引入混合高斯分布和低秩矩阵分解进行更细致的背景建模。然而，由于传统方法主要依赖浅层特征，仍然无法完全避免泛化能力差的问题。

近年来，随着深度学习的发展，深度卷积神经网络（CNN，convolutional neural network）已经在各个领域中取得了成功的应用，其中也包括铁路的异物侵限检测。相比传统图像处理方法，深度学习有检测精度高、鲁棒性强等优势，能在不同光线条件、不同天气、不同背景环境中都实现较好的检测性能，引起了研究人员的关注。首先，部分工作采用了双阶段目标检测算法，如文献［5］采用 Faster R-CNN（Faster Region-based Convolutional Neural Network）模型进行铁轨异物检测，并使用全局平均池化层代替全连接层，针对性优化模型。而单阶段目标检测相对双阶段有速度优势，文献［6］在 SSD（Single Shot MultiBox Detector）模型的基础上引入了特征金字塔，使模型在铁路小目标异物检测上取得了更好的效果；文献［7］同样针对小目标，采用空洞卷积替换 YOLOv3（You Only Look Once）中的部分卷积层，以提升检测性能；而文献［8］则是在 YOLOv8 模型中引入了 CBAM（Convolutional Block Attention Module）注意力机制，增强了检测模型的特征提取能力。此外，文献［21］研制了基于 YOLOv8 与 Faster R-CNN 的高速铁路线路环境异物入侵视频检测系统，结合双模态补偿技术，在高速运行场景下实现了高精度的动态视频检测。另文献［22］则针对铁路作业人员的安全防护，提出了基于 YOLOv5 改进算法的安全帽与反光背心检测方法，在提升检测准确率的同时有效降低了模型复杂度。这些研究为铁路场景下深度学习目标检测方法的工程化应用提供了有力支撑，同时也表明异物检测技术已经从依赖手工特征的传统方法逐步发展到以深度学习为核心的方法。前者为早期研究提供了基础框架与思路启发，但在面对光照变化和复杂背景时存在泛化与鲁棒性不足的瓶颈；后者虽然凭借强大的数据驱动与自动特征学习能力，在检测精度和稳定性方面取得了显著提升，一定程度上突破了传统方法的局限，但其感知和推理仍主要局限于像素级的视觉特征，缺乏对场景全局语义和复杂逻辑关系的深层理解，这导致它们在面对非典型、弱特征或需要背景知识判断的异常情况时，鲁棒性不足，仍存在漏检和误检的风险。

目前基于卷积神经网络的检测模型都规模较小，且依赖于 COCO、VOC 等常见检测数据集进行预训练，数据较少且标注相对单一，不能充分利用图像中的深层语义信息，漏检率和误检率仍有优化空间。从最初的主要应用于自然语言处理的大语言模型（如 GPT[9-11] 系列）发展到多模态大模型（如 CLIP[12]、DALL-E[13] 等），多模态大模型的优越性能和强大的泛化性在多个领域（如自动驾驶、医疗影像分析、内容生成等）得到了广泛的应用和验证。多模态大模型可以应用与多种视觉与语言交互任务，包括视觉问答（VQA，visual question answering）、图文生成、图文检索等。其中，BLIP 模型（Bootstrapping Language-Image Pre-training[14]）是文献［14］提出的一个用于视觉问答和其他多模态任务的大模型。BLIP 通过结合图像和文本数据进行预训练，旨在提升模型对视觉和语言的理解能力。而在铁轨异物侵限检测与识别任务中，通过微调 BLIP 大模型并向其提问的方式，可以提取图像更深层的语义信息，从而提升告警的准确率，降低异常事件的漏检和误检。

本文提出了一种基于全局语义与局部特征融合的铁路异物检测算法，在图像通过智能侵限区域划分之后，利用 YOLOv7 专用检测模型进行检测框精确位置的预测，同时充分利用通用预训练大模型中的深层语义信息，提升检测系统对不同天气、不同光线条件、不同背景环境的鲁棒性，通过模型串联的方式挖掘低分检测框，从而降低异物漏检率和误检率，更全面的保障铁轨运行的安全。

1 YOLOv7 检测算法与 BLIP 多模态模型概述

近年来，深度学习在铁路异物检测领域取得了显著进展。传统机器学习方法，如背景差分和支持向量机，在特定场景下表现良好，但由于浅层特征的限制，难以应对复杂环境。深度卷积神经网络的引入显著提升了检测性能，其中双阶段检测算法（如 Faster R-CNN[15]）在精度上占优，而单阶段检测算法（如 YOLO[16] 和 SSD[17]）则以实时性著称。YOLOv7[18] 通过增强 ELAN 结构和引入路径聚合网络（PAN，path aggregation network），在保证高精度的同时兼顾速度，已被广泛应用

于铁路监控场景。此外，多模态模型 BLIP 的出现进一步扩展了检测方法的语义处理能力，通过结合视觉和语言特征，可有效识别复杂环境中的异常。结合这些先进技术，本文提出了一种基于全局语义与局部特征融合的铁路异物检测方法，以应对复杂场景下的异物侵限和遗留问题。

1.1　YOLOv7 检测算法

YOLOv7 是 YOLO 系列的版本之一，该算法专注于实时对象检测，具有较高的精度和速度。YOLOv7 在 YOLOv4[19] 和 YOLOv5 的基础上进行了多项改进，包括更高效的模型架构、优化的训练策略和更好的推理性能。

