中图法分类号:TP391.41; TH74　　**文献标识码**: A　　**文章编号**: 1006-8961(2026)10-3653-18

论文引用格式:Wu J, Cai G Z, Chu H X, Xu G, Zhao X M and Yin H. 2026. Small-object detection network for wide-field surveillance video: SOD-YOLO. Journal of Image and Graphics, 31(10):3653-3670(吴军, 蔡广震, 楚和轩, 徐刚, 赵雪梅, 尹恒. 2026. 用于大视场监控视频的小目标检测网络SOD-YOLO. 中国图象图形学报, 31(10):3653-3670)[DOI:10.11834/jig.250491]

用于大视场监控视频的小目标检测网络SOD-YOLO

吴军[1,2],蔡广震[1],楚和轩[1],徐刚[1],赵雪梅[1,3],尹恒[1*]

1. 桂林电子科技大学电子工程与自动化学院,桂林　541004; 2. 广西科学院,南宁　530000;
3. 智能综合自动化广西高校重点实验室,桂林　541004

摘　要:**目的**　针对大视场监控视频中小目标检测存在的样本稀缺、特征微弱与定位不准等难题,提出一种用于大视场监控视频的小目标检测网络SOD-YOLO(small object detection-you only look once)。**方法**　该方法从3个层面系统性地进行改进:1)提出结合SAM2(segment anything model 2)语义指导与UE虚拟仿真的虚实融合样本生成策略,以低成本获取大量高质量标注数据;2)设计包含视频差分预处理、多尺度特征融合及双层路由注意力的特征增强机制,以提升模型对微小特征的感知与分辨能力;3)采用解耦头结构并创新α-CIoU(alpha-complete intersection over union)损失函数,以优化小目标边界框的回归精度。**结果**　在建设工地、高速公路服务区及大学校园3个真实场景数据集上的实验结果表明,SOD-YOLO的综合性能显著优于当前主流模型,在建筑工地场景取得最高AP75(13.5%)与AP50s(53.9%);在高速公路服务区场景获得最优AP(42.6%)与AP75(29.5%);尤其在极具挑战的大学校园场景(小目标像素占比约0.007 5%),其AP、AP75与AP50s相较基准模型YOLOv7分别提升了4.9%、2.5%与5.0%。**结论**　本研究为解决低分辨率监控视频下的小目标检测问题提供了一套有效且可靠的技术方案。

关键词:视频监控;卷积神经网络(CNN);小目标检测;小目标边界框定位;样本增广

Small-object detection network for wide-field surveillance video: SOD-YOLO

Wu Jun[1,2], Cai Guangzhen[1], Chu Hexuan[1], Xu Gang[1], Zhao Xuemei[1,3], Yin Heng[1*]

1. School of Electronic Engineering and Automation, Guilin University of Electronic Technology, Guilin　541004, China;
2. Guangxi Academy of Sciences, Nanning　530000, China; 3. Key Laboratory of Intelligent Integrated Automation, Guangxi Higher Education Institutions, Guilin　541004, China

Abstract: **Objective** Small-object detection in large-field-of-view (FOV) surveillance videos remains a persistent challenge in real-world applications such as construction-site safety monitoring, traffic supervision in highway service areas, and public-space management on university campuses. Compared with conventional close-range datasets, targets in large-FOV imagery are often extremely small and are heavily affected by background clutter, illumination variation, compression noise, motion blur, and frequent occlusion. Therefore, detectors trained on generic benchmarks typically face three intertwined bottlenecks: 1) scarce and costly annotations for tiny objects across diverse surveillance scenes; 2) weak and

收稿日期:2025-10-03;**修回日期**:2026-03-02;**预印本日期**:2026-03-09

*通信作者:尹恒　yinh@guet.edu.cn

基金项目:国家自然科学基金项目(42361071,42261061);广西重点研发计划资助(ZG2504240008);广西研究生教育创新计划项目(YCBZ2024165)

Supported by: National Natural Science Foundation of China (42361071, 42261061); Guangxi Key R&D Program, China (ZG2504240008); Guangxi Postgraduate Education Innovation Program (YCBZ2024165)

ambiguous visual features caused by limited pixel coverage and low signal-to-noise ratio, which makes tiny targets difficult to distinguish from the background; and 3) inaccurate localization, where even a deviation of a few pixels can substantially reduce intersection over union (IoU), thereby harming high-precision metrics and downstream tracking or analysis. To address these issues in a unified manner, this study proposes SOD-YOLO, a small-object-oriented detector designed for low-resolution, large-FOV surveillance videos. The primary goal is to enhance supervision signals, strengthen tiny-feature representation, and improve high-precision localization simultaneously, thereby delivering reliable performance under practical surveillance constraints. **Method** SOD-YOLO follows a data-feature-localization co-design philosophy and introduces improvements in three aspects. 1) Virtual-real fusion sample generation. To alleviate the scarcity of real annotations and increase the diversity of tiny-object appearances and backgrounds, we propose a cost-effective data-generation strategy that combines SAM2 (segment anything model v2) semantic guidance with unreal engine (UE) virtual simulation. SAM2 is used to obtain fine-grained object masks and semantic cues, enabling high-quality object cutout, composition, and accurate label transfer for small objects. UE is used to generate controllable virtual surveillance scenes with precise ground-truth annotations, in which camera viewpoints, object-scale distributions, lighting conditions, and scene layouts are configured to mimic real surveillance patterns. These synthetic data are mixed with real samples to expand the training distribution, improving robustness to background variation and rare target configurations while keeping labeling costs manageable. 2) Tiny-object feature enhancement. Because tiny targets provide weak cues that are easily overwhelmed in large-FOV frames, we design a feature-enhancement mechanism with three components. First, video differential preprocessing explicitly highlights subtle foreground changes and suppresses irrelevant static textures. Second, strengthened multi-scale feature fusion better integrates fine-grained spatial detail from shallow layers with semantic context from deeper layers, allowing the network to preserve small-object signals while retaining discriminative context. Third, bilevel routing attention (BRA) is incorporated to allocate attention adaptively to informative regions and improve the model's ability to perceive and distinguish tiny targets under clutter and occlusion. 3) Localization-oriented optimization. To address the strong localization sensitivity of small objects, SOD-YOLO adopts a decoupled head that separates classification and regression learning, reducing task conflict and stabilizing optimization when supervision is weak. In addition, we propose an α-CIoU loss to improve regression precision and encourage better alignment at higher IoU thresholds, directly targeting the common failure mode of loose or shifted bounding boxes for tiny objects. Collectively, these designs form an end-to-end detector that improves learning from limited data, enhances tiny-feature representation, and refines bounding-box regression. **Result** Extensive experiments were conducted on three real-scene large-FOV surveillance datasets: construction site, highway service area, and university campus, covering multi-category detection of pedestrians, e-bikes, cars, trucks, and buses, as well as extremely small targets under complex backgrounds. Evaluation followed COCO-style metrics, with particular emphasis on high-IoU and small-object performance. On the construction-site dataset, SOD-YOLO achieved the best localization-sensitive performance, with an AP75 of 13.5% and an AP50s of 53.9%, indicating superior high-precision bounding-box regression and sensitivity to small targets in cluttered scenes. On the highway service-area dataset, which is characterized by dense traffic and large intra-class variation, SOD-YOLO reached the best overall AP of 42.6% and an AP75 of 29.5%, demonstrating robust general performance and high-IoU accuracy under complex multi-class conditions. On the especially challenging university-campus dataset, where the smallest objects may occupy only about 0.007 5% of image pixels and minor offsets severely degrade IoU, SOD-YOLO clearly outperformed the YOLOv7 baseline, improving AP, AP75, and AP50s by approximately 4.9%, 2.5%, and 5.0%, respectively. To further verify generalization, we evaluated SOD-YOLO on the public VT-TOD visible benchmark under the official split. The model achieved leading performance, with a precision of 61.4%, an AP50:95 of 52.3%, an APs of 63.5%, and an APm of 58.2%, outperforming representative methods such as QueryDet, YOLOC, and Drone-YOLO. Ablation studies confirmed that every proposed component contributes positively: differential preprocessing produced the most notable gain by enhancing foreground saliency; multi-scale fusion and BRA provided additional gains by strengthening discriminative representations; and the decoupled head together with α-CIoU consistently improved high-IoU localization quality. Experiments on virtual-real mixing ratios showed that a balanced fusion strategy best exploits synthetic diversity while preserving realism, whereas overreliance on synthetic data can reduce generalization because of domain gaps. **Conclusion** This study proposes SOD-YOLO as an effective and reliable

solution for small-object detection in low-resolution, large-FOV surveillance videos. By integrating virtual-real fusion data generation, tiny-feature enhancement, and localization-oriented optimization, SOD-YOLO improves both robustness and precision across diverse real surveillance scenarios and a public benchmark, particularly under high-IoU and small-object settings. The proposed framework provides a practical pathway toward deploying accurate small-object detectors in real monitoring systems where annotation scarcity, weak features, and localization sensitivity are critical constraints.

Key words: video surveillance; convolutional neural network (CNN); small object detection; small object bounding-box localization; sample augmentation

0 引　言

相比于窄视场监控摄像机，大视场(large field of view，FOV)监控摄像机可覆盖更大区域、有利于捕捉低速运动物体以便实施智能化分析(Shen 等，2017；Ye 等，2022)，在交通、工业监控(Liang 等，2024)以及安全等场景应用中获得广泛重视。然而，低成本的大视场监控摄像机多通过超短焦距获取宽广视野(Cao 等，2025)，导致其场景目标影像分辨率偏低，使得诸如行人、小汽车等监控画面中仅占少量像素区域的小尺寸目标检测存在特征表达弱、边界模糊及背景干扰(Liu 等，2021a；Wang 等，2021)等诸多挑战。近年来，以卷积神经网络(convolutional neural network，CNN)为代表的深度学习技术在目标检测视觉任务中表现出色，成为当前主流研究方向，从这一角度出发并考虑场景特点，现有的监控视频小目标检测方法和技术可概述分为两类：室内特定场景小目标检测和室外复杂场景小目标检测(Zou 等，2023；Bai 等，2018；潘晓英 等，2023)。

