

文章编号 1004-924X(2026)19-3119-19

SA-DINO: 面向卫星遥感图像的电力杆塔检测

石超君^{1,2}, 曹煜捷², 李 喆², 张 珂^{1,2*}, 王晓阳², 霍梦成²

(1. 华北电力大学 燕赵电力实验室, 河北 保定 071003;

2. 华北电力大学 电子与通信工程系, 河北 保定 071003)

摘要: 针对高分辨率卫星遥感图像中电力杆塔尺度差异显著, 小型配电塔特征弱, 复杂背景干扰严重以及边界框定位精度不足等问题, 提出一种基于尺度感知 (Scale-Aware, SA) Grounding DINO 的电力杆塔检测方法 (SA-DINO)。首先, 设计小目标细节保持分支, 在视觉特征进入 Transformer 编码器之前增强高分辨率特征层中的局部边缘、高频纹理和上下文信息, 以缓解小型配电塔特征在深层传递过程中的细节衰减问题。其次, 提出尺度感知语言引导 Query 选择模块, 根据不同尺度特征层的空间分辨率自适应分配 Query 数量, 使高分辨率层中的小目标候选区域获得更多进入解码器的机会, 同时保留全局高置信 Query 以兼顾大型输电塔检测。最后, 引入尺度感知小目标定位损失, 利用归一化 Wasserstein 距离度量预测框与真实框之间的几何差异, 并结合目标尺度自适应调整定位监督强度, 从而提升小目标边界框回归的稳定性。在 SRSPTD 数据集上的实验结果表明, SA-DINO 的 mAP50 和 mAP50-95 分别达到 80.9% 和 42.2%, 相比基线模型分别提升 3.5% 和 1.7%, 参数量增加 0.94%。实验结果验证了本文方法在复杂遥感场景下电力杆塔检测任务中的有效性。

关键词: 卫星遥感; 杆塔检测; Grounding DINO; 尺度感知; 小目标

中图分类号: TP391 **文献标识码:** A

doi: 10.3724/OPE.20263419.3119

CSTR: 32169.14.OPE.20263419.3119

SA-DINO: Detecting power towers based on satellite remote sensing imagery

SHI Chaojun^{1,2}, CAO Yujie², LI Zhe², ZHANG Ke^{1,2*}, WANG Xiaoyang², HUO Mengcheng²

(1. Yanzhao Electric Power Laboratory, North China Electric Power University,
Baoding 071003, China;

2. Department of Electronic and Communication Engineering, North China Electric Power University,
Baoding 071003, China)

* Corresponding author, E-mail: zhangkeit@ncepu.edu.cn

Abstract: Power tower detection in high-resolution satellite remote sensing imagery is hindered by substantial variations in tower scale, weak feature representations of small distribution towers, complex background interference, and limited bounding-box localization accuracy. To address these challenges, SA-DINO, a scale-aware power tower detection method based on Grounding DINO, is proposed. First, a detail-

收稿日期: 2026-07-18; **修订日期:** 2026-08-31.

基金项目: 国家自然科学基金资助项目 (No. 62206095, No. 62506129); 河北省自然科学基金“京津冀基础研究合作专项”资助项目 (No. F2024502017); 河北省中央引导地方科技专项资助项目 (No. 264Z2101G); 中央高校基本科研业务费面上项目 (No. 2026MS145)

preserving small-object feature branch is designed to enhance local edges, high-frequency textures, and contextual information in high-resolution feature maps before they are passed to the Transformer encoder, mitigating the degradation of small distribution tower features during deep feature propagation. Second, a scale-aware, language-guided query selection module is introduced to adaptively allocate queries according to the spatial resolution of feature maps at different scales. Small-object candidates in high-resolution feature maps are thus given greater access to the decoder, while global high-confidence queries are retained to preserve detection performance for large transmission towers. Finally, a scale-aware small-object localization loss is formulated using the normalized Wasserstein distance to quantify geometric discrepancies between predicted and ground-truth bounding boxes. The strength of localization supervision is adaptively adjusted according to object scale to improve the stability of small-object bounding-box regression. Experiments on the SRSPTD dataset demonstrate that SA-DINO achieves an mAP50 of 80.9% and an mAP50-95 of 42.2%, exceeding the baseline by 3.5% and 1.7%, respectively, with only a 0.94% increase in parameter count. These results demonstrate that scale-aware feature enhancement, query selection, and localization supervision can effectively improve power tower detection in complex remote sensing scenarios with minimal additional model parameters.

Key words: satellite remote sensing; tower detection; Grounding DINO; scale awareness; small object

1 引 言

随着电网规模持续扩大和输电走廊不断延伸,电力基础设施的智能化巡检已成为保障能源安全,降低运维成本和提升灾害响应能力的重要支撑^[1]。电力杆塔作为输配电系统中的关键承载结构,其空间分布、运行状态和周边环境变化将直接关系到电网的稳定运行^[2]。相比人工巡检和无人机巡检,卫星遥感能够在大范围、跨区域和复杂地理环境下获取连续观测数据,为电力杆塔快速定位、资产管理和风险评估提供了新的技术路径^[3]。因此,如何从高分辨率卫星遥感图像中准确识别不同尺度的电力杆塔,已成为遥感智能解译和电力运维交叉领域中的重要研究问题。