YOLO 系列算法将目标检测任务定义为一个回归问题，通过 CNN 网络，直接在输出层回归目标的位置和类别信息。YOLOv7 引入了多种模型架构，模型的深度与宽度影响其精度与效率。模型越深、越宽，其识别精度越高，但相应的计算复杂度也会增加。YOLOv7 网络由主干网络（Backbone）、网络的颈部（Neck）以及检测头（Head）三部分组成，如图 1 所示。主干网络用于图像的特征提取，颈部负责对主干网络提取的特征进行高效融合，以结合高层语义信息和底层空间信息，检测头则基于融合后的特征图进行最终的预测，输出目标的类别和位置。在数据上，YOLOv7 引入了 Mosaic 数据增强技术，帮助模型学习更多的图像位置和像素变化，从而有效提升了预测的准确性和模型性能。而主干骨架部分则采用了增强的 ELAN 结构，与传统的逐层堆叠不同，增强的 ELAN 在设计中在原有层次化连接的基础上引入更多梯度流分支，使特征在不同深度的路径间能够充分传递与融合。在保持整体网络轻量化的同时，这一改进有效提升了梯度信息的流动性与表达能力，避免了因单一路径过深而导致的梯度消失或训练不稳定问题。具体而言设输入特征为 f_{in}，首先，输入特征 f_{in} 会被送入两个并行的卷积层，以变换通道并产生两个初始特征分支 f_0 和 f_1。f_0 分支通常经过较少处理，作为直接传递信息的“短路”分支，保留了最原始的梯度流。而 f_1 分支则会进入一个由 k 个计算单元组成的扩展块。在 E-ELAN 中，这些计算单元采用了分组卷积 $G_i(\cdot)$，以极小的计算成本增加特征的多样性。其内部特征传递过程可以看作是一系列的特征变换与累积：

$$f_i(n+1) = G_n(f_n)$$

最终，模块的输出 f_{out} 是将所有层级的特征图进行拼接的结果，该操作聚合了不同感受野和计算深度的信息：

$$f_{out} = \mathrm{Concat}(f_0, f_1, f_2, \cdots, f_{k+1})$$

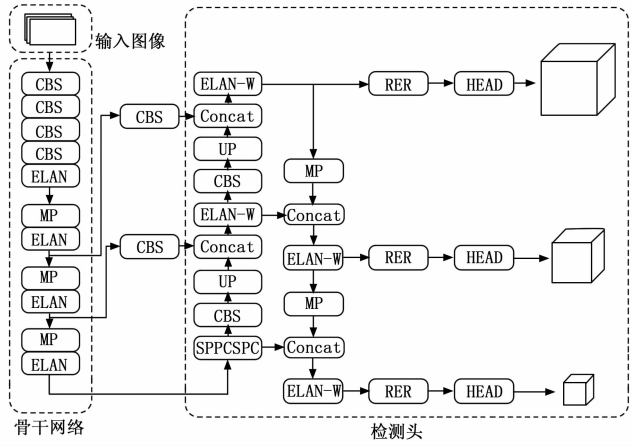


图 1　YOLOv7 模型结构

颈部部分通过引入 PAN 结构，其结构如图 2 所示，其构建了一个高效的双向特征融合网络。它在传统 FPN 自顶向下传递强语义信息的基础上，创新性地增加了一条自底向上传递强定位信息的路径。通过上采样与下采样的交替进行，高层语义特征与底层空间细节得以在所有层级的特征图上充分流动与融合。这种设计确保了最终送入检测头的特征金字塔同时富含精确的定位线索和丰富的语义信息，从而极大地增强了模型在多尺度目标检测任务上的性能。

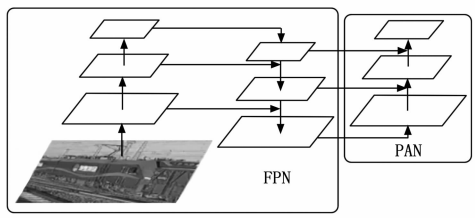


图 2　FPN＋PAN 结构图

头部部分采用解耦的检测头，分类和回归的任务分别通过不同的卷积模块进行处理。分类分支预测目标类别，另一个分支负责边界框的位置以及交并比（IoU，intersection over union）的预测。YOLOv7 作为专用检测模型有很多优点，首先它保持了 YOLO 系列的高速度，在检测大规模图像或实时视频流时非常适合。这对于铁路异物检测，特别是在高速铁路场景下，需要快速反应和高效处理的任务来说是一个重大优势。同时，相较于前代版本，在精度上有显著提升，能够更好地检测小物体和复杂背景中的目标，这对于铁路异物检测中一些细微的标志、信号等关键目标尤为重要。除此以外，由于改进的训练策略和数据增强，YOLOv7 在不同环境下（如白天和夜晚，晴天和雨天）的泛化能力相对于此前的 YOLO 系列有所提升，更适合处理天气和光线条件多变的铁轨异物检测任务。

1.2 BLIP 多模态大模型

BLIP 是一种用于视觉语言理解的多模态模型，其结构设计旨在解决图像和文本的联合建模任务。BLIP 的主要架构包括三个关键部分：图像编码器、文本编码器，以及用于不同任务的头部网络。模型通过联合视觉特征和语言特征，能有效完成包括图像标注（Image Captioning）、视觉问答和图文匹配等任务。