室内特定场景中目标运动范围较小且轨迹规律(如仓储 AGV、人员固定动线)，其视频监控设备视野受限但成像质量稳定，可通过静态背景建模实现高效的(小)目标检测，主要面临特定视角下的目标尺寸变化、遮挡等挑战，如：Dosovitskiy 等人(2021)建立基于 Transformer 的检测框架并通过自注意力机制增强特征表示能力，在低对比度和动态背景下提高了小目标检测精度；Rasheed 等人(2024)结合特征金字塔网络和 EfficientNet 提升工业零件检测精度，在复杂背景和微小缺陷检测方面表现突出；Bryan 等人(2019)利用 CenterNet 增强仓储物流中小物品定位精度，可有效应对物品遮挡和复杂背景问题；Jiang 等人(2024)提出多尺度特征融合小目标检测网络，以增强无人机影像中不同尺度小目标的特征表达；Kulambayev 等人(2023)联合 MobileNetV3 与 Faster R-CNN (faster region-based convolutional neural network)(Ren 等，2017)建立轻量级检测架构并在嵌入式设备中实现实时检测；Liu 等人(2021b)提出一种基于 Swin Transformer 的目标检测方法并利用层次化注意力机制提高模型对小目标的敏感度；针对多物体遮挡问题，Cai 和 Vasconcelos(2018)通过增强级联 R-CNN 上下文感知能力使模型在多物体遮挡情况下更准确地识别小目标。

室外复杂场景小目标检测涉及智能交通、安全监控、无人机侦察以及遥感影像分析等众多应用领域，面临动态背景、光照变化及远距离目标带来的特征模糊、易受遮挡等诸多挑战，如 Wang 等人(2025)针对无人机航拍场景设计了一种多尺度特征融合网络以提升远距离小目标的检测性能；为改善卫星图像中的检测效果，Hao 等人(2024)结合超分辨率技术增强小目标特征，Zhang 等人(2023)则引入注意力机制有效抑制复杂背景干扰；为解决轻量化部署问题，Liu 等人(2023)研究 YOLO(you only look once)轻量化技术问题并在边缘设备上实现目标识别，从而有效服务于智能交通领域应用；针对野生动物监测需求，Zhou 等人(2024)提出了一种基于红外图像深度学习算法以提升夜间和恶劣天气下的检测精度，Zeng 等人(2024)则通过图像增强与自适应锚框优化提高小目标检测的鲁棒性。

为进一步突破小目标特征提取的技术瓶颈，研究者们在特征增强与结构创新方向持续探索：早期经典工作中，Deng 等人(2022)提出的扩展特征金字塔网络(extended feature pyramid network，EFPN)通过强化高分辨率特征层的语义信息传递，缓解了传统特征金字塔中浅层细节与深层语义脱节的问题，为多尺度小目标检测奠定了重要基础；近年研究更注重结构适配性结合动态优化，Yang 等人(2025)设计的风车形卷积(PConv)针对小目标高斯像素分布

特性优化感受野，配合尺度动态损失根据目标尺寸自适应调整权重，显著提升红外场景中小目标的特征分辨能力；Wu等人（2026）则面向资源受限场景，提出融合空间通道注意力的轻量级多尺度网络，通过动态特征聚焦机制在保证推理效率的同时强化微小目标特征，为大视场监控这类需实时检测的场景提供了参考。在网络架构与场景适配层面，Chen等人（2024）将嵌套式Mamba结构（Mamba-in-Mamba for efficient infrared small-target detection，MiM-ISTD）应用于红外小目标检测，通过外层Mamba捕获全局上下文、内层Mamba建模局部特征的协同模式，以线性计算复杂度实现高效特征关联，突破了传统CNN在长距离依赖捕捉上的局限；针对与大视场监控相似的远距离复杂光照场景，Zheng等人（2025）提出的LAM-YOLO引入光照—遮挡注意力（lighting-occlusion attention，LAM）模块，结合SIB-IoU（intersection over union，IoU）损失与专用小目标检测头，在VisDrone数据集上较YOLOv8实现7.1%的平均精度（average precision，AP）提升，验证了注意力机制对复杂环境的适配价值。跨模态融合与高质量数据集构建也成为重要研究方向，Ying等人（2025）构建的可见光—热红外小目标基准数据集（RGB-T-Tiny）包含1.2 M条标注，并提出尺度自适应评估指标（scale-adaptive fitness，SAFit），为跨模态小目标检测的公平对比提供了关键支撑。此外，针对本文关注的建筑工地场景，Zhang和Zhao（2025）虽然基于RT-DETR（real-time detection Transformer）（Zhao等，2024）改进了小目标检测算法，但在极低像素占比目标的定位精度上仍存在提升空间。

相对于室内特定场景，室外目标（车辆、行人）运动范围大且具有随机性，依赖广角镜头（水平视场角≥90°）实现大视场监控目的且需处理多目标遮挡（如人群密集区域）和跨尺度运动（如近远距离目标切换）的挑战，故其小目标检测结果面临更多不确定性，尤其是网络模型场景泛化。

目前，YOLO系列网络广泛用于视频目标检测任务，该网络处理大尺寸目标时表现良好且计算高效，但大视场监控视频小目标检测性能仍有待提升（Diwan等，2023；Redmon等，2016；Hussain，2023）；另一方面，目前小目标检测任务普遍面临数据集稀缺、样本分布不均及标注精度差等问题，如常用公开数据集COCO（common objects in context）（Lin等，2014）中仅52.3%的图像中包含小目标且因像素占比非常小而导致其人工标注质量较低，从而影响模型训练精度及场景泛化性能。针对上述问题，本文以YOLOv7（Wang等，2023）为基础框架设计小目标检测网络SOD-YOLO（small object detection-you only look once）。选择YOLOv7主要基于以下3个方面考量：1）在性能基线平衡方面，以大学校园场景为例，YOLOv7在多个场景的小目标平均精度（AP）均优于YOLOv8/v10。同时其较小的参数量为后续集成视频差分、双层路由注意力（bilevel routing attention，BRA）等新增模块预留了充足的计算冗余，而其基础推理速度仍能满足大视场监控的实时性需求。2）在架构扩展性方面，YOLOv7的主干网络与检测头结构相对独立，无强耦合设计，这为自定义修改损失函数、替换解耦合头及插入MSFP多尺度特征融合模块提供了便利。相比之下，YOLOv8/v10等后续版本的模块化封装程度更高，在一定程度上限制了此类底层定制化改进。3）在场景适配性方面，YOLOv7本身采用的ELAN-W模块增强了小目标特征的梯度传播能力，其原生的多尺度特征图也与小目标的尺度分布更为匹配，使得其原始框架已具备优于后续版本的微小目标检测鲁棒性。综合来看，尽管YOLOv8/v10等版本在通用场景的精度与速度平衡上有所推进，但其核心优化并非针对小目标任务的定制化改进。因此，YOLOv7是在基础性能、改进灵活性与实时效率三者间取得最佳平衡的理想基准模型，为本研究的系统性增强提供了坚实基础。

综上所述，本文设计的SOD-YOLO创新之处主要在于以下3个方面：1）通过大型语义分割模型SAM2（segment anything model）语义指导下的静态影像小目标样本复制粘贴生成、结合实景三维模型与虚拟人物资产（unreal engine，UE）的动态视频小目标样本透视投影生成，有效扩充小目标训练数量、多样性并提高其标注质量；2）通过视频差分预处理增强小目标的可学习特征、融合不同网络层多尺度特征图增强小目标检测敏感性、引入双层路由注意力（BRA）（Zhu等，2023）捕获多尺度特征间上下文关系以优化小目标特征提取与表达，显著增强低分辨率监控视频下网络模型的小目标特征检测精度和鲁棒性；3）通过解耦合头设计将分类、回归任务分别处理以避免任务间的干扰，并在传统目标框回归损失函数CIoU（complete IoU）中引入可调节参数α以增

加高 IoU 目标的损失和梯度，从而有效克服模型对小目标位置、尺寸细微变化的敏感性以提高目标框回归精度。

1 网络结构

本文网络 SOD-YOLO 整体结构如图 1 所示，主要包括 4 个模块：1）视频背景差分预处理旨在对数据集进行背景差分处理以减少背景对小目标干扰，从而增强网络模型对小目标的识别能力；2）双层路由注意力机制旨在通过双层路由的自注意力计算以有效捕获多尺度特征间的上下文关联，从而优化小目标特征的提取与表达；3）多尺度特征融合旨在通过融合不同网络层多尺度特征图增强小目标检测敏感性；4）多任务解耦旨在通过解耦合头设计分别处理分类任务和回归任务以避免任务冲突对模型性能的影响，从而通过改善多任务学习过程中的权重竞争问题进一步提升小目标检测精度和鲁棒性；5）α-CIoU 定位损失函数旨在通过增强高 IoU 目标的梯度更新，提升小目标边界框的回归精度与训练收敛效率，从而有效改善小目标的定位性能。

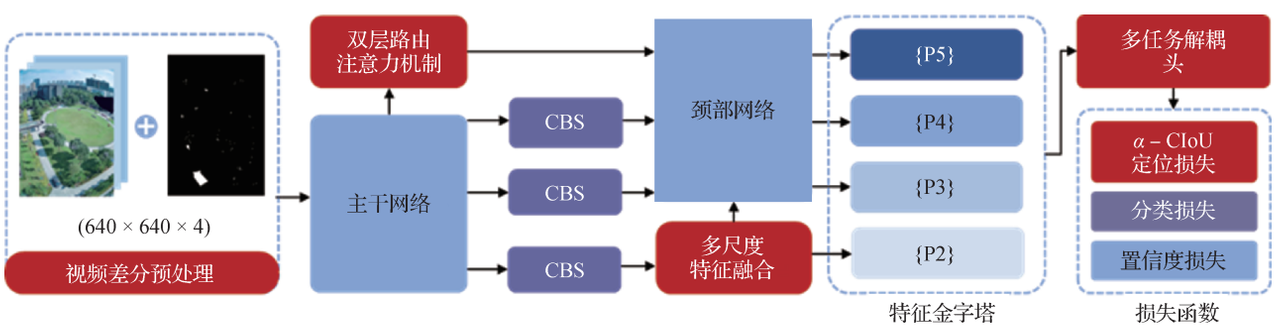


图 1　SOD-YOLO 网络示意

Fig. 1　Diagram of SOD-YOLO network

1.1 小目标特征提取增强

小目标像素占比过低导致深度学习模型难以从 RGB 图像中有效提取小目标的关键特征信息。针对这一问题，本文采取以下 3 种策略以增强网络模型的小目标特征提取能力。

1）视频差分预处理。旨在通过提升监控视频高频分量以增强小目标的可学习特征并减少背景对小目标干扰。如图 2 所示，本文设计了一种背景差分与边缘检测相结合的视频预处理方法。

当原始监控视频 $\boldsymbol{I}_{\mathrm{RGB}}$ 存在稳定背景图像 $\boldsymbol{B}(x,y)$ 时，按式（1）计算可得差分图 $\boldsymbol{D}(x,y)$，该差分图保留前景目标显著区域的同时消除了大部分背景干扰。具体为

$$\boldsymbol{D}(x,y)=\left|\boldsymbol{I}_{\mathrm{RGB}}(x,y)-\boldsymbol{B}(x,y)\right| \tag{1}$$