近年来,深度学习推动了遥感目标检测技术的快速发展。卷积神经网络和基于 Transformer 的目标检测模型已被广泛用于舰船、飞机、车辆与建筑物等遥感目标识别任务,并在复杂背景下表现出较强的特征建模能力^[4-5]。小目标检测研究表明,多尺度特征表达、注意力机制、上下文增强和定位损失优化是提升小尺寸目标检测精度的主要方向^[6-7]。在通用目标检测领域,DETR 系列^[8]方法通过端到端集合预测减少了人工设计组件,同时相较于传统目标检测模型检测精度得到了进一步提升。DINO^[9]进一步通过去噪训练和

改进的查询初始化提升了收敛效率与检测性能。在此基础上,Grounding DINO^[10]将语言信息引入目标检测过程,使模型能够根据文本提示进行开放词汇目标感知,从而为遥感场景下具有明确类别语义的电力杆塔检测提供了一种新的思路。

然而,直接将 Grounding DINO 应用于高分辨率卫星遥感图像中的电力杆塔检测仍面临明显挑战。首先,配电塔通常尺寸较小,纹理结构弱,在多层特征传递过程中容易发生细节衰减,导致浅层空间信息难以充分保留。其次,输电塔与配电塔在尺度和外观上差异显著,原始 Grounding DINO 采用全局 Top-k 查询选择策略,容易使语义响应更强的大目标或复杂背景区域占据更多查询名额,从而降低小目标进入解码器的概率^[10]。最后,小目标边界框对轻微位置偏移高度敏感,传统 IoU 及其扩展度量在目标尺度较小时容易因少量位置偏差产生较大的数值波动^[11],而 Grounding DINO 所采用的 L1 和 GIoU^[12]定位约束缺少针对不同目标尺度的差异化优化,仍难以充分适应配电塔等小尺度目标的定位需求。

为解决上述问题,本文提出一种名为 SA-DINO 的基于尺度感知 Grounding DINO 的卫星遥感图像电力杆塔检测方法。首先,设计小目标细节保持分支,在多尺度视觉特征进入 Transform-

er 编码器之前增强高分辨率层中的局部细节、高频边缘和上下文信息,以缓解小型配电塔特征在深层传播中的弱化问题。然后构建尺度感知语言引导 Query 选择模块,根据不同特征层的空间分辨率分配查询数量,并结合文本提示生成的语言引导分数,使更多查询关注高分辨率小目标区域,同时保留全局高置信查询以兼顾大型输电塔。最后,引入尺度感知小目标定位损失,将预测框和真实框建模为二维高斯分布,并基于归一化 Wasserstein 距离衡量二者的中心和尺度差异,同时根据真实框尺寸自适应调整定位权重,从而提升小目标框回归的稳定性。

2 相关工作

2.1 遥感目标检测与小目标检测

遥感目标检测旨在从卫星或航空影像中定位并识别具有特定语义类别的地物目标,在交通监测、城市管理、灾害评估和电力基础设施巡检等领域具有重要应用价值^[13]。近年来,深度学习的蓬勃发展推动了遥感图像目标检测领域的突破性进展,已广泛应用于电力杆塔等关键电力基础设施检测的下游任务中。目前,基于电力基础设施目标检测的主流算法主要分为以 R-CNN 系列^[14]为代表的二阶段检测算法和以 YOLO 系列^[15-16]为代表的一阶段检测算法。Lu 等^[17]整合了深度学习和量子神经网络技术,提出了一种量子增强的深度轻量级物体检测模型,用于电力塔基座遮挡检测。Huang 等^[18]针对多场景遥感图像提出了一种名为 DLR-YOLO 的轻量级电力塔检测网络,该网络利用动态低秩训练技术,在显著减少模型参数和计算复杂度的同时,实现了高检测精度。

与自然图像相比,遥感图像通常具有成像范围大、目标尺度变化明显、背景复杂和目标分布稀疏等特点,这使得检测模型不仅需要具备较强的语义表达能力,还需要有效保留小目标的空间细节信息^[19]。在电力杆塔检测任务中,上述问题更加突出。输电塔通常尺寸较大,结构特征相对明显,而配电塔在高分辨率卫星遥感图像中往往表现为小尺度、弱纹理和细长结构,其有效判别信息主要集中在局部边缘和高频细节区域。传统检测器虽然可以通过特征金字塔结构融合不

同层级特征,但在多次下采样和深层语义聚合过程中,小型杆塔的空间细节仍容易被削弱。同时,遥感图像中的道路、建筑物、阴影和植被纹理可能与杆塔结构产生相似响应,进一步增加了小目标识别难度。

针对小目标检测,现有研究通常从多尺度特征融合、注意力增强、上下文建模和损失函数优化等方面进行改进^[20]。其中,多尺度特征融合方法通过整合浅层空间信息和深层语义信息,提高模型对不同尺度目标的适应能力。注意力机制则通过通道或空间重标定突出目标相关区域,抑制复杂背景干扰。上下文建模能够利用目标周围环境信息增强小目标判别能力。Zha 等^[21]提出了一种基于分组可变形卷积的双向特征金字塔网络,并将其应用于输电塔的遥感目标检测。它基于分组可变形卷积构建的特征提取网络,通过上采样和最大池化操作,对不同尺度的特征图进行加权特征融合,从而获得更丰富的特征表示。Fu 等^[22]也提出了一种特征融合架构来生成多尺度特征层次结构,通过自上而下的路径用语义表示增强浅层的特征,并通过自下而上的路径将顶层的特征图与低级信息相结合,可以为多尺度对象形成强大的特征表示。Shi 等^[23]提出了一种用于高分辨率卫星遥感图像的电力杆塔检测网络,通过引入大型空间选择性核机制和注意力特征融合模块增大感受野提取更多上下文信息的同时,增强了模型的特征表示能力。