BLIP 的基本结构包括图像编码器、文本编码器和针对不同任务的头部网络，如图 3 所示。其中图像编码器使用了基于自注意力机制的前馈网络（例如，Vision Transformer）来提取图像的视觉特征。自注意力模块可以聚合全局信息，使得模型在处理大尺度目标和背景时能够捕捉更丰富的上下文信息。文本编码器使用预训练的语言模型，例如 BERT（Bidirectional Encoder Representations from Transformers）或 GPT 样的结构，负责将输入的文本信息编码成与视觉特征对应的嵌入向量。文本编码器同时支持双向自注意力机制来捕捉语句之间的依赖关系和上下文信息。而根据任务的不同，BLIP 使用了不同的头部网络。例如图文对比学习（ITC，image-text contrastive）用于图文匹配，通过对比学习将图像和文本的对应关系最大化。图文匹配（ITM，image-text matching）用于图像－文本匹配任务。视觉问答任务主要基于图像和问题的联合特征，通过对这两者的融合，来回答视觉相关的问题。其具体工作机制是在图像和文本编码之后，通过交叉注意力模块将两者结合，最终通过解耦的输出层给出答案。

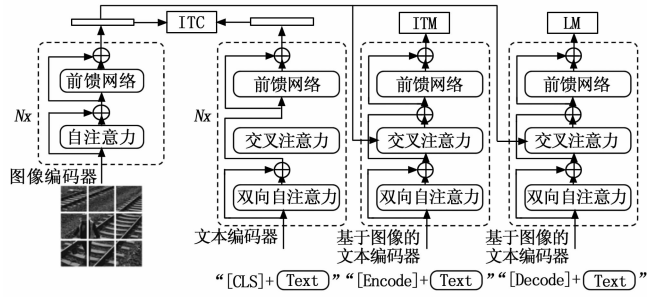


图 3 BLIP 模型结构

在 BLIP 的 VQA 任务中，模型将输入的图像特征和问题文本通过交叉注意力机制进行特征融合。通过这一融合机制，模型可以捕捉到图像和问题之间的相关性，从而在输出层给出答案。在本文的实际应用场景中，模型会根据问题“Is there any foreign object on the railway track?”，提取图像中的关键信息，结合问题理解进行推理，并输出“yes”“no”或“train”的答案。

2 基于通专结合的铁轨异物检测算法

本文提出了一种基于通专结合的铁路异物检测算法，通过模型解耦处理不同光照条件与摄像头模态（如白天可见光、夜晚可见光和红外），显著提升了系统在复杂场景下的检测精度与鲁棒性。

同时，结合 YOLOv7 专用检测模型和 BLIP 多模态模型的深层语义理解能力，设计了双阈值动态调整策略，有效降低漏检率和误检率，为铁路安全监控提供了全面支持。

2.1 实验框架的搭建

如图 4 所示，本文采用模型解耦和通专结合的铁路

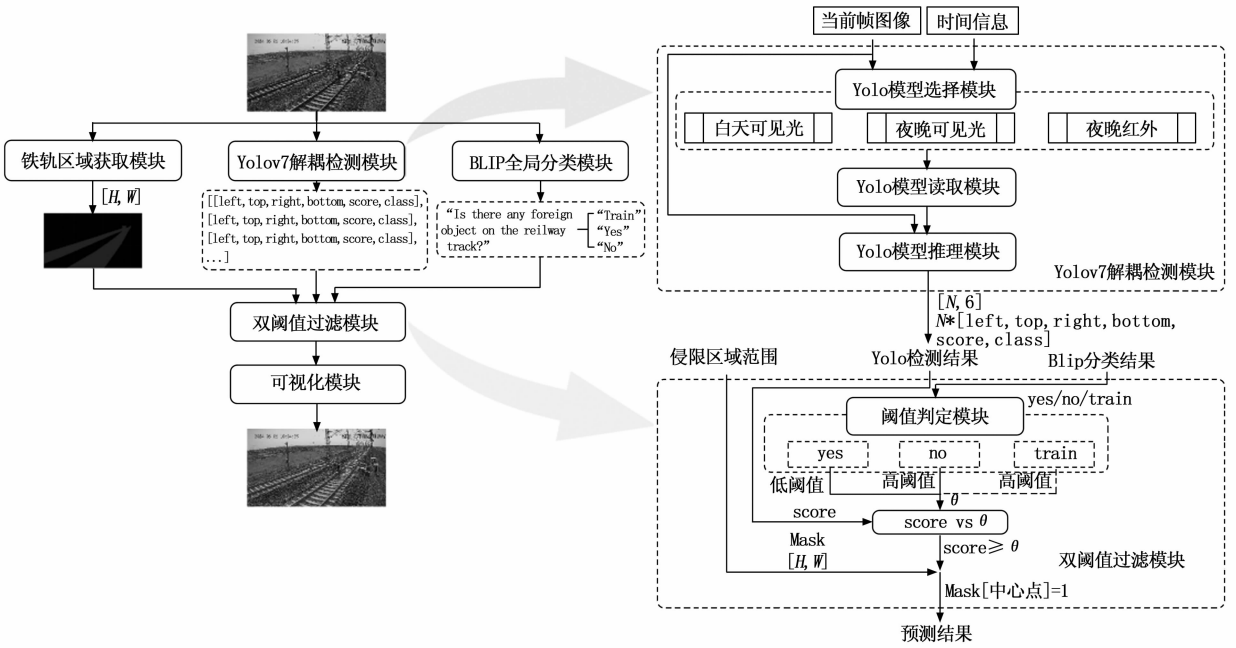


图 4 基于通专结合的铁轨异物检测算法整体框架图

异物检测算法，输入图像在通过智能侵限区域划分之后，利用多个 YOLOv7 专用检测模型进行检测框精确位置的预测，同时充分利用通用预训练大模型中的深层语义信息，提升检测系统对不同天气、不同光线条件、不同背景环境的鲁棒性，通过模型串联的方式挖掘低分检测框，从而降低异物漏检率和误检率，更全面的保障铁轨运行的安全。