当原始监控视频不存在稳定背景图像时，则按式（2）进行边缘检测计算以获取目标的边缘特征图 $\boldsymbol{E}(x,y)$。具体为

$$\boldsymbol{E}(x,y)=\sqrt{\left(\frac{\partial \boldsymbol{I}_{\mathrm{RGB}}(x,y)}{\partial x}\right)^2+\left(\frac{\partial \boldsymbol{I}_{\mathrm{RGB}}(x,y)}{\partial y}\right)^2} \tag{2}$$

式中，$\frac{\partial \boldsymbol{I}_{\mathrm{RGB}}(x,y)}{\partial y}$ 和 $\frac{\partial \boldsymbol{I}_{\mathrm{RGB}}(x,y)}{\partial y}$ 分别表示图像在 x 和 y 方向上的梯度分量。进一步地，将 $\boldsymbol{D}(x,y)$ 或 $\boldsymbol{E}(x,y)$ 与原始 RGB 视频图像堆叠，可生成同时保留颜色信息、显著区域和边缘特征信息的四通道输入数据。具体为

$$\boldsymbol{I}_{\mathrm{input}}=Concat(\boldsymbol{I}_{\mathrm{RGB}},\boldsymbol{C}) \tag{3}$$

式中，$\boldsymbol{C}$ 表示 $\boldsymbol{D}(x,y)$ 或 $\boldsymbol{E}(x,y)$。由于 $\boldsymbol{C}$ 代表了视频的高频分量，故 $\boldsymbol{I}\in\mathbf{R}^{H\times W\times 4}$ 为小目标检测提供了更丰富的特征信息。在实际监控场景中，背景并非始终稳定。为此，本文在视频预处理阶段显式引入了背景稳定性判定步骤，并将其量化为运动像素比例 ρ。设第 t 帧灰度图像为 $\boldsymbol{I}_t(x,y)$，相邻帧的像素级绝对差分为

$$\boldsymbol{D}_t(x,y)=\left|\boldsymbol{I}_t(x,y)-\boldsymbol{I}_{t-1}(x,y)\right| \tag{4}$$

通过给定差分阈值 τ_d，定义运动掩膜 $M_t(x,y)$ 为

$$M_t(x,y)=\begin{cases}1 & \boldsymbol{D}_t(x,y)>\tau_d\\ 0 & \text{其他}\end{cases} \tag{5}$$

对同一摄像视角，在连续 T 帧（本文取 $T=30$）上统计运动像素占比 ρ，具体为

$$\rho=\frac{1}{(T-1)HW}\sum_{t=2}^{T}\sum_{x,y}M_t(x,y) \tag{6}$$

式中，H，W 分别为图像高和宽。若 $\rho<\rho_0$，背景稳定

判定阈值，则判定该时间段内场景背景基本保持静止（本文实验中取$\tau_d = 20$，$\rho_0 = 5\%$），可以估计稳定背景$\boldsymbol{B}(x, y)$，采用式（1）进行背景差分。当$\rho \geqslant \rho_0$时，说明场景中存在大范围运动（如人群密集流动、车辆频繁进出等），此时难以可靠建模静态背景，故直接采用式（2）进行边缘检测，以获得边缘特征图$\boldsymbol{E}(x, y)$。

需要指出的是，视频差分方法在处理完全静止或移动极其缓慢的目标时，可能因目标与背景差异不显著而导致漏检。为缓解该问题，本文在实际部署中采用了多帧累积差分与背景建模相结合的策略：当检测到场景中存在长期静止目标时，系统自动切换至边缘增强模式即式（2），并辅以光流分析，从而在保留运动目标的同时，对表观特征明显的静止小目标仍具有一定识别能力。此外，在训练阶段，本文通过合成数据引入不同运动状态的目标样本，以提升模型对静止与运动目标的整体判别鲁棒性。

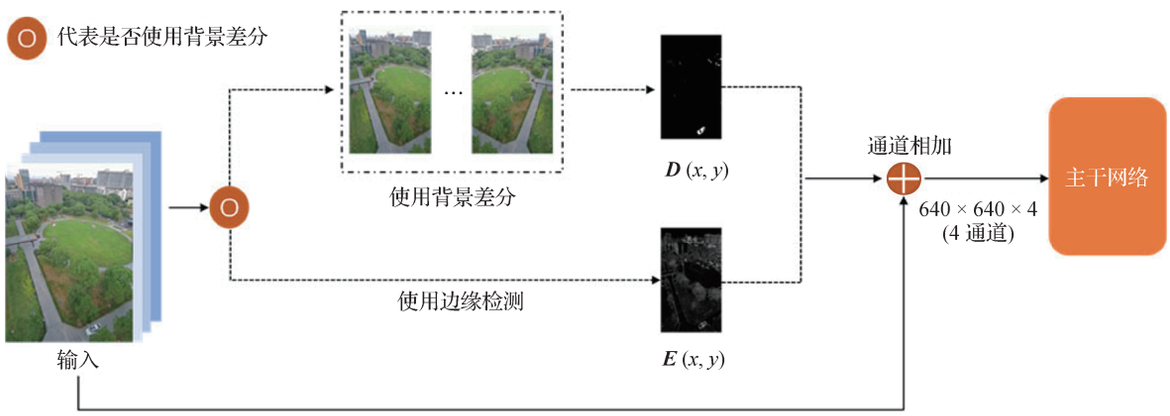


图2　视频背景差分四通道输入示意

Fig. 2　Diagram of four-channel input under video background differential preprocessing

2）多尺度特征融合。旨在融合不同网络层输出的多尺度特征图以增强小目标检测敏感性。容易理解，深度学习网络输出的浅层特征图提供了丰富的局部细节信息，而深层特征图则主要含有较为抽象的语义信息，故网络层数增加会导致小目标特征逐渐消失或淹没于背景信息之中。针对这一问题，本文设计了小目标检测结构多尺度特征金字塔融合（multi-scale feature pyramid fusion，MSFP）以精确表达小目标细节，如图3所示，其基本思想是整合自上而下的深层特征图及自下而上的浅层特征图，并通过与主干网络同尺度特征进行拼接形成可更好保留浅层次特征图中丰富细节信息的多尺度融合特征。该融合特征结合了浅层特征图、深层特征图两者优势且保留更多细节信息，故有助于增强网络对小目标检测的敏感性，进而实现不同尺度目标的精确检测。需要指出的是，尽管深层特征因感受野增大而可能损失部分小目标的细节信息，但其蕴含的高级语义上下文（如目标与场景的关联、遮挡关系等）对区分真实目标与背景噪声至关重要。本文所设计的MSFP结构并非简单地将深层特征与浅层特征拼接，而是通过上采样与通道调整后与同尺度浅层特征进行加权融合，使深层语义信息作为引导，增强浅层特征图中潜在目标的显著性。换言之，深层特征在此起到“语义校准”作用，帮助模型在复杂背景下更准确地判断哪些局部细节应被保留与强化，而非直接用于定位像素级细节。因此，融合深层特征有助于提升小目标检测的整体鲁棒性，尤其是在背景杂乱、目标密集的监控场景中。

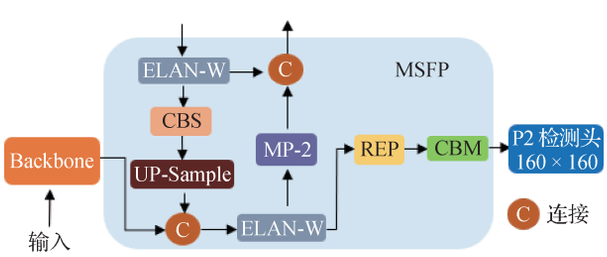


图3　MSFP网络结构

Fig. 3　Diagram of MSFP structure

3）双层路由注意力机制。旨在捕获多尺度特征间上下文关系以优化小目标特征提取与表达，尤其是在背景复杂、小目标密集的场景。本文引入双层

路由注意力机制(BRA)以帮助网络同时理解全局特征和局部特征(贾可心 等,2022),其基本思想是全局特征负责捕捉长距离依赖和背景上下文信息以更好地理解目标与环境的关系;局部特征则专注于局部区域的细粒度特征建模。两者通过加权融合动态平衡全局和局部信息,从而使模型能够兼顾全局背景和局部细节,关键在于如何自适应地为输入特征分配权重来增强模型对特定信息或区域的关注。这里首先将输入特征 $\boldsymbol{X} \in \mathbf{R}^{(W \times H \times C)}$ 划分为 $S \times S$ 个区域,每个区域包含 WH/S^2 个特征向量,即将 $\boldsymbol{X}$ 转变为 $\boldsymbol{X}^r \in \mathbf{R}^{S^2 \times \frac{WH}{S^2} \times C}$;然后,通过线性映射得到 $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 并以之为基础计算区域级特征表示 $\boldsymbol{Q}^r$,$\boldsymbol{K}^r$ 及其邻接矩阵 $\boldsymbol{A}^r$;最后,保留每个区域前 k 个连接以度量区域间的相关性并在局部窗口内计算 Token-to-Token 注意力以学习细粒度特征,该注意力权重可表示为

$$\boldsymbol{A}^r = soft(\boldsymbol{K}_g^{\mathrm{T}} / \sqrt{d}) \tag{7}$$

式中,$\boldsymbol{K}_g^{\mathrm{T}}$ 为条件 $\boldsymbol{O} = \boldsymbol{A}\boldsymbol{V}_g$ 下的新特征表示。BRA 区域级注意力和局部 Token 关注的协同作用确保了特征恢复至原始分辨率后可实现全局上下文信息和局部细节建模的有效平衡,故有助于优化小目标特征提取与表达;此外,限制计算范围(局部窗口)及降采样操作也大大降低了计算复杂度,从而能适用于高性能、高效率要求的高分辨率图像小目标检测任务。

1.2 小目标边界框定位增强

视觉定位任务普遍要求网络模型对物体位置变化具有敏感性,即图像中物体位置发生变化时,模型应能感知并适应这些变化以准确定位物体。需要指出的是,小目标边界框定位通常比大目标更加困难,原因在于小的位移可能对模型损失度量 IoU 产生较大影响(Yu 等,2016);此外,目前用于目标边界框定位的网络耦合头(Hussain,2023)通常设计为多任务共享模式,而同一耦合头进行多任务学习时不同任务信息(如分类信息、回归信息等)可能相互干扰,尤其是当小目标出现在复杂背景或密集场景中时,由于其边界框定位和目标分类之间没有足够区分度时,耦合头将无法有效区分和提取目标的细节特征。针对这一问题,本文采取以下两种策略以增强网络模型的小目标边界框定位能力:

1)多任务解耦。通过解耦合头设计将分类、回归任务分开处理以避免任务间的干扰。如图 4 所示,首先使用 1 × 1 卷积调整输入特征图的通道数以匹配后续任务的需求,随后引入两个并行的 3 × 3 卷积层分别用于分类、回归任务特征提取,最后利用 1 × 1 卷积分别完成目标分类、边界框回归及置信度预测,其中,每个特征点的目标类别置信度被单独评估,同时通过回归系数的学习来描述该特征点对应目标的边界框坐标;分类、回归和置信度 3 部分结果综合处理得到最终检测结果。后续实验结果表明,引入解耦合头结构不仅通过任务独立优化降低分类任务对回归任务的干扰、显著提升检测精度,还通过优化计算路径和特征处理提高了网络收敛速度,使得模型在处理复杂场景时更加高效、更具鲁棒性。

2)α-CIoU 损失函数设计。以 YOLO 为代表的目标检测网络广泛采用损失函数 CIoU(Zheng 等,2020)优化目标边界框的回归过程。尽管 CIoU 综合

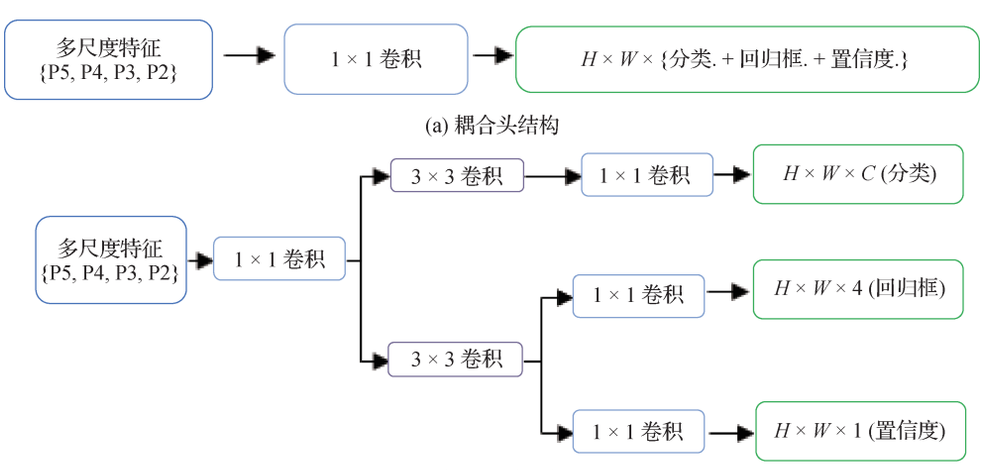


图 4 SOD-YOLO 解耦合头结构

Fig. 4 Diagram of SOD-YOLO decoupled head ((a) coupled head structure; (b) decoupled head structure)

考虑了边界框重叠度、中心点距离和长宽比等因素，且在一般目标检测中表现良好，但其损失计算依赖于对小目标边界框定位精度不够敏感的IoU，尤其是小目标附近的长宽比、中心点距离误差容易被忽视。针对这一问题，本文借鉴α-IoU（He等，2021）思想设计损失函数α-CIoU以优化小目标边界框回归过程。α-IoU是在传统IoU基础上引入一个可调节参数α（通常为幂指数）以获得一个更为灵活的损失函数，具体形式为

$$CIoU = IoU - \frac{d^2}{c^2} - av \tag{8}$$

$$a = \frac{v}{(1 - IoU) + v} \tag{9}$$

$$v = \frac{4}{\pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h}\right)^2 \tag{10}$$

$$L_{\alpha\text{-IoU}} = 1 - \left(\frac{A \cap B}{A \cup B}\right)^{\alpha} \tag{11}$$

式中，A和B分别表示预测框和真实框的面积，α是一个控制损失函数敏感度的超参数，故网络模型损失计算通过调节α的值可自适应于不同大小的目标。

由式（11）看出，当$IoU > 0.5$时，α-IoU的梯度大于1，故相对于原始IoU损失，α-IoU增加了高IoU目标的损失和梯度，从而有利于提高目标边界框的回归精度。受此启发，将由式（8）定义的损失函数CIoU改写为

$$L_{\alpha\text{-IoU}} = 1 - IoU^{\alpha} - \left(\frac{d^2}{c^2}\right) - (av)^{\alpha} \tag{12}$$

由式（12）看出，当CIoU > 0.5时，α-CIoU的梯度大于1，这意味着，α-CIoU相较于原始CIoU损失能施加更大的梯度更新，从而增强对高IoU目标的优化力度。相对于大目标，小目标的边界框通常难以精确对齐且回归误差的微小变化都会显著影响最终检测结果，而α-CIoU通过增加高IoU目标的损失和梯度可进一步缩小预测框与真实框之间的偏差，确保即便是小目标也能得到更准确的定位，故α-CIoU特性对于小目标检测尤为关键；此外，增强高IoU目标的梯度还可加速模型训练过程的收敛，从而提升整体检测性能。

2 样本生成

监控视频分辨率较低导致其小目标像素占比很小且存在边界模糊，这使得小目标样本视频数据集人工制作效率低、标注质量差。针对这一问题，本文提出以下两种方法自动获取小目标样本和对应的标注框。

1）结合复制粘贴技术与SAM2掩膜约束的静态影像小目标样本生成。复制粘贴（copy-and-paste）技术已广泛用于增强数据分布多样性，其核心思想是从已有标注数据中提取小目标区域图像片段并将其粘贴到随机选择的背景图像，该技术虽然有助于网络模型更好地学习小目标特征分布且简单高效，但粘贴过程的随机性可能导致输出结果违背目标与背景间的物理约束，例如，道路场景中小目标（如行人或车辆）通常应出现在地面上而非悬浮在空中。本文采用复制粘贴技术实现监控视频小目标样本增广目的，并针对其不足引入大型语义分割模型SAM2（Kirillov等，2023）动态生成掩膜以指导复制粘贴过程，其优点在于两方面：一是SAM2能够精准区分交通场景中的道路、天空和建筑等语义类别，其生成的掩膜边界清晰、能准确反映场景语义结构，可确保目标粘贴位置符合物理意义，例如，基于SAM2生成的道路区域掩膜可以有效限制车辆目标仅粘贴在道路区域中以避免传统随机粘贴可能导致的车辆悬浮空中或与建筑重叠等不合理现象；二是SAM2生成的掩膜具有较高的场景适应性，能够广泛应用于多种复杂场景，并且确保粘贴目标的语义一致性，从而为目标生成提供强有力的指导。图5给出了SAM2掩膜引导下的静态影像小目标样本复制粘贴生成流程，有效增强了同一目标场景的样本多样性。

2）结合实景三维模型与虚拟人物资产（Unreal Engine，UE）的动态视频小目标样本生成。大视场监控视频中样本数据筛选困难且传统人工标注质量低，针对这一问题，本文将UE虚拟人物资产（Sanders，2016）添加到监控场景对应的三维模型并渲染生成包含小目标样本的视频数据，从而通过使用虚拟人物资产获取大量训练样本。以某大校园大视场监控场景为例，其样本生成过程为：首先，采用无人机倾斜摄影建模技术构建一个精确且真实的三维场景；然后，将其导入虚幻引擎（UE4）并利用UE4自带的人物资产来模拟校园中的行人活动，即通过添加不同性别、年龄和服装风格的虚拟人物使场景更加贴近真实样本；最后，以无人机视角模拟生成的校园场景的监控视频并使用UE4自动函数获取人物资产

在视频中的检测框坐标、导出带有资产位置信息文件，从而直接获取人物资产的标注文件。图 6 给出了 UE 合成视频在不同时刻的小目标样本数据及其标注，相同背景下显著提高了小目标样本的多样性。

由上述可看出，本文小目标样本生成方法或通过掩膜设计引入合乎物理逻辑的更强语义约束成功地克服了传统随机粘贴技术在样本数据增广过程中的不足，或结合场景三维建模、UE 中虚拟人物资产、虚拟行人行为模拟及三维图像渲染技术实现视频小样本数据模拟生成并自动获得高质量的标注框，从而有效解决了监控视频小目标样本稀少、标注质量低等问题，为后续模型训练提供可靠的数据支持。需要强调的是，上述两类样本生成方式仅作为对第 3 节所述 3 个真实监控场景数据集的补充，而非替代；第 3. 2. 3 节的消融实验也进一步表明，在“真实数据为主、适量合成数据辅助”的配置下，既可以显著提升小目标检测精度，又能避免过多合成数据导致的泛化性能下降。

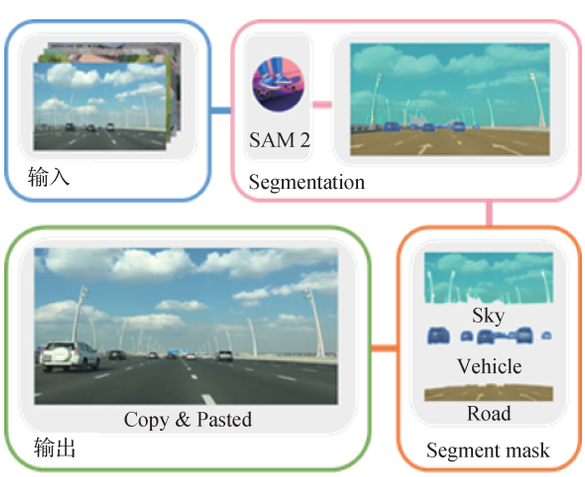


图 5　SAM2 掩膜下的静态影像小目标样本复制粘贴生成流程

Fig. 5　Generation pipeline of small object samples in static images through copy-and-paste technique guided by semantic segmentation model SAM2

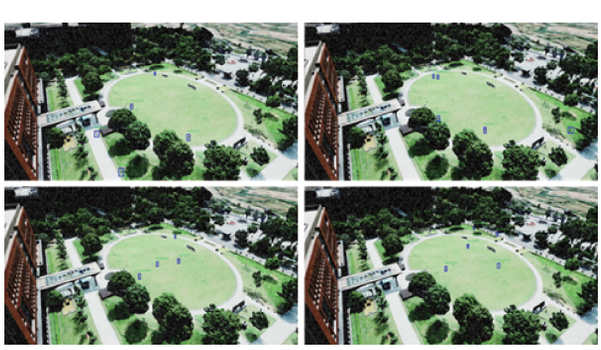


图 6　基于 UE 虚拟人物资产的视频小目标样本合成示意

Fig. 6　Diagram of generating small target samples for dynamic videos using UE virtual character assets

3　实验与分析

由于目前缺乏公开的监控视频小目标检测样本数据集，本文网络 SOD-YOLO 性能测试选用涉及小目标检测应用的建设工地、高速公路服务区及大学校园等 3 个真实场景并构建相应监控视频样本数据集用于训练测试目的，该场景监控视频由多个不同视角摄像机拍摄，相机距离地面垂直高度约 20～30 m，视频分辨率为 300 万像素。由于相机距离地面较高、视场大，待检测目标尺寸较小且不同类别目标存在较大的尺度差异。3 个场景数据具体说明如下：