除特征表示优化外,边界框回归损失也是影响小目标检测性能的重要因素。传统 IoU 及其扩展度量主要依赖预测框与真实框之间的重叠区域,当目标仅占少量像素时,轻微的位置偏移就可能引起较大的 IoU 波动。针对这一问题,Xu 等^[11]将边界框建模为二维高斯分布,并提出归一化 Wasserstein 距离(Normalized Wasserstein Distance, NWD),通过连续的分分布距离衡量微小边界框之间的几何差异,同时结合排序式标签分配策略改善微小目标的训练样本分配。此后,研究人员进一步将 NWD 应用于边界框回归。Shi 等^[24]将 NWD 与 CIoU 结合构建 NWD-CIoU 回归损失,以提高无人机遥感图像中小目标定位的稳定性。Xiao 等^[25]将 NWD 引入 DINO 的定位损失,在 Transformer 检测框架下增强小目标的边

界框回归能力。Zhang 等^[26]进一步提出尺度敏感损失,根据目标尺度动态调节 NWD 与 CIoU 的相对贡献,以适应不同尺度目标的定位需求。上述研究表明,NWD 能够缓解 IoU 类度量对小目标位置偏移过度敏感的问题,但如何结合具体任务中的目标尺度分布,对不同尺度实例施加更有针对性的定位监督,仍具有进一步研究空间。

尽管上述方法能够从特征增强或定位优化等方面改善小目标检测性能,但现有方法大多面向通用 遥感或微小目标场景,尚未充分考虑同一电力塔数据集中配电塔与输电塔之间显著的尺度差异。因此,仍需要结合电力塔目标的尺度分布特点,从特征保持、查询分配和实例级定位监督等方面进行针对性优化。

2.2 视觉-语言目标检测模型

近年来,Transformer 类型的网络架构在目标检测领域得到了广泛应用。DETR 将目标检测建模为集合预测问题,通过 Transformer 编码器和解码器实现端到端检测,减少了传统检测器中人工设计的候选框和后处理步骤。随后,Deformable DETR^[27]、DINO^[9]等方法通过多尺度可变形注意力、去噪训练和改进的查询(Query)初始化策略进一步提升了模型收敛速度和检测精度。与传统卷积检测器相比,基于 Transformer 的检测器能够更好地建模全局关系和长距离依赖,这对于背景复杂、目标稀疏的遥感图像具有一定的优势^[28]。

在此基础上,视觉-语言目标检测进一步将文本语义引入目标检测过程,使模型能够根据给定文本提示定位相应目标。GLIP^[29],OWL-ViT^[30],OV-DETR^[31],Grounding DINO^[10]等方法通过图像-文本对齐或开放词汇学习,使检测模型具备更强的语义泛化能力。其中,Grounding DINO 将 DINO 检测框架与语言引导机制相结合,通过文本提示增强目标感知能力,能够根据类别名称或短语描述完成目标定位。因此,Grounding DINO 被广泛应用于输电线路检测^[32]、遥感图像目标检测^[33]等各种下游任务中。对于电力杆塔检测任务而言,数据集的类别名称“Distribution tower”和“Transmission tower”具有明确的语义类别,因此 Grounding DINO 为遥感图像中的电力杆塔检测提供了跨模态建模基础。

然而,原始 Grounding DINO 直接应用于电力杆塔遥感图像时仍存在一定局限。首先,Grounding DINO 的视觉特征需要经过主干网络、特征融合模块和 Transformer 编码器处理,小型配电塔的局部细节可能在特征传递过程中逐渐弱化。其次,在 Query 初始化阶段,模型通常依据语言引导分类分数从所有尺度层 Token 中进行全局 Top-k 选择。该策略能够选择语义响应较强的候选区域,但没有显式考虑 Token 所属的尺度层级。对于电力杆塔检测任务,大型输电塔或复杂背景区域往往更容易获得较高响应,而小型配电塔主要依赖高分辨率层中的局部细节,其候选 token 可能在全局竞争中被忽略。最后,原始边界框回归损失主要依赖 L1 和 GIoU 约束,对小目标轻微偏移的稳定性不足,难以充分适应配电塔这类小尺度目标的定位需求。因此,本文将基于上述分析,以 Grounding DINO 为基线模型进行针对性改进。

3 原 理

3.1 总体框架

针对高分辨率卫星遥感图像中电力杆塔尺度差异显著,小型配电塔特征稀疏以及复杂背景干扰严重等问题,提出一种基于尺度感知 Grounding DINO 的电力杆塔检测方法。该方法以 Grounding DINO 为基础框架,引入视觉特征增强、查询选择优化和定位损失约束 3 个层面的改进,以提升模型对不同尺度电力杆塔的检测能力,尤其增强小型配电塔的定位精度。

在原始 Grounding DINO 框架中,模型主要依赖多尺度特征、语言引导分类分数和全局 Top-k 查询选择机制完成目标定位。然而,对于遥感图像中的电力杆塔检测任务,配电塔通常尺寸较小,容易在深层特征传播和全局 Query 竞争中被削弱。为此,本文从 3 个方面进行改进,如图 1 所示。首先,设计小目标细节保持分支(Detail-Preserving Small-object Feature Branch, DPSF),在视觉特征进入 Transformer 编码器前增强高分辨率特征中的局部细节和边缘信息。其次,提出尺度感知语言引导 Query 选择模块(Scale-Aware Language-Guided Query Selection, SA-LGQS),根据不同尺度特征层的空间分辨率自适应分配