具体而言，铁路异物侵限检测系统采用从监控视频流中捕获的图像以及其他相关信息（时间等）作为输入，分别经过侵限区域获取模块、YOLOv7 检测模块、BLIP 全局分类模块对图像进行特征提取，并完成相应的预测。其中，侵限区域获取模块采用 YOLOv8 目标分割模型，完成对图像中的侵限关注区域的像素级预测；YOLOv7 检测模块对全图进行异物检测，在根据输入图像的光线条件和当前系统时间判断当前状态，选择合适的模型并加载，返回异常目标的边界框和分类结果；而 BLIP 全局分类模块对全局进行状态判断。随后，双阈值过滤模块模块根据 BLIP 的结果动态调整检测结果的过滤阈值，同时根据分割的结果进行检测结果的区域过滤。通过这种双阈值的模式，可以充分挖掘 YOLO 中的低分异常检测框，降低漏检率。最后，在将过滤后的 YOLO 结果输入到可视化模块，进行异常帧的可视化标注与帧图像存储。

2.2　铁轨区域的获取

铁轨区域的获取是基于通专结合的铁轨异物检测算法的关键步骤之一，其目的是精确地识别并提取图像中的铁轨区域，以便对铁路沿线的安全状况进行重点监控的同时，排除因为铁轨旁道路车辆、行人经过造成的干扰。铁轨区域的准确识别对于后续的异常目标检测、设备检测和轨道状况评估具有重要意义。传统的铁轨检测方法多依赖于直线检测技术，通常通过霍夫变换等算法检测图像中的直线，假设铁轨是直线状的，或者采取多段直线进行拼接的方式。然而，这种方法在复杂的环境中表现不佳，尤其是在铁轨受到遮挡、铁轨区域形状不规则、不同光照条件下或场景复杂时，传统方法常常无法有效识别并提取铁轨区域。

为了解决这些问题，我们采用了 YOLOv8 目标分割模型来进行铁轨区域的获取。与传统的直线检测方法不同，YOLOv8 能够通过训练集学习到更多复杂的图像特征，使其能够适应不同的环境变化，不仅能够检测识别图像中的铁轨，还可以精确地定位铁轨的具体像素区域，进行像素级的分类，从而在后续的处理过程中，提供更加细粒度的信息，确保对铁路设施的监控更加全面和精确。

YOLOv8 模型在进行目标分割时，首先通过输入图像进行特征提取，利用卷积神经网络从图像中提取高维特征。这些特征包含了图像的局部和全局信息，有助于后续的铁轨检测和分割任务。在特征提取完成后，通过多尺度融合的策略将不同层次的特征进行组合，以更好地处理复杂的场景变化。多尺度特征融合使得模型能够同时捕捉图像中的大物体和小物体，从而提高了对不同尺度目标的检测能力。这些多尺度的特征图通过拼接或加权求和的方式进行融合，增强了模型在复杂环境中的鲁棒性。通过这种融合，YOLOv8 能够学习到更多细粒度的特征，尤其是在铁轨这种形状复杂且容易变形的目标上，能够准确地处理各种变化。随后，通过头部对生成的特征图进行预测，在目标分割任务中，YOLOv8 不仅进行边界框回归，还通过像素级的分类精确标定目标的形状。通过这一过程，YOLOv8 能够为每个像素分配一个类别标签，从而实现像素级的目标分割。这种像素级分割方法，显著提高了对铁轨区域的精确度，尤其是在复杂环境中，能够应对遮挡、变形以及光照变化等挑战。

相较于传统的基于霍夫变换的直线检测方法，YOLOv8 分割模型的优势在于其强大的特征学习能力和对复杂场景的适应性。霍夫变换等传统方法虽然能够有效识别简单场景中的铁轨直线，但在光照变化、铁轨弯曲、遮挡以及噪声影响下，精度和鲁棒性较差。而 YOLOv8 模型能够通过大量数据进行端到端的训练，从而获得针对不同环境和场景的丰富特征表示，大大提高了检测的准确性和可靠性。此外，YOLOv8 分割模型支持实时预测，其推理速度较快，适用于实际监控系统中的实时图像处理需求。因此，YOLOv8 分割模型在铁轨区域获取中的应用，为铁路监控系统提供了更为先进和高效的解决方案。

2.3　YOLOv7 解耦检测

针对铁路监控场景中，光线条件（白天、夜晚）与摄像头模态（可见光、红外）的变化，本文提出了一种基于 YOLOv7 的模型解耦方案。该方案通过将不同光照和传感器模态下的目标检测任务分别解耦为独立的子模型，从而优化了模型在各类场景下的表现。具体而言，针对白天的可见光图像、夜间的可见光图像以及红外图像，分别设计了适配性更强、计算更加高效的检测模型，使得每个子模型能够专注于特定的视觉特征，提高了整体检测精度与推理速度，同时有效减少了资源消耗。