1）建筑工地。约 10 000 幅图像构成，图像尺寸为 1 536 × 2 048 像素，其中，小目标样本（尺寸小于 32 × 32 像素）占比为 56. 3%，小目标类别为行人、电动自行车和小轿车。

2）高速公路服务区。约 25 000 幅图像构成，图像尺寸为 1 536 × 2 048 像素，其中，小目标样本（尺寸小于 32 × 32 像素）占比为 64. 7%，小目标类别为行人和小轿车，其他类别为货车、油罐车和公交车。

3）大学校园。约 20 000 幅图像构成，其中，尺寸为 1 536 × 2 048 像素的为宽视场图像约 15 000 幅，尺寸为 1 920 × 1 280 像素的窄视场图像 5 000 幅；小目标样本约 70 000 个，占样本总量的 83. 5%，类别全部为行人。

本文所有实验均将输入图像统一缩放至 640 × 640 像素，以适配 SOD-YOLO 的网络结构。具体而言，对原始分辨率为 1 536 × 2 048 像素或 1 920 × 1 280 像素的图像，采用双线性插值进行等比例缩放，并在训练过程中引入随机裁剪与色彩增强等数据增广手段，以保持尺度不变性与模型泛化能力。在测试阶段，图像同样缩放至 640 × 640 像素后进行前向推理，最终输出结果再映射回原始图像坐标系以进行评估。

为进一步验证 SOD-YOLO 在公开小目标检测基准数据集上的泛化性能，本文选取可见光—热红外微小目标基准数据集 VT-TOD 中的可见光通道开展对比实验。VT-TOD 中目标尺度普遍较小，且分布于复杂的自然场景与城市场景背景之上，具有典型的小目标检测难度。实验中仅使用可见光图像部

分数据集及其标注信息。

本文网络使用4块NVIDIA GeForce 3080 Ti GPU训练，batchsize大小为8，初始学习率为0.01，epochs为250个。网络性能评估采用COCO数据集中标准指标AP、特定IoU阈值下的AP（AP50、AP75）以及不同物体大小的总体AP：APs（小）、APm（中）和APl（大）。

3.1 网络性能评估

本文选取多种主流目标检测深度学习网络进行对比分析，表1—表3分别给出了3个场景数据集的对比实验结果，其中：YOLOv7（Wang等，2023）、YOLOv8和YOLOv10（Wang等，2024）属于单阶段检测网络，以检测速度快、实时性强为特点；RT-DETR是基于Transformer的端到端实时检测器；Faster R-CNN是典型的双阶段检测网络，以高精度和复杂场景适应性见长；YOLC（you only look clusters）（Liu等，2024）和Drone-YOLO（Zhang等，2023）为专门针对小目标检测网络。

表1 建筑工地场景对比实验结果

Table 1 Comparative experimental results of construction site scenes

/%

模型	参数/M	AP	AP50	AP75	AP50s
YOLOv7	28.1	22.8	52.6	12.1	46.3
YOLOv8	28.6	21.3	51.1	12.1	48.8
YOLOv10	21.6	20.3	50.5	12.7	44.1
RT-DETR	45.9	25.3	**57.5**	12.0	44.6
Faster R-CNN	-	21.5	54.3	13.1	48.5
YOLC	43.3	22.9	55.2	12.6	45.6
Drone-YOLO	40.5	**25.4**	56.9	11.9	52.7
SOD-YOLO（本文）	36.4	24.5	56.2	**13.5**	**53.9**

注：加粗字体表示各列最优结果。“-”为不适用或未测量。

由表1—表3可以看出，建筑工地场景下本文SOD-YOLO在AP75和AP50s上表现突出，尤其是AP75这一高IoU阈值下的检测精度优于所有其他模型，这表明SOD-YOLO采用的α-CIoU定位损失和解耦合头提高了小目标边界框的回归精度；SOD-YOLO的AP和AP50精度分别为24.5%和56.2%，仅次于RT-DETR和Drone-YOLO，但SOD-YOLO训练得到的模型参数量却较RT-DETR降低20.7%，这表明SOD-YOLO在保持高效训练的同时能达到与其他网络相当甚至更好的性能。高速公路服务区场景包含多种待检测类别、样本数量多且目标间包含较严重的遮挡，本文SOD-YOLO在AP精度上超过YOLC和RT-DETR，同时在AP75指标上精度依旧最高，这表明SOD-YOLO在复杂类别及遮挡条件下的小目标高精度检测任务具备较强的鲁棒性。大学校园场景中小目标像素占比仅为0.007 5%，特征提取难度最大，但SOD-YOLO在AP、AP75、AP50s 3个指标仍取得最高精度，相较于基础网络YOLOv7分别提高了4.9%、2.5%、5%，除AP50指标外模型性能也均优于其他目标检测网络，这表明SOD-YOLO采

表2 高速公路服务区场景对比实验结果

Table 2 Comparative experimental results of highway service area scenes

/%

模型	参数/M	AP	AP50	AP75	AP50s
YOLOv7	73.1	38.5	80.2	28.4	40.3
YOLOv8	75.6	38.9	81.3	28.0	42.1
YOLOv10	70.5	37.7	79.6	27.5	41.6
RT-DETR	102.1	40.1	82.5	28.8	42.1
Faster R-CNN	-	39.5	80.4	27.6	41.8
YOLC	89.3	42.4	**83.3**	28.1	43.6
Drone-YOLO	80.9	41.6	82.5	27.5	42.5
SOD-YOLO（本文）	92.7	**42.6**	82.4	**29.5**	**44.3**

注：加粗字体表示各列最优结果。“-”为不适用或未测量。

表3 大学校园场景对比实验结果

Table 3 Comparative experimental results of college campus scenes

/%

模型	参数/M	AP	AP50	AP75	AP50s
YOLOv7	86.3	33.7	70.6	22.9	36.1
YOLOv8	78.5	32.4	72.5	22.5	35.2
YOLOv10	74.2	32.0	69.3	24.4	33.5
RT-DETR	104.6	37.2	**72.8**	22.7	39.9
Faster R-CNN	-	35.2	72.7	24.1	32.6
YOLC	91.8	38.3	71.5	25.2	40.4
Drone-YOLO	84.1	36.6	71.4	24.5	**41.1**
SOD-YOLO（本文）	95.4	**38.6**	71.5	**25.4**	**41.1**

注：加粗字体表示各列最优结果。“-”为不适用或未测量。

用的小目标特征提取增强策略可有效增强低分辨率监控视频下网络模型的小目标特征检测精度和鲁棒性。对比YOLC和Drone-YOLO专门设计的两种小目标检测网络，本文SOD-YOLO在小目标平均精度APs最高，显示出在小目标检测和高精度检测任务中的显著优势。

图7展示了本文网络SOD-YOLO与Faster R-CNN、YOLC、Drone-YOLO及基础网络YOLOv7在某时刻视频上的小目标检测结果对比示意。从图7可发现，YOLOv7在检测小目标时存在严重的漏检和误检问题，其主要原因在于其特征提取模块对远距离小目标的信息表达能力有限，导致高层特征中小目标被弱化而难以有效识别；基于Region Proposal机制且多用于通用目标检测的Faster R-CNN虽然具备较强的检测能力，但在面对分辨率较低或尺寸较小的目标时，因候选区域生成易受干扰而导致部分背景区域被误判为目标（图中建筑物影子、电瓶车、雪糕筒等）；相比之下，Drone-YOLO与YOLC作为专门面向小目标检测任务设计的网络，因在特征提取及目标感知能力上进行了优化，能够较好地保留小目标的空间信息并提升检测精度，但二者在复杂场景下仍存在对多尺度目标识别能力不足的问题（图

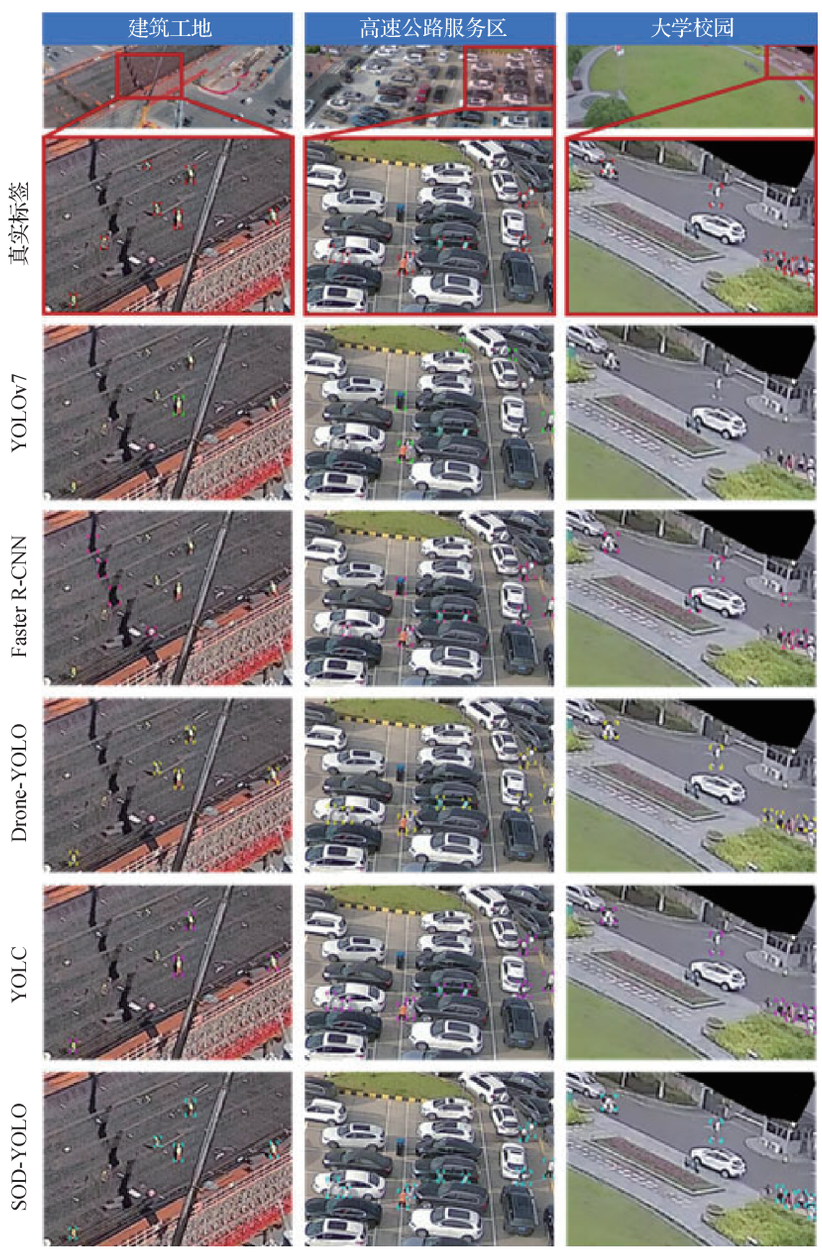


图7　不同网络模型推理结果示意

Fig. 7　The inference results of different network models

中垃圾桶等)。与上述网络相比,本文SOD-YOLO针对小目标特征提取增强、小目标边界框定位增强所采取的策略有效减少了漏检问题(如图7中人员聚集区域)和误检问题(如图7中建筑物影子、垃圾桶、电瓶车、雪糕筒),各场景中除指标AP50外均最优,指标AP50也处于次优或接近次优结果,表明本文网络小目标检测具有良好的准确性和鲁棒性。