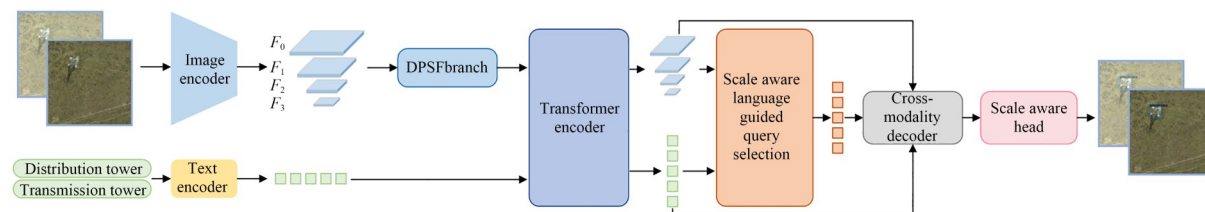


图 1 SA-DINO 模型总体结构

Fig. 1 Overview structure of SA-DINO model

Query 数量,使高分辨率层中的小目标候选区域获得更多进入解码器的机会。最后,引入尺度感知小目标定位损失 (Scale-Aware Small-object Localization Loss, SA-LM Loss),利用归一化 Wasserstein 距离度量预测框与真实框之间的几何差异,并根据目标尺寸自适应调整定位监督强度。

3.2 小目标细节保持分支

在高分辨率卫星遥感图像中,配电塔通常呈现出目标尺寸小、结构细长、纹理响应弱以及背景干扰强等特点。与输电塔相比,配电塔在图像中所占像素比例更低,其判别信息主要集中在局部边缘、细线结构和弱纹理区域。然而,在 Grounding DINO 的视觉特征提取过程中,图像特征需要经过主干网络和多尺度特征融合模块

进行逐层下采样与语义聚合。该过程虽然有助于增强深层语义表达能力,但也容易导致浅层高分辨率特征中的细节信息被削弱,尤其是小型配电塔的局部结构在深层特征传播中可能逐渐被复杂背景或大尺度语义信息淹没。因此,仅依赖原始多尺度特征输入 Transformer 编码器,可能难以充分保留小目标的空间细节信息。针对这一问题,设计了小目标细节保持分支,用于在视觉特征进入 Transformer 编码器之前对高分辨率特征层进行细节增强,从而提高模型对小型电力杆塔的感知能力。

DPSF 位于 Transformer 编码器之前,其核心思想是在不破坏原有多尺度语义表达的基础上,对包含丰富空间信息的高分辨率特征层进行轻量化增强,其结构如图 2 所示。设骨干网络输出

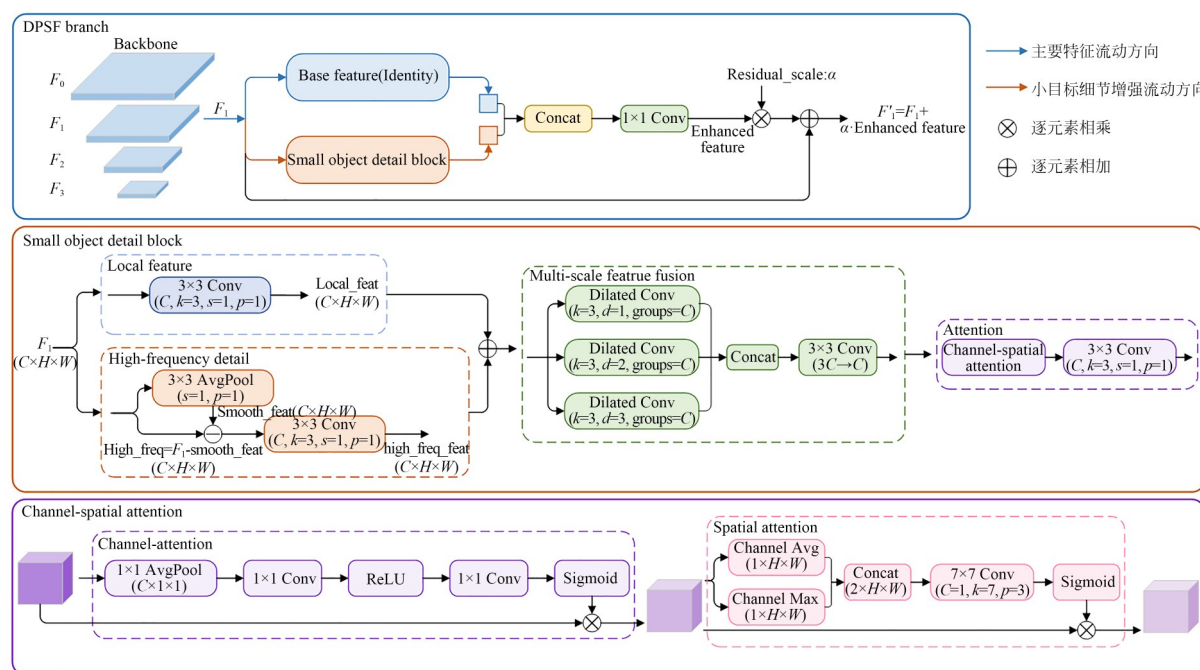


图 2 小目标细节保持分支结构

Fig. 2 Structure of detail-preserving small-object feature branch

的多尺度视觉特征为 $\mathcal{F} = F_0, F_1, F_2, F_3$, 其中 F_0 和 F_1 表示空间分辨率较高的特征层, 包含更多边缘、纹理和局部结构信息, 更适合表征小型配电塔。由于对 F_0 进行处理会极大增加模型的参数量, 从而导致显存占用巨大, 因此本文将 DPSF 主要作用于 F_1 。同时, 对其他分辨率特征层保持原始语义表达, 以避免对大目标检测和全局语义建模产生过强干扰。通过这种方式, DPSF 能够有针对性地增强小目标敏感层, 同时保持 Grounding DINO 原始多尺度特征结构的稳定。