一般而言，白天的可见光图像光照条件较好，夜晚可见光和红外光照条件较差，且常常伴随较强的噪声。在分别使用白天可见光、夜间可见光、夜间红外的数据训练相对应的检测模型的基础上，采用低光增强和噪声

抑制的方式对夜间可见光和夜间红外的检测算法进行针对性改进。具体而言，对于夜间可见光图像，首先利用低光增强技术（如直方图均衡化、Retinex 算法等）提升图像的亮度和对比度，增强细节表现；然后，应用噪声抑制算法，如高斯滤波或非局部均值去噪，去除图像中的噪声，避免噪声对目标检测性能的干扰。通过这些针对性的改进，解耦模型在夜间可见光和红外图像的目标检测中能更好地适应不同光照环境，提升检测准确性，进而提高铁路监控系统在各种复杂环境下的稳定性和可靠性。

另外，需要说明的是，在本研究中，我们针对分割与检测任务分别选择了 YOLOv8 和 YOLOv7。YOLOv8 引入了原生的分割头结构，具备端到端的实例分割能力，适用于需要精确轮廓提取的任务。尽管 YOLOv8 在某些场景下表现优异，但在铁路场景的目标检测任务中，我们通过实测发现 YOLOv7 在精度和稳定性方面更具优势。YOLOv7 采用 anchor-based 架构，实现了高精度和实时性的平衡。因此，为了充分发挥各自模型的优势，我们在分割任务中采用了 YOLOv8，而在检测任务中选择了 YOLOv7。

2.4 通专结合：双阈值异常判定

为了提高铁路轨道监控中的异常检测精度，本系统结合了 BLIP 全局分类模块与 YOLOv7 目标检测模块，通过双阈值判定策略来优化异常目标的识别和过滤。首先，BLIP 模型通过全局分类任务对图像进行状态评估，具体来说，模型会被问到：“Is there any foreign object on the railway track?”，根据输入的图像，BLIP 模型会给出三个可能的答案：“yes”，“no”或“train”。根据 BLIP 的判断结果，系统将在 YOLOv7 检测结果中应用动态的阈值策略，从而确定是否需要进一步处理检测到的目标。

具体地，设 BLIP 给出的答案为 A，系统根据 BLIP 的答案来设定不同的阈值 θ，θ 的值定义如下。

$$\theta=\begin{cases}\theta_h, A=\text{‘no’ or } A=\text{‘train’}\\ \theta_l, A=\text{‘yes’}\end{cases} \tag{1}$$

式中，θ_h 为高阈值，θ_l 为低阈值，对于 YOLOv7 给出的第 i 个检测目标 o_i，其置信度为 s，则可得双阈值过滤后的检测目标 p_i。

$$p_i=\begin{cases}o_i, s\geqslant\theta\\ \varnothing, s<\theta\end{cases} \tag{2}$$

式中，$\varnothing$ 表示当置信度低于阈值时将该目标删除，不计入过滤后结果中。通过这种基于 BLIP 分类结果的双阈值策略，系统能够根据实际情况动态调整检测的敏感度。在“yes”的情况下，较低的阈值能够有效挖掘 YOLOv7 中的低分异常检测框，降低漏检率，提升对微小或模糊异常目标的检测能力。而在“train”和“no”情况下，系统能减少误报，提升了整体系统的效率和准确性。最后，经过双阈值判断过滤后的 YOLOv7 检测结果会被传递至可视化模块进行后续的异常标注与帧图像存储，便于人工查看与系统后续分析。

3 实验与结果分析

3.1 实验环境与参数设置

算法模型训练及测试环境如表 1 所示。

表 1 实验环境

配置名称	参数（版本）
操作系统	Ubuntu 20.04.1
CPU	Intel（R）Xeon（R）Gold 6226R CPU @ 2.90 GHz
GPU	NVIDIA GeForce RTX 3090 Ti * 2
算法框架	PyTorch（1.11.0）
CUDA	11.3

在基于 YOLOv7 的解耦检测模型训练的过程中，采用随机梯度下降（SGD，stochastic gradient descent）优化器以减少模型陷入局部最优的风险。网络初始学习率设为 0.001，动量因子设为 0.937，采用线性衰减（Linear Decay）学习率调整策略，学习率的衰减因子设置为 0.1。批次大小（batch-size）设置为 16，输入图像大小调整为 1 280×1 280，总共训练 250 轮。

在对 BLIP 模型进行微调时 batch size 设置为 16，学习率设置为，迭代轮数设置为 20；利用 UPop 进行剪枝时注意力掩码矩阵正则化参数设置为 0.014 4，多层感知器掩码矩阵正则化参数设置为，再训练中模型学习率设置为；对 YOLOv8 进行训练时 batch size 为 16，迭代次数为 200，IoU 阈值设置为 0.5，学习率设置为 0.01，图像尺寸为 640 * 640。

3.2 数据集采集与处理

本研究聚焦于铁路环境下的异物侵限与遗留检测问题。该场景具有高度专业性和结构特性，铁路轨道具有线性延展、对称双边、背景单一却易被干扰（如异物、残留器件、环境设施遮挡）等特点。因此，为了评估本文改进的通专结合铁轨异物检测算法的优越性，本文自行采集、处理并构建了铁路异物侵限与遗留数据集进行训练与测试。与 COCO 等开源通用数据集相比，我们的铁路场景数据集具有更高的任务耦合度。开源通用数据集更侧重多类别、多尺度、日常场景下的通用识别，而非面向单一工业结构中异常目标的精准判断。因此，若使用开源通用数据集，不仅会导致任务目标泛化、评估失真，还无法体现所提方法在光照模态变化、天气变化的复杂场景下的优势。