从表1—表3也可以看出,在部分指标上其他方法仍具有一定优势。例如,在建筑工地和高速公路服务区场景中,RT-DETR在总体AP指标上略高于SOD-YOLO,这与其基于Transformer的全局特征建模和采用端到端集合预测机制有关,更有利于中等尺度目标和结构复杂背景的检测;在AP50指标上,YOLC、Drone-YOLO等面向小目标的网络在个别场景与SOD-YOLO相当甚至略优,其锚框设计和损失加权对"是否检出"这一粗粒度指标更为敏感。相较之下,SOD-YOLO更侧重于在高IoU阈值和APs指标下优化远距离小目标的精确定位,因此在AP75、AP50s等更关注定位精度的指标上优势更为明显。上述结果表明,本文方法在远距离小目标精细检测方面具有明显优势,但在中大目标和通用场景检测上仍有进一步提升空间,后续工作可在保持当前框架的基础上考虑引入更强的全局建模模块或多任务蒸馏策略,以进一步提高模型的综合性能。

3.2 公开数据集对比实验

为进一步验证所提SOD-YOLO在公开小目标检测基准数据集上的泛化能力,本文在可见光—热红外微小目标检测数据集VT-TOD上开展了对比实验。VT-TOD数据集由国防科技大学等单位构建,面向可见光与热红外双模态场景,对复杂自然环境和城市环境中的大量微小目标进行了精细标注,具有目标尺度小、背景复杂以及样本数量大等特点,是当前具有代表性的小目标检测基准之一。同时提供配准的可见光与热红外图像,本文仅选取其中的可见光图像及对应标注进行实验,保持与原始数据集相同的训练/测试划分。实验中,SOD-YOLO在VT-TOD可见光训练集上从头训练,输入分辨率统一为640×640像素,优化器、初始学习率、batchsize等超参数与前文在自建监控数据集上的设置基本一致,仅根据VT-TOD图像数量适当调整了训练轮次以避免过拟合。为进行公平对比,本文选取了3种具有代表性的小目标检测方法作为对比模型:基于查询机制的QueryDet、面向无人机航拍小目标的YOLOC,以及针对无人机监控场景设计的Drone-YOLO,所有模型均在相同的训练/测试划分下训练与评估。

表4给出了各方法在VT-TOD可见光测试集上的检测性能对比结果。可以看出,SOD-YOLO在所有指标上均取得了最优性能,其中精确率达到61.4%,AP50:95为52.3%,APs为63.5%,APm为58.2%。与当前表现较好的Drone-YOLO相比,SOD-YOLO的精确率提升了2.5%,AP50:95提升4.2%,APs提升2.7%,APm提升6.2%。相较QueryDet和YOLOC,SOD-YOLO都有较大幅度提升,进一步说明本文方法在公开小目标基准数据集上的检测精度优势。

表4 VT-TOD公共数据集上的小目标检测性能对比

Table 4 Performance comparison of small object detection on the VT-TOD public dataset

/%

模型	参数量/M	精确率	AP50:95	APs	APm
QueryDet	38.6	43.6	33.8	45.2	39.8
YOLOC	68.9	56.8	42.4	59.2	54.6
Drone-YOLO	78.8	58.9	48.1	60.8	52.0
SOD-YOLO	74.4	**61.4**	**52.3**	**63.5**	**58.2**

注:加粗字体表示各列最优结果。

从图8可以看到,QueryDet在远距离船只等微小目标区域存在较多漏检;YOLOC和Drone-YOLO虽然能够检出更多目标,但仍然出现部分目标边界不准或多个相邻目标被合并为一个框的情况。相比之下,SOD-YOLO在放大区域内能够给出数量更完整、边界更贴合的检测结果。

3.3 消融实验

3.3.1 功能模块

为验证视频差分预处理、双层路由注意力机制、多尺度特征融合、α-CIoU损失函数及多任务解耦合头设计在SOD-YOLO小目标特征提取、边界框定位增强方面的作用,本文将上述模块不同组合引入YOLOv7基础模型并利用高速公路服务区、大学校园两个场景数据进行消融实验,实验结果如表5所示,其中首行为YOLOv7基础模型的训练精度。

由表5可以看出,基础网络引入视频差分预处理后性能提升最为显著,原因在于,从RGB三通道



图 8　VT-TOD 数据集上的小目标检测可视化结果

Fig. 8　Visualization results of small object detection on the VT-TOD dataset

到包含差分信息的四通道训练使得网络通过捕获目标与背景之间的显著差异能更加精准地提取前景目标的特征，同时降低复杂背景对检测任务的干扰；分别引入双层路由注意力机制、多尺度特征融合、α-CIoU 损失函数后，网络各项评价指标均实现了小幅提升，其中视频差分预处理通过去除无关背景信息可使前景目标更加突出，而双层路由注意力机制则能更精准地聚焦目标区域，故两者协同能进一步提升小目标检测能力。

需要指出的是，单独引入多任务解耦合头虽然使 AP 和 AP50 指标出现了小幅下降，但结合 α-CIoU 和多尺度特征融合却显著提升目标精测精度，其原因在于，解耦合头设计改变了分类、回归任务的特征学习方式，因分开学习的特征在训练初期未能充分融合而导致检测精度下降，但多尺度特征融合对小目标检测敏感性的增强及 α-CIoU 损失计算对小目标特征的关注共同弥补了多任务解耦合头设计带来的不足，故能显著提升目标边界框回归精度。总体上，基础网络结合上述全部模块后，其各项指标均得到了显著提升，其中小目标平均精度 APs 表现最突出，证明了 SOD-YOLO 小目标检测任务中各模块设计的有效性。

表 5　消融实验

Table 5　Ablation experiment

/%

视频背景差分	双层路由注意力机制	多尺度特征融合	α-CIoU 损失函数	多任务解耦	AP	AP50	AP75	APs
–	–	–	–	–	33.7	69.6	21.9	36.1
√	–	–	–	–	37.4	70.3	24.4	40.8
–	√	–	–	–	36.8	**71.3**	22.6	40.2
–	–	√	–	–	35.1	71.1	23.2	38.9
–	–	–	√	–	34.5	70.8	24.7	37.8
–	–	–	–	√	32.9	70.4	22.7	36.4
√	√	–	–	–	37.9	70.5	24.6	40.6
–	–	√	√	√	36.6	71.0	23.8	40.0
√	√	√	√	√	**38.6**	**71.3**	**25.4**	**41.1**

注：加粗字体表示各列最优结果。“√”表示加入该策略，“–”为未加入该策略。

图 9 给出了基础网络在不同功能模块下推理生成的大学校园场景行人目标检测热力图，用于反映网络对图像中各区域的关注强度，热力图数值越高（红）表示网络对该区域越敏感、关注度越高，其中，图 9(a) 为原始输入图像，背景中包含大量干扰因素（如树木、道路、栏杆等复杂结构）且目标行人尺寸较小；图 9(b) 为基础网络 YOLOv7 的热力图，可以看出该网络对目标区域的关注较为分散，在背景区域，如树叶、地面等处出现了大量高响应区域，表明其易受到背景干扰，难以准确聚焦待检测的小目标；图 9(c) 为引入视频预差分与双层路由注意力机制模块后的热力图，从中可发现网络对无关区域（如树木、道路等）的响应明显减弱，原因在于背景差分增强了时序变化区域的显著性，而注意力机制有效引导了网络

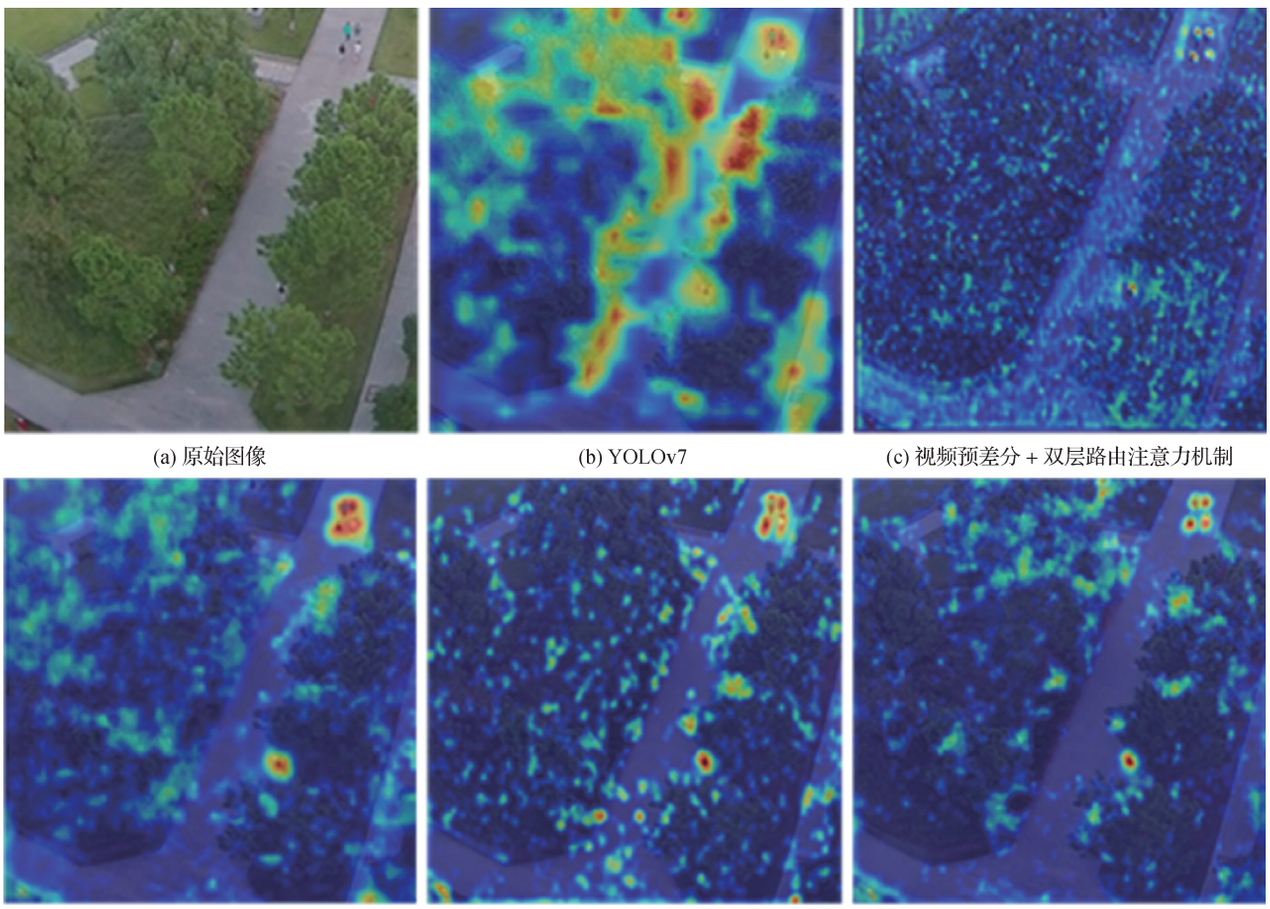


图9　不同模块组合下大学校园场景行人目标检测热力图

Fig. 9　Heat map of object detection in college campus scene under different module combinations ((a) original image; (b) YOLOv7 heat map; (c) heat map of video pre-differential and routing attention mechanism; (d) heat map of multi-scale feature fusion module; (e) heat map of multi-scale feature fusion, α-CIoU loss, and decoupled head; (f) heat map of all the modules)