对于输入的高分辨率特征 F_1 主要是通过 Small Object Detail Block 从而进行细节增强。该模块首先通过局部细节提取操作增强邻域范围内的空间响应。由于小型配电塔通常只覆盖有限像素区域, 其有效信息往往表现为局部边缘和线状纹理, 因此局部卷积能够在较小感受野内捕获目标的细粒度结构。该过程可表示为:

$$F_{\text{local}} = \text{Conv}_{3 \times 3}(F_1), \quad (1)$$

其中: $\text{Conv}_{3 \times 3}(\cdot)$ 表示 3×3 卷积操作。与大尺度卷积相比, 局部卷积能够更好地保持小目标的空间位置和结构边界, 避免过大的感受野引入过多背景信息, 从而为后续细节增强提供更加稳定的局部特征基础。

为了进一步突出配电塔的高频结构, DPSF 引入高频残差信息提取过程。遥感图像中的大面积地物背景通常具有较平滑的纹理变化, 而杆塔、导线支撑结构及其阴影边界往往表现为局部突变和细线响应。因此, 可以通过原始特征与局部平滑特征之间的差异获得高频信息:

$$F_{\text{hf}} = F_1 - \text{AvgPool}_{3 \times 3}(F_1), \quad (2)$$

其中: $\text{AvgPool}_{3 \times 3}(\cdot)$ 表示局部平均池化操作。该操作能够在一定程度上抑制平坦背景区域, 同时保留边缘、角点和细粒度纹理变化。对于小型配电塔而言, 这类高频信息有助于强化目标轮廓和局部结构, 使其在后续特征交互过程中不易被背景噪声覆盖。随后将其与 F_{local} 逐元素相加得到 F_{th} 。

仅依赖局部高频信息可能会放大遥感图像中的道路边缘、建筑轮廓及阴影边界等干扰区域。为提升增强特征的上下文判别能力, DPSF

进一步引入多尺度上下文聚合结构。具体而言, 模块采用不同膨胀率的空洞卷积对局部增强特征进行并行建模, 以同时捕获小目标本身的细节信息及其邻域上下文关系。该过程可表示为:

$$F_{\text{ctx}} = \text{Concat}(\text{Conv}_{d_1=1}(F_{\text{th}}), \text{Conv}_{d_2=2}(F_{\text{th}}), \text{Conv}_{d_3=3}(F_{\text{th}})), \quad (3)$$

其中: $\text{Conv}_{d_i=i}(\cdot)$ 表示膨胀率为 i 的卷积操作, $\text{Concat}(\cdot)$ 表示通道维度拼接。通过不同尺度感受野的联合建模, DPSF 能够在保持小目标细节的同时引入适度上下文约束, 使模型不仅关注局部强响应区域, 还能结合周围结构判断其是否符合电力杆塔的空间形态特征。

在完成局部细节和上下文信息提取后, DPSF 进一步利用注意力机制对增强特征进行自适应筛选。由于高分辨率遥感特征中包含大量背景纹理, 仅对所有区域进行均匀增强可能会引入噪声。因此, 本文采用通道和空间联合重标定策略, 使网络能够自动突出与电力杆塔相关的响应通道和空间位置。通道注意力用于建模不同特征通道的重要性, 增强与杆塔结构、边缘纹理和局部形态相关的特征响应。空间注意力则用于强调潜在目标区域, 并抑制无关背景区域。通过注意力重标定, DPSF 能够在复杂背景中更加有选择性地增强小目标特征, 降低无效纹理对检测结果的影响。

为了保持预训练 Grounding DINO 特征分布的稳定性, DPSF 没有直接替换原始高分辨率特征, 而是采用残差融合方式将增强信息注入原始特征层。通过残差系数 α 控制增强分支对原始特征的影响强度。当 α 较小时, 模块能够以渐进方式补充小目标细节信息, 避免在训练初期对原有特征表示造成过大扰动。

总体而言, 小目标细节保持分支的作用主要体现在 3 个方面。首先, 该分支通过局部卷积和高频残差提取增强了小型配电塔的边缘、纹理和细线结构, 有助于缓解小目标在深层特征传播过程中的细节衰减。其次, 多尺度上下文聚合使增强特征不仅包含局部细节, 还具备一定范围的空间上下文信息, 从而提升模型对复杂遥感背景中真实杆塔目标的判别能力。最后, 注意力重标定和残差融合机制保证了增强过程的选择性和稳定性, 使模型能够在保持原始 Grounding DINO

语义建模能力的同时,提高对小尺度目标的感知能力。通过在 Transformer 编码器之前引入 DPSF,后续的过程能够获得更加清晰的小目标视觉表征,从而为电力杆塔的准确检测提供更可靠的特征基础。

3.3 尺度感知语言引导 Query 选择模块

在 Grounding DINO 中,Transformer 编码器输出的视觉特征 Token 通常会根据其与其与文本语义之间的匹配分数进行全局 Top-k 筛选(如图 3(a)所示),并将得分最高的若干 Token 作为 Transformer 解码器的初始 Queries。这种查询初始化方式能够利用语言信息引导目标区域选择,使模型具备较强的开放词汇目标感知能力。然