本数据集来源于 53 个铁轨沿线摄像机拍摄的监控

视频数据，由于现有的带监控摄像头的铁路线路大多数在工作状态，能采集到的视频数据中，铁轨上几乎没有侵限事件和遗留物事件发生，目标种类十分有限，且正负样本数量差异过大。因此，本文使用实地采集的铁轨图像作为背景、铁轨场景可能出现的异物为前景，采用 Copy-Paste 方案[20]进行数据合成。

具体而言，首先我们从数据中筛选出无异物、无遮挡且能清晰展现铁轨结构的图像作为背景池。为保证后续合成的标准化，所有背景图像均被统一缩放至 1 280 ×1 280 像素，并保留了不同天气（雨天，雪天，雾天，晴天）与光照（白天，夜晚以及夜晚红外）下的多样化场景。

前景物体的选取主要遵循“高危性”与“场景相关性”两大原则。高危性目标依据铁路安全规程，涵盖了对行车安全构成严重威胁的类别，如行人、车辆等；场景相关性目标则模拟了线路中常见的遗留物或障碍物，如施工工具、自然掉落物及生活废弃物等，具体类别如表 2 所示。

表 2　铁路异物前景类别与相关信息

异常目标种类	重要程度	样本数
车辆	重要	200
行人	重要	217
自行车	重要	99
摩托车	重要	102
动物	次要	148
煤块	次要	77
石头	重要	100
树枝	一般	89
铁锹	重要	108
水桶	次要	101
瓶子	一般	135
纸箱	一般	99
袋子	一般	100

为最大限度地减少合成痕迹、确保合成图像的真实性，我们设计了如下精细化的多阶段融合流程。

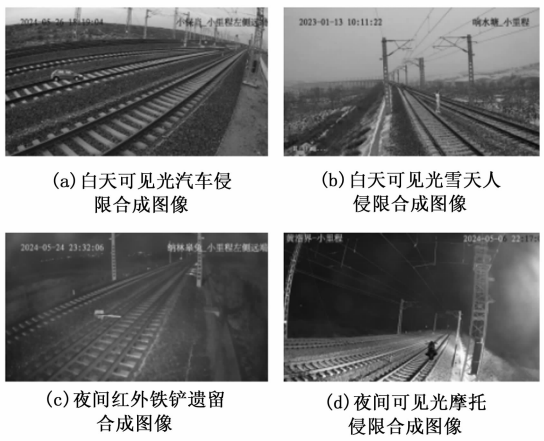
1）尺度与透视匹配：根据相机视角定义的透视规则，前景物体会依据其在图像中的纵向粘贴位置，在合理的尺度范围内进行随机缩放，以模拟“近大远小”的真实距离感。

2）光照与颜色融合：在粘贴前，系统会分析前景粘贴位置的局部背景邻域的颜色直方图。随后，通过直方图匹配算法对前景物体的亮度、对比度和饱和度进行归一化，使其色调分布与局部背景光照环境趋于一致。

3）边缘平滑处理：为消除前景物体边缘的尖锐感，我们对其掩码的边缘进行高斯模糊处理，实现前景与背景之间更加柔和、自然的过渡。

4）阴影模拟生成：为提升合成物体的空间真实感，我们根据场景主光源方向，为前景物体生成了半透明的模拟阴影。该阴影由前景掩码经过压扁、倾斜和高斯模糊后形成，并叠加在前景与背景之间。

部分合成图像示例如图 5 所示。综上所述，在我们采集的真实数据的基础上加入合成数据，共获得白天可见光图像 4 367 张、夜晚红外图像 3 395 张、夜晚可见光图像 2 232 张，对其进行检测目标标注和全局分类标注后，分别构建检测数据集与 VQA 数据集。



(a) 白天可见光汽车侵限合成图像　(b) 白天可见光雪天人侵限合成图像

(c) 夜间红外铁铲遗留合成图像　(d) 夜间可见光摩托侵限合成图像

图 5　Copy-Paste 与多种增强合成铁轨异物图像示例

3.3　性能评估指标

为了更全面的评估改进后的异物检测算法的效果，采用常用的目标检测评估指标和全局漏报率、误报率对方法进行准确性性能衡量。常用的目标检测评估指标包括精确率（Precision）、召回率（Recall）、平均精度（AP，average precision）平均精度均值（mAP，mean average precision），其表达式如下：

$$Precision = \frac{TP_{\text{object}}}{TP_{\text{object}} + FP_{\text{object}}} \tag{3}$$

$$Recall = \frac{TP_{\text{object}}}{TP_{\text{object}} + FN_{\text{object}}} \tag{4}$$

$$AP = \frac{1}{N}\sum_{i=1}^{N} Precision(r_i) \tag{5}$$

$$mAP = \frac{1}{C}\sum_{i=1}^{C} AP_i \tag{6}$$

式中，TP_{object}、FP_{object}和 FN_{object}为针对目标的统计个数，而对于全局而言的漏报率 R_{miss}和误报率 R_{false}表达式如下：

$$R_{\text{miss}} = \frac{FN_{\text{image}}}{TP_{\text{image}} + FN_{\text{image}}} \tag{7}$$

$$R_{\text{false}} = \frac{FP_{\text{image}}}{TN_{\text{image}} + FP_{\text{image}}} \tag{8}$$