聚焦于具有目标特征的区域;图9(d)为引入多尺度特征融合模块后的热力图,可以发现网络在感知目标区域的同时,对道路等大面积背景区域的关注显著减少,对小目标样本的关注度明显提高;图9(e)为引入多尺度特征融合、α-CIoU损失函数及多任务解耦模块后的热力图,相比图9(d)对目标区域的响应强度明显增强,目标轮廓更为清晰;图9(f)为引入所有功能模块后的最终热力图,该热力图中响应最强的区域集中于目标上且目标轮廓明确、分布紧凑,表明网络在整体感知、特征表达与检测精度方面均得到了显著提升,验证了各模块在小目标检测任务中的协同增益作用。

3.3.2　样本生成

为验证UE合成数据对模型训练性能的影响,本文将大学校园场景的真实数据及UE合成数据按不同比例混合进行网络训练以评估其在不同数据组合下的性能表现,实验结果如表6和表7所示。其中,表6的实验为以大学校园场景的6 000幅真实数据为基础逐步添加不同比例UE合成数据进行训练,并以500幅真实数据为验证集;表7的实验为在保持大学校园场景图像训练数据量8 000幅不变的前提下,逐步引入合成数据并同步减少真实数据,直至真实数据完全被合成数据替代,从而探究合成数据在小目标检测任务中的可替代性及其对模型性能的影响。

从表6和表7的实验结果可以看出,适量引入UE合成图像可有效缓解真实数据不足的问题,提升模型在小目标检测任务中的精度表现。这表明合成数据具备较高的训练价值,尤其在真实数据获取成本较高或难以覆盖场景变化的条件下,能够为模型提供有力的数据支撑。然而,实验同时也发现,合成数据的数量并非越多越好。在某些比例配置下,随

表 6　添加 UE 合成数据模型的性能变化实验

Table 6　Experimental results on model performance with addition of UE synthetic data

模型	真实数据	UE 数据	AP/%	AP50/%
YOLOv7	6 000	–	32.7	76.6
		1 500	**39.3**	**80.8**
		3 000	37.1	78.7
YOLOv8	6 000	–	33.5	77.9
		1 500	38.2	**81.5**
		3 000	**39.2**	80.4
RT-DETR	6 000	–	36.2	78.6
		1 500	39.4	**81.9**
		3 000	**40.3**	81.5
SOD-YOLO	6 000	–	37.5	79.8
		1 500	41.0	**82.7**
		3 000	**42.1**	82.3

注：加粗字体表示各模型各精度指标的最优结果。“–”为不适用或未测量。

表 7　保持训练数据总量不变减少真实数据比例实验

Table 7　Experimental results on model performance under a constant training data volume with gradually reduced real data proportion

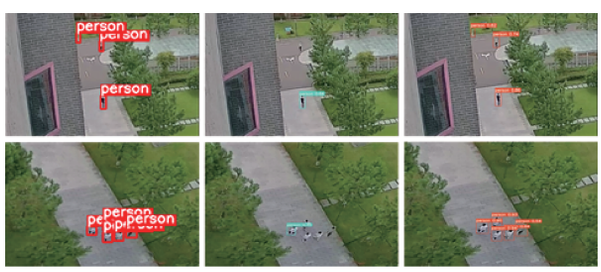
模型	真实数据	UE 数据	AP/%	AP50/%
YOLOv7	6 000	2 000	**36.8**	**72.3**
	4 000	4 000	34.4	66.9
	–	8 000	18.5	38.8
YOLOv8	6 000	2 000	**39.2**	**75.5**
	4 000	4 000	32.5	69.6
	–	8 000	20.8	42.8
RT-DETR	6 000	2 000	39.5	**78.2**
	4 000	4 000	**40.8**	70.3
	–	8 000	28.3	50.6
SOD-YOLO	6 000	2 000	39.9	**78.7**
	4 000	4 000	**41.0**	70.8
	–	8 000	29.0	51.2

注：加粗字体表示各模型各精度指标的最优结果。“–”为不适用或未测量。

着合成数据占比的持续上升，模型性能反而出现下降趋势。推测其原因在于合成数据的真实性：当合成数据与真实数据相似度较高时，添加合成数据对模型训练贡献明显；否则，合成数据过多可能导致的模型过拟合将影响模型在真实数据上的泛化能力。

图 10—图 12 展示了 3 个场景添加 UE 合成数据后的 YOLOv7 推理可视化结果。

从图 10 可以看出，YOLOv7 在未引入合成数据



(a) 真实标签　(b) 未添加合成数据　(c) 添加合成数据

图 10　大学校园场景添加 UE 合成数据前后的网络推理结果对比可视化示意图

Fig. 10　Visualization of network inference results before and after adding UE synthetic data in the college campus scenario ((a) ground truth; (b) inference without synthetic data; (c) inference with synthetic data)



(a) 真实标签　(b) 未添加合成数据　(c) 添加合成数据

图 11　高速公路服务区场景添加 UE 合成数据前后的网络推理结果对比可视化示意图

Fig. 11　Visualization of network inference results before and after adding UE synthetic data in the highway service area scenario ((a) ground truth; (b) inference without synthetic data; (c) inference with synthetic data)

时对远处行人小目标的检测能力较弱，易出现漏检和误检，但加入UE合成数据训练后有效提高了行人小目标识别敏感性，主要原因为合成数据增强了网络对复杂背景中小目标分布的学习能力，同时补充了现实校园中难以获取的多样视角和边缘样本，使得模型在相似背景条件下的识别能力得以提升。

图11对应的高速公路服务区场景中，YOLOv7因车辆、行人小目标常处于部分遮挡状态易出现漏检，但加入UE合成数据训练后即使在遮挡较重区域也能保持良好的检测效果（如图中聚集的人群）。

图12对应的建筑工地场景存在大量非结构化背景（如脚手架、设备、材料堆积等），导致YOLOv7产生较多漏检，但引入合成数据进行训练后，尤其是合成数据中包含更多边界模糊、光照变化和遮挡干扰下的样本，使得网络在面对建筑工地这一具有强干扰性的场景时表现出更好的鲁棒性。



(a) 真实标签　(b) 未添加合成数据　(c) 添加合成数据

图12　建筑工地场景添加UE合成数据前后的网络推理结果对比可视化示意图

Fig. 12　Visualization of network inference results before and after adding UE synthetic data in the construction site scenario ((a) ground truth; (b) inference without synthetic data; (c) inference with synthetic data)

4　结　论

小目标检测技术可为无人驾驶、安防监控以及智慧医疗等场景应用提供更加精准、可靠的技术支持。大视场监控摄像机覆盖区域广泛，但由于小目标可用特征少、定位精度低，包括深度神经网络在内的、依赖于显著性特征（如纹理、形状）信息的目标检测方法在小目标检测任务中面临诸多挑战；此外，训练数据集稀缺、样本分布不均及标注精度差等困难也是小目标检测网络训练亟待解决的问题。针对上述问题，本文构建大视场监控视频小目标检测网络SOD-YOLO，通过引入视频差分预处理、双层路由注意力机制、多尺度特征融合，增强低分辨率监控视频下的小目标特征检测精度，同时有效克服模型对小目标位置、尺寸细微变化的敏感性以提高目标框回归精度；提出结合复制粘贴技术与SAM2掩膜约束的静态影像小目标样本生成策略、结合实景三维模型与UE虚拟人物资产的动态视频小目标样本生成策略，克服了视频监控数据收集难度大、标注质量差等问题，显著提高了小目标数据集的制作效率。针对涉及大视场监控视频小目标检测应用的3个典型场景数据集构建及网络训练，测试结果证明了本文方法的有效性，对于提升监控系统智能化水平并拓展其应用场景具有重要意义。后续将结合更多场景数据对网络进行性能测试并探索在低功耗嵌入式系统上的应用部署与推理实现，为研制高性价比的视频检测设备终端奠定基础。

参考文献（References）

Bai Y C, Zhang Y Q, Ding M L and Ghanem B. 2018. SOD-MTGAN: small object detection via multi-task generative adversarial network//Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer: 210-226 [DOI: 10.1007/978-3-030-01261-8_13]

Bryan B, Gong Y, Zhang Y Z and Poellabauer C. 2019. Second-order non-local attention networks for person re-identification//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea(South): IEEE: 3759-3768 [DOI: 10.1109/ICCV.2019.00386]

Cai Z W and Vasconcelos N. 2018. Cascade R-CNN: delving into high quality object detection//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, USA: IEEE: 6154-6162 [DOI: 10.1109/CVPR.2018.00644]

Cao M, Ikehata S and Aizawa K. 2025. Field-of-view IoU for object detection in 360° images. IEEE Transactions on Image Processing, 34: 6813-6822 [DOI: 10.1109/TIP.2023.3296013]

Chen T X, Ye Z, Tan Z T, Gong T, Wu Y, Chu Q, et al. 2024. MiM-ISTD: Mamba-in-Mamba for efficient infrared small-target detection. IEEE Transactions on Geoscience and Remote Sensing, 62: #5007613 [DOI: 10.1109/TGRS.2024.3485721]

Deng C F, Wang M M, Liu L, Liu Y and Jiang Y L. 2022. Extended feature pyramid network for small object detection. IEEE Transactions on Multimedia, 24: 1968-1979 [DOI: 10.1109/TMM. 2021. 3074273]