而,对于高分辨率卫星遥感图像中的电力杆塔检测任务,不同类别目标之间存在显著的尺度差异。输电塔通常尺寸较大、结构较完整,在深层语义特征中容易获得较高的语言匹配分数;而配电塔通常尺寸较小、纹理弱、空间占比低,其有效特征更多保留在高分辨率特征层中。若仍采用原始全局 Top-k 查询选择策略,所有尺度层的 Token 会在同一候选集合中竞争 Query 名额,导致语义响应更强的大目标区域或复杂背景区域更容易被选中,而小型配电塔对应的高分辨率 Token 可能被忽略。为缓解这一问题,提出尺度感知语言引导 Query 选择模块,如图 3(b)所示,在语言语义引导的基础上进一步引入尺度层级

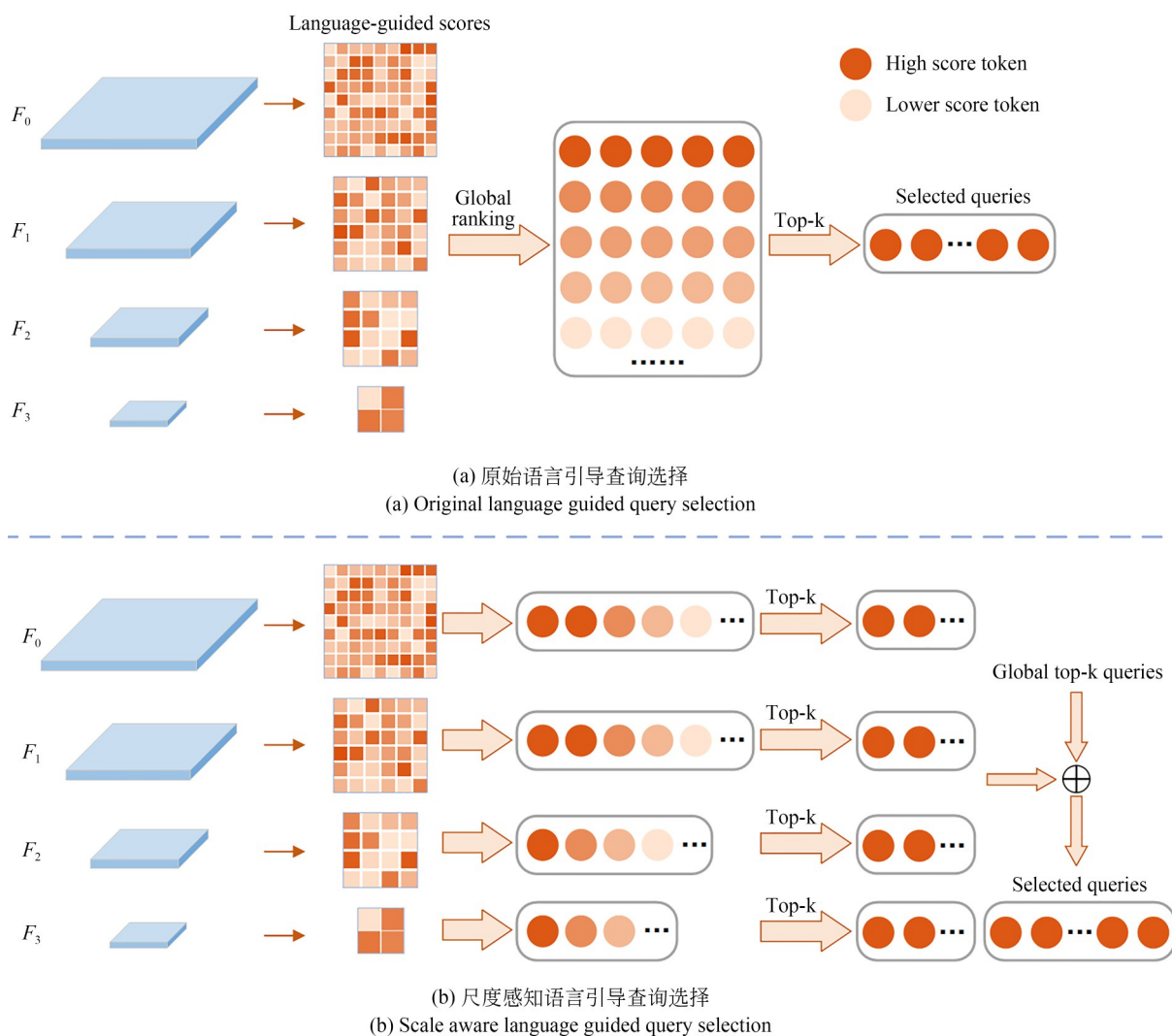


图 3 原始语言引导查询选择机制和本文提出的尺度感知语言引导查询选择机制

Fig. 3 Original language-guided query selection mechanism and scale-aware language-guided query selection mechanism proposed in this paper

约束,使模型能够根据不同尺度特征层的空间分辨率合理分配 Queries,从而提高小目标候选区域进入解码器的概率。

SA-LGQS 位于 Transformer 编码器和 Transformer 解码器之间,用于替代原始 Grounding DINO 中单纯依赖全局 Top-k 的 Query 选择过程。该模块的输入包括 Transformer 编码器输出的视觉特征序列、语言引导分类分数、多尺度特征层的空间尺寸以及各尺度层 Token 在特征序列中的起始索引。设编码器输出的视觉特征序列表示为 $M \in \mathbb{R}^{N \times C}$, 其中 N 为所有尺度层 Token 的总数, C 为特征维度。对应的语言引导分类分数表示为 $S \in \mathbb{R}^{N \times C_l}$, 其中 C_l 表示文本提示相关的类别或文本 Token 维度。对于第 i 个视觉 Token, 其语言引导分数可表示为:

$$s_i = \max_{c \in \mathcal{C}} S_{i,c}, \quad (4)$$

其中: $\mathcal{C}$ 表示文本提示对应的类别集合。对于本文任务而言, 文本提示主要包括“Distribution tower”和“Transmission tower”, 因此 s_i 反映了第 i 个视觉 Token 与电力杆塔语义之间的最大匹配程度。该得分仍然保留了 Grounding DINO 的语言引导优势, 使 Query 选择过程能够围绕目标语义展开, 而不是仅依赖视觉特征响应。