式中，TP_{image}、FP_{image}和 FN_{image}为针对图像的统计个数。

除此以外，为了衡量模型的实时性性能，还采用了

帧率（FPS，frames per second）、参数量、处理耗时、每秒浮点运算次数（FLOPs，floating point operations per second）等速度指标。

3.4 实验结果分析

3.4.1 检测模型对比实验

为了验证本文提出的通专结合铁轨异物检测算法在铁路异物侵限检测中的有效性，我们将其与主流的目标检测算法进行了对比实验，具体包括 SSD、YOLOv5 和 YOLOv7。实验结果如表 3 所示。

表 3 对比实验结果

实验算法	*Precision*	*Recall*	*mAP*
SSD	0.860	0.866	0.860
YOLOv5	0.894	0.709	0.816
YOLOv7	0.953	0.886	0.884
Ours	0.953	0.891	0.889

首先，实验结果表明，基于通专结合方法的改进 YOLOv7 架构在检测精度上显著优于当前主流的目标检测算法。与 SSD 相比，本文模型在 *mAP* 指标上提高了 2.9%，与 YOLOv5 相比，本文方法的 *mAP* 提升了 7.3%，且在复杂背景下的目标检测性能表现更为稳定。此外，和 YOLOv7 相比，改进后的检测算法在 *mAP* 上提升了 0.5%。综上所述，基于通专结合的异物检测算法在铁路异物侵限检测任务中展现出了较强的性能优势，相较于现有主流检测模型具有显著的提升。

3.4.2 细分场景下的性能分析实验

为了更加全面评估基于通专结合算法的性能优势，本文从目标尺度和光照条件两个关键维度，对各模型在多个细分场景下的性能进行了系统的对比分析。

1）多尺度目标的检测性能分析：

铁路场景中异物在尺寸和形态上差异显著，从大型车辆到微小瓶罐不等，这给检测算法的尺度适应性带来了严峻挑战。针对这一问题，我们选取了覆盖大、中、小三种尺度的六类典型目标，包括大尺度目标（车辆，行人），中尺度目标（铁锹，纸箱）和小尺度目标（石头，瓶子），并对各算法在这些目标上的检测结果进行了系统比较，结果如表 4 所示。

表 4 各模型针对多尺度目标的 *AP* 结果

	车辆	行人	铁锹	纸箱	石头	瓶子
SSD	0.925	0.881	0.865	0.870	0.832	0.815
v5	0.928	0.875	0.841	0.842	0.753	0.738
v7	**0.955**	0.916	0.891	0.901	0.842	0.825
Ours	**0.948**	**0.931**	**0.898**	**0.910**	**0.863**	**0.851**

实验结果表明：在车辆、行人等大尺度目标上，由于其特征信息较为丰富，各类先进检测算法均能取得较优的性能，因此本文方法与基线 YOLOv7 相比优势并不明显。其次，在铁锹、纸箱等中等尺度目标上与 YOLOv7 基线相比，本文方法在 *AP* 值上表现出稳定提升，平均增幅约为 0.8%。

本文方法在小尺度目标的检测上展现出相比于其余模型的核心优势。对于石头、瓶子等像素占比极小、且易与背景混淆的高难度目标，基线 YOLOv7 的性能明显受限，而本文方法得益于“通专结合”策略，在这两个类别上平均精度分别实现了 2.1% 和 2.6% 的显著提升。

2）不同光照条件下的检测性能分析：

铁路沿线监控系统需实现 7×24 小时连续工作，因此光照条件随时间变化而极为复杂。为验证各算法在不同光照环境下的鲁棒性，我们将测试集划分为“白天可见光”“夜晚可见光”与“夜晚红外”三个子集，并分别评估了各算法在不同子集上的 *mAP* 表现，结果如表 5 所示。

表 5 不同算法在不同光照条件下的 *mAP* 结果

	白天可见光	夜晚可见光	夜晚红外
SSD	0.875	0.792	0.801
v5	0.831	0.758	0.765
v7	0.898	0.825	0.833
Ours	0.911	0.863	0.869

实验结果表明：在进入夜晚可见光和夜晚红外等低光照、高噪声的复杂场景时，本文提出的方法表现出更强的鲁棒性，*mAP* 分别较 YOLOv7 提升 3.8% 和 3.6%。

综上，通过对多尺度目标以及多样化光照条件的深入分析，进一步验证了本文所提出的基于全局语义与局部特征融合的方法在复杂铁路场景中具备显著的性能优势与鲁棒性，尤其在小尺度目标检测和低光照环境下表现突出。

3.4.3 系统消融实验

为了验证 BLIP 全局分类模块对改进 YOLOv7 模型在铁路异物侵限检测中的有效性，我们设计了消融实验，分别评估 YOLOv7 单独模型、BLIP 单独模型、YOLOv7 与 BLIP 联合模型对系统异物检测与告警能力的影响。为了验证现实场景的效果，我们采用了现场采集的真实数据构建测试集，评估系统的误报率和漏报率结果，实验结果如表 6 所示。通过在 YOLOv7 的输出结果中加入 BLIP 分类的全局判断，我们能够在全局出现异常的情况下挖掘低分异常框，有效降低漏检率。实验中，YOLOv7 与 BLIP 联合模型的漏报率上较 BLIP 单一模型下降了 1.92%，较 YOLOv7 单一模型下降了 0.44%，说明 BLIP 的全局信息处理能力和 YOLOv7 对