Diwan T, Anirudh G and Tembhurne J V. 2023. Object detection using YOLO: challenges, architectural successors, datasets and applications. Multimedia Tools and Applications, 82(6): 9243-9275 [DOI: 10.1007/s11042-022-13644-y]

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, et al. 2021. An image is worth 16 × 16 words: transformers for image recognition at scale//Proceedings of 2021 International Conference on Learning Representations. Vienna, Austria: OpenReview.net

Hao X Y, Luo S J, Chen M Y, He C H, Wang T and Wu H. 2024. Infrared small target detection with super-resolution and YOLO. Optics and Laser Technology, 177: #111221 [DOI: 10.1016/j.optlastec. 2024.111221]

He J B, Erfani S M, Ma X J, Bailey J, Chi Y and Hua X S. 2021. α-IoU: a family of power intersection over union losses for bounding box regression. Advances in Neural Information Processing Systems, 34: 20230-20242

Hussain M. 2023. YOLO-v1 to YOLO-v8, the rise of YOLO and its complementary nature toward digital manufacturing and industrial defect detection. Machines, 11(7): #677 [DOI: 10.3390/machines11070677]

Jia K X, Ma Z H, Zhu R and Li Y G. 2022. Attention-mechanism-based light single shot multiBox detector modelling improvement for small object detection on the sea surface. Journal of Image and Graphics, 27(4): 1161-1175 (贾可心, 马正华, 朱蓉, 李永刚. 2022. 注意力机制改进轻量SSD模型的海面小目标检测. 中国图象图形学报, 27(4): 1161-1175) [DOI: 10.11834/jig.200517]

Jiang L J, Yuan B X, Du J W, Chen B Y, Xie H F, Tian J, et al. 2024. MFFSODNet: multiscale feature fusion small object detection network for UAV aerial images. IEEE Transactions on Instrumentation and Measurement, 73: #5015214 [DOI: 10.1109/TIM. 2024. 3381272]

Kirillov A, Mintun E, Ravi N, Mao H Z, Rolland C, Gustafson L, et al. 2023. Segment anything//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE: 3992-4003 [DOI: 10.1109/ICCV51070.2023.00371]

Kulambayev B, Nurlybek M, Astaubayeva G, Tleuberdiyeva G, Zholdasbayev S and Tolep A. 2023. Real-time road surface damage detection framework based on mask R-CNN model. International Journal of Advanced Computer Science and Applications, 14(9): 757-765 [DOI: 10.14569/IJACSA.2023.0140979]

Liang M F, Su J C, Schulter S, Garg S, Zhao S Y, Wu Y, et al. 2024. AIDE: an automatic data engine for object detection in autonomous driving//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 14695-14706 [DOI: 10.1109/CVPR52733.2024.01392]

Lin T Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. 2014. Microsoft COCO: common objects in context//Proceedings of the 13th European Conference on Computer Vision (ECCV). Zurich, Switzerland: Springer: 740-755 [DOI: 10.1007/978-3-319-10602-1_48]

Liu C G, Gao G S, Huang Z Y, Hu Z H, Liu Q J and Wang Y H. 2024. YOLC: you only look clusters for tiny object detection in aerial images. IEEE Transactions on Intelligent Transportation Systems, 25(10): 13863-13875 [DOI: 10.1109/TITS.2024.3386928]

Liu S H, Zha J L, Sun J, Li Z and Wang G. 2023. EdgeYOLO: an edge-real-time object detector//Proceedings of the 42nd Chinese Control Conference (CCC). Tianjin, China: IEEE: 7507-7512 [DOI: 10.23919/CCC58697.2023.10239786]

Liu Y, Sun P, Wergeles N and Shang Y. 2021a. A survey and performance evaluation of deep learning methods for small object detection. Expert Systems with Applications, 172: #114602 [DOI: 10.1016/j.eswa.2021.114602]

Liu Z, Lin Y T, Cao Y, Hu H, Wei Y X, Zhang Z, et al. 2021b. Swin Transformer: hierarchical vision transformer using shifted windows//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE: 9992-10002 [DOI: 10.1109/ICCV48922.2021.00986]

Pan X Y, Jia N X, Mu Y Z and Gao X R. 2023. Survey of small object detection. Journal of Image and Graphics, 28(9): 2587-2615 (潘晓英, 贾凝心, 穆元震, 高炫蓉. 2023. 小目标检测研究综述. 中国图象图形学报, 28(9): 2587-2615) [DOI: 10.11834/jig.220455]

Rasheed M R, Coleman S, Gardiner B, Vance P, McAteer C and Nguyen K. 2024. An EfficientNet-based transfer learning system for defect classification in manufacturing//Proceedings of the 22nd IEEE International Conference on Industrial Informatics (INDIN). Beijing, China: IEEE: 1-7 [DOI: 10.1109/INDIN58382.2024.10774443]

Redmon J, Divvala S, Girshick R and Farhadi A. 2016. You only look once: unified, real-time object detection//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE: 779-788 [DOI: 10.1109/CVPR.2016.91]

Ren S Q, He K M, Girshick R and Sun J. 2017. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6): 1137-1149 [DOI: 10.1109/TPAMI.2016.2577031]

Sanders A. 2016. An Introduction to Unreal Engine 4. New York, USA: AK Peters/CRC Press.

Shen D G, Wu G R and Suk H I. 2017. Deep learning in medical image analysis. Annual Review of Biomedical Engineering, 19: 221-248 [DOI: 10.1146/annurev-bioeng-071516-044442]

Wang A, Chen H, Liu L H, Chen K, Lin Z J, Han J G, et al. 2024.

YOLOv10: real-time end-to-end object detection//Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: #3429

Wang C Y, Bochkovskiy A and Liao H Y M. 2023. YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE: 7464-7475 [DOI: 10.1109/CVPR52729.2023.00721]

Wang J W, Xu C, Yang W and Yu L. 2021. A normalized Gaussian Wasserstein distance for tiny object detection [EB/OL]. [2021-10-26]. https://arxiv.org/pdf/2110.13389.pdf

Wang X D, Peng Y X and Shen C M. 2025. Efficient feature fusion for UAV object detection [EB/OL]. [2025-01-29]. https://arxiv.org/pdf/2501.17983.pdf

Wu K P, Xu Y F and Zhang J P. 2026. Lightweight multi-scale dynamic feature focusing network integrating spatial channel attention mechanism for autonomous driving object detection. Digital Signal Processing, 170: #105770 [DOI: 10.1016/j.dsp.2025.105770]

Yang J N, Liu S L, Wu J J, Su X Y, Hai N and Huang X L. 2025. Pinwheel-shaped convolution and scale-based dynamic loss for infrared small target detection. Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia, USA: AAAI Press: 9202-9210 [DOI: 10.1609/aaai.v39i9.32996]

Ye M, Shen J B, Lin G J, Xiang T, Shao L and Hoi S C H. 2022. Deep learning for person re-identification: a survey and outlook. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(6): 2872-2893 [DOI: 10.1109/TPAMI.2021.3054775]

Ying X Y, Xiao C, An W, Li R J, He X, Li B Y, et al. 2025. Visible-thermal tiny object detection: a benchmark dataset and baselines. IEEE Transactions on Pattern Analysis and Machine Intelligence, 47(7): 6088-6096 [DOI: 10.1109/TPAMI.2025.3544621]

Yu J H, Jiang Y N, Wang Z Y, Cao Z M and Huang T. 2016. UnitBox: an advanced object detection network//Proceedings of the 24th ACM International Conference on Multimedia. Amsterdam, the Netherlands: ACM: 516-520 [DOI: 10.1145/2964284.2967274]

Zeng S, Yang W Z, Jiao Y Y, Geng L and Chen X T. 2024. SCA-YOLO: a new small object detection model for UAV images. The Visual Computer, 40(3): 1787-1803 [DOI: 10.1007/s00371-023-02886-y]

Zhang G and Zhao X. 2025. An improved RT-DETR model for small object detection on construction sites//Proceedings of the 11th International Conference on Computing and Artificial Intelligence (ICCAI). Kyoto, Japan: IEEE: 47-54 [DOI: 10.1109/ICCAI66501.2025.00015]

Zhang Y F, Yu J Y, Wang Y Y, Tang S, Li H, Xin Z Y, et al. 2023. Small object detection based on hierarchical attention mechanism and multi-scale separable detection. IET Image Processing, 17(14): 3986-3999 [DOI: 10.1049/ipr2.12912]

Zhao Y A, Lyu W Y, Xu S L, Wei J M, Wang G Z, Dang Q Q, et al. 2024. DETRs beat YOLOs on real-time object detection//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 16965-16974 [DOI: 10.1109/CVPR52733.2024.01605]

Zheng Y C, Jing Y X, Zhao J F and Cui G M. 2025. LAM-YOLO: drones-based small object detection on lighting-occlusion attention mechanism YOLO. Computer Vision and Image Understanding, 261: #104489 [DOI: 10.1016/j.cviu.2025.104489]

Zheng Z H, Wang P, Liu W, Li J Z, Ye R G and Ren D W. 2020. Distance-IoU loss: faster and better learning for bounding box regression//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA: AAAI Press: 12993-13000 [DOI: 10.1609/aaai.v34i07.6999]

Zhou Y, Jiang Y T, Yang Z L, Li X X, Yang Y, Wang Y, et al. 2024. A small object detection method based on the attention mechanism and multi-level feature fusion//Proceedings of 2024 International Conference on Image Processing and Artificial Intelligence. [s.l.]: SPIE: #132132R [DOI: 10.1117/12.3035229]

Zhu L, Wang X J, Ke Z H, Zhang W and Lau R. 2023. BiFormer: vision transformer with bi-level routing attention//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE: 10323-10333 [DOI: 10.1109/CVPR52729.2023.00995]

Zou Z X, Chen K Y, Shi Z W, Guo Y H and Ye J P. 2023. Object detection in 20 years: a survey. Proceedings of the IEEE, 111(3): 257-276 [DOI: 10.1109/JPROC.2023.3238524]

作者简介

吴军,男,教授,博士生导师,主要研究方向为测绘遥感、视频安防、计算机视觉及嵌入式系统构建。
E-mail: wujun@guet.edu.cn

尹恒,通信作者,男,博士后,主要研究方向为计算机视觉和多源遥感数据智能处理与分析。E-mail: yinh@guet.edu.cn

蔡广震,男,硕士研究生,主要研究方向为人工智能和视频目标检测。E-mail: cgz@mails.guet.edu.cn

楚和轩,男,硕士研究生,主要研究方向为视频目标检测。
E-mail: chuhexuan03@mails.guet.edu.cn

徐刚,男,博士研究生,主要研究方向为计算机视觉。
E-mail: xugang@nimte.ac.cn

赵雪梅,女,教授,主要研究方向为测绘遥感和人工智能。
E-mail:zhaoxm@guet.edu.cn