由于编码器的视觉特征序列由多个尺度层的视觉 Token 拼接而成, 不同 Token 实际对应不同空间分辨率的特征层。通常情况下, 高分辨率特征层包含更多局部纹理、边缘和空间位置信息, 更有利于小型配电塔的感知。低分辨率特征层具有更强的语义抽象能力, 更适合大型输电塔和全局上下文建模。因此, SA-LGQS 首先根据多尺度特征图的空间尺寸以及每一层在特征序列中的起始位置, 将所有视觉 Token 重新划分为不同尺度层的 Token 集合。通过该划分过程, 模型能够明确每个候选 Token 的尺度来源, 为后续尺度感知 Query 分配提供基础。

在完成尺度层划分后, SA-LGQS 根据预设的尺度分配比例为不同特征层分配 Query 数量。设解码器所需的总 Query 数量为 K , 其中用于尺度感知层级分配的 Query 数量为 K_s , 各尺度层的分配比例为 $\{r_l\}_{l=0}^L$ ($0 \leq L \leq 3$), 则第 l 个尺度层分配到的 Query 数量可表示为:

$$K_l = \lfloor K_s r_l \rfloor. \quad (5)$$

在实际设置中, 高分辨率层通常被赋予更大的分配比例, 使其能够获得更多 Query 名额。例如, 最高分辨率层对应更多小目标候选区域, 因此可分配较高比例的 Query, 而低分辨率层主要用于补充大目标和全局语义信息, 因此分配相对较少的 Query。与原始全局 Top-k 策略相比, 这种层级分配机制能够避免高分辨率小目标 Token 在全局竞争中被大目标或背景区域压制, 从结构上提高小型配电塔被解码器关注的机会。

为了进一步缓解小目标语义响应偏弱的问题, SA-LGQS 在语言引导分数的基础上引入尺度层得分偏置。对于属于第 l 个尺度层的第 i 个 Token, 其修正后的选择分数定义为:

$$\tilde{s}_i = s_i + b_l, i \in \Omega_l, \quad (6)$$

其中: b_l 表示第 l 个尺度层对应的得分偏置, Ω_l 表示第 l 个尺度层 Token 的索引集合。对于高分辨率特征层, 可以设置较小的正偏置, 以适度提高小目标候选 Token 的选择优先级。对于低分辨率特征层, 则可以保持原始分数或设置较弱偏置, 以防止大尺度语义 Token 过度占用 Query 数量。该得分偏置并不是直接改变模型的分类结果, 而是在 Query 初始化阶段对不同尺度层进行轻量补偿, 使弱响应的小目标区域在候选筛选中获得更公平的竞争机会。由于偏置幅度较小, 模块仍然主要依赖语言引导分数进行目标相关性判断, 从而避免过度引入人为先验。

随后, SA-LGQS 在每个尺度层内部独立执行 Top-k 查询选择。对于第 l 个尺度层, 其选出的 Query 索引集合表示为:

$$Q_l = \text{TopK}(\tilde{s}_i, K_l), i \in \Omega_l. \quad (7)$$

该过程保证了每个尺度层都能够根据自身内部的语言引导响应选择最具目标相关性的 Token, 而不是让所有尺度层在全局范围内无约束竞争。对于小型配电塔而言, 高分辨率层内部的 Top-k 选择能够保留更多局部候选区域, 使其在后续解码过程中获得更充分的交互和回归优化。对于大型输电塔而言, 中低分辨率层仍然保留一定 Query 分配, 因此模型不会完全丢失对大目标语义区域的关注。

考虑到电力杆塔检测中不仅存在小型配电塔, 也存在尺寸较大、语义响应较强的输电塔, 若

仅依赖尺度层级分配,可能会在一定程度上削弱原始 Grounding DINO 对全局高置信目标的捕获能力。因此,SA-LGQS 额外设计了全局 Top-k 保留分支,用于从所有尺度层 Token 中直接选择语言引导分数最高的一部分候选 Queries。该过程可表示为:

$$Q_g = \text{TopK}(s_i, K_g), i \in \Omega, \quad (8)$$

其中: Ω 表示所有尺度层 Token 的索引集合, K_g 表示全局保留 Query 数量。全局 Top-k 分支能够保留原始查询选择策略中对高置信目标区域的优势,特别是对大型输电塔和显著目标区域具有重要作用。通过引入该分支,SA-LGQS 在增强小目标 Query 分配的同时,避免模型过度偏向高分辨率层,从而保持不同尺度目标之间的检测平衡。

最终,SA-LGQS 将各尺度层选出的 Queries 与全局保留 Queries 进行融合,形成最终输入解码器的 Query 集合:

$$Q = \text{Merge}(Q_0, Q_1, \dots, Q_{L-1}, Q_g), L = 4. \quad (9)$$

在融合过程中,如果不同分支选中了相同的 Token,则进行去重处理。若去重后 Query 数量不足,则根据语言引导分数从剩余候选 Token 中进行补齐,以保证最终 Query 数量与解码器设定保持一致。获得最终索引集合后,模型从编码器输出的视觉特征序列中提取对应 Token 作为解码器的 Queries,并送入跨模态 Transformer 解码器,与文本特征和视觉特征序列进一步交互,最终完成目标分类和边界框回归。

SA-LGQS 的作用主要体现在语言引导和尺度感知两个方面。语言引导机制保证了 Query 选择过程与目标语义相关,使模型能够围绕“Distribution tower”和“Transmission tower”两类电力杆塔进行候选区域筛选。尺度感知机制则根据不同特征层对不同尺度目标的表达能力进行 Query 分配,使高分辨率层中的小目标候选区域获得更多进入解码器的机会。二者结合后,模型不仅能够利用文本提示增强目标相关性判断,还能够缓解小目标在全局 Query 竞争中被忽略的问题。对于高分辨率卫星遥感图像而言,该模块能够有效改善小型配电塔的召回能力,同时通过全局 Top-k 保留分支维持大型输电塔的检测稳定性。

总体而言,SA-LGQS 在不改变 Grounding DINO 主体检测流程的基础上,通过轻量化的尺度感知查询重分配策略,提高了模型对多尺度电力杆塔目标的适应性,为复杂遥感场景下的小目标检测提供了更加合理的 Query 初始化机制。

3.4 尺度感知小目标定位损失

在高分辨率卫星遥感图像的电力杆塔检测任务中,边界框定位精度直接影响模型对杆塔目标的识别可靠性。与输电塔相比,配电塔通常具有更小的空间尺度和更弱的纹理响应,其目标框在图像中所占像素比例较低。当预测框发生轻微中心偏移或宽高变化时,小目标的重叠区域会迅速减小,导致 IoU 类指标产生较大波动。对于小型配电塔而言,即使预测框与真实框在视觉上已经较为接近,轻微的位置偏差也可能导致 IoU 明显下降,从而使训练过程中的定位监督不够平滑。

针对该问题,归一化 Wasserstein 距离通过将边界框建模为二维高斯分布,从中心位置和尺度两个方面连续度量预测框与真实框之间的几何差异。由于其不完全依赖边界框重叠区域,NWD 已被后续研究应用于小目标标签分配和边界框回归,并进一步与 IoU 类损失结合,以提高小目标定位的稳定性。然而,标准 NWD 通常对不同尺度目标采用统一的损失权重,当同一数据集中同时存在尺度差异明显的目标时,难以根据目标尺寸自适应调整定位监督强度。对于本文电力塔检测任务,配电塔尺寸较小,而输电塔通常具有更大的空间尺度,因此若对两类目标施加相同的 NWD 监督强度,仍不能充分突出小目标在定位优化中的贡献。基于此,本文不改变 NWD 的基本距离度量形式,而是在其基础上引入与真实目标尺度相关的自适应权重动态调整定位监督强度,构建尺度感知小目标定位损失。该方法根据真实框尺寸连续调整 NWD 项的监督强度,使尺度越小的目标获得越大的定位权重,而对于较大的输电塔,权重逐渐趋近于常规值。同时保留 Grounding DINO 原有的 L1 Loss 和 GIoU Loss,以维持整体边界框回归的稳定性。

SA-LM Loss 的基本思想是将边界框由传统的矩形表示转换为二维高斯分布表示。设预测

框和真实框分别为:

$$B_p = (x_p, y_p, w_p, h_p), B_g = (x_g, y_g, w_g, h_g), \quad (10)$$

其中: x 和 y 表示边界框中心点坐标, w 和 h 分别表示边界框宽度和高度。对于每个边界框, 其中心点可以看作二维高斯分布的均值, 宽度和高度则可以用于描述该分布在水平方向和垂直方向上的扩散范围。因此, 预测框和真实框可分别表示为:

$$\mathcal{N}_p = \mathcal{N}(\mu_p, \Sigma_p), \mathcal{N}_g = \mathcal{N}(\mu_g, \Sigma_g), \quad (11)$$

均值向量定义为:

$$\mu_p = \begin{bmatrix} x_p \\ y_p \end{bmatrix}, \mu_g = \begin{bmatrix} x_g \\ y_g \end{bmatrix}. \quad (12)$$

协方差矩阵定义为:

$$\Sigma_p = \begin{bmatrix} \frac{w_p^2}{4} & 0 \\ 0 & \frac{h_p^2}{4} \end{bmatrix}, \Sigma_g = \begin{bmatrix} \frac{w_g^2}{4} & 0 \\ 0 & \frac{h_g^2}{4} \end{bmatrix}. \quad (13)$$

在该表示方式下, 边界框不再仅由离散的四个坐标确定, 而是被描述为具有中心位置和尺度扩展的连续分布。这样可以在一定程度上降低小目标定位对重叠面积的过度依赖, 使预测框与真实框之间的差异度量更加平滑。

在二维高斯分布表示的基础上, 标准 NWD 采用 Wasserstein 距离衡量预测框与真实框之间的几何差异。由于上述协方差矩阵为对角矩阵, 两个边界框之间的二阶 Wasserstein 距离可以简化为:

$$W_2^2(B_p, B_g) = (x_p - x_g)^2 + (y_p - y_g)^2 + \frac{(w_p - w_g)^2 + (h_p - h_g)^2}{4}. \quad (14)$$

该距离由两部分组成: 第一部分为预测框与真实框中心点之间的距离, 用于刻画目标位置偏移; 第二部分为宽度和高度之间的差异, 用于刻画目标尺度变化。与 IoU 主要依赖重叠区域不同, Wasserstein 距离直接度量两个框在中心位置和尺寸形态上的连续差异, 因此对于小目标轻微