小目标的特征提取能力均对异常的识别起到了关键作用。

表 6　消融实验结果

实验编号	YOLOv7	BLIP	误报率	漏报率
1	×	√	0.00%	2.14%
2	√	×	1.09%	0.66%
3	√	√	1.09%	0.22%

综上所述，BLIP 模块的加入显著提高了 YOLOv7 模型的检测性能，特别是在复杂环境下，提升了模型的鲁棒性和检测精度，证明了全局分类模块与局部目标检测模型结合的有效性。

3.4.4　速度指标测试实验

为了评估 YOLOv7 与 BLIP 通专结合模型的推理速度，我们在不同分辨率下对模型进行了速度测试。测试环境为 NVIDIA GTX 3090 Ti GPU，输入图像分辨率分别为 1 280×1 280 和 640×640。在标准分辨率下，YOLOv7 模型耗时 30.8 ms，而加入 BLIP 模块后，由于 BLIP 耗时 135.1 ms，整体的 FPS 将略有下降，如表 7 所示。YOLOv7 与 BLIP 通专结合模型在精度提升的同时，仍能维持一定的推理速度。特别是在检测精度和小目标检测能力上，本文提出的模型在保证检测精度的同时，推理速度和计算资源消耗均能够满足铁路轨道安全监控的需求。

表 7　各部分模型参数统计及速度测试

	参数量/M	处理速度/ms	FPS	FLOPs/G
BLIP	361.59	135.1	7.4	186.12
YOLOv7	70.4	30.8	32.4	360.0
YOLOv8-Seg	46	40.61	24.6	220.5

4　结束语

本文提出了一种基于通专结合的铁路异物侵限及遗留检测，针对铁路轨道监控中常见的环境复杂性、光照变化以及小目标检测的挑战，设计了一种多层次、动态调整的检测策略。通过对不同光照条件和摄像头模态（如白天可见光、夜间可见光、红外）进行模型解耦，本文的方法在不同环境下展现出了较高的适应性和检测性能，尤其在低光照和复杂背景下具有显著的优势。此外，本文结合专用模型与通用大模型，提出了双阈值判定策略，能够根据全局分类结果动态调整检测的敏感度。这种策略不仅有效提升了低分目标的检测精度，降低了漏检率，同时还减少了误报，确保了系统在多种实际应用场景下的稳定性和可靠性。实验表明，改进后的模型在准确性、计算效率和资源消耗之间达到了较好的平衡，且具有较强的实用价值和扩展性。

尽管本文提出的方法有效改善了铁路场景异物检测的效果，但仍面临一些挑战，如极端天气条件下的检测效果和数据集的多样性问题。未来，随着数据集的不断扩展和模型算法的优化，特别是在低照度和噪声抑制方面的进一步研究，本文将进一步改善算法，为铁路交通的安全运行提供更加可靠的技术支持。

参考文献：

[1] 陈小屹. 基于图像处理的铁路轨道异物入侵检测研究[D]. 兰州：兰州交通大学，2023.

[2] 史红梅，柴　华，王　尧，等. 基于目标识别与跟踪的嵌入式铁路异物侵限检测算法研究[J]. 铁道学报，2015，37 (7)：58－65.

[3] TENG Z，LIU F，ZHANG B. Visual railway detection by superpixel based intracellular decisions[J]. Multimedia Tools and Applications，2016，75 (5)：2473－2486.

[4] 侯　涛，伍海萍，牛宏侠. 改进 MOG-LRMF 的铁轨动态异物检测[J]. 交通运输系统工程与信息，2020，20 (2)：91－100.

[5] 徐　岩，陶慧青，虎丽丽. 基于 Faster R-CNN 网络模型的铁路异物侵限检测算法研究[J]. 铁道学报，2020，42 (5)：91－98.

[6] 李建国，陈敬涛，张　伟，等. 基于改进型 SSD 算法的铁路货场异物侵限小目标检测研究[J]. 铁道通信信号，2024，60 (7)：57－62.

[7] 张　剑，王等准，莫光健，等. 基于改进 YOLOv3 的高铁异物入侵检测算法[J]. 计算机技术与发展，2022，32 (2)：69－74.

[8] 陈伟迅，柯旭能，孟思明. 基于改进 YOLOv8 的铁路异物侵限检测方法[J]. 机电工程技术，2024，53 (11)：211－214.

[9] RADFORD A，NARASIMHAN K，SALIMANS T，et al. Improving language understanding by generative pre-training[Z]. 2018.

[10] RADFORD A，WU J，CHILD R，et al. Language models are unsupervised multitask learners[J]. OpenAI blog，2019，1 (8)：9.

[11] BROWN T，MANN B，RYDER N，et al. Language models are few-shot learners[J]. Advances in neural information processing systems，2020，33：1877－1901.

[12] RADFORD A，KIM J W，HALLACY C，et al. Learning transferable visual models from natural language supervision[C]//International conference on machine learning，PmLR，2021：8748－8763.

[13] RAMESH A，PAVLOV M，GOH G，et al. Zero-shot text-to-image generation[C]//International conference on machine learning，Pmlr，2021：8821－8831.

（下转第 50 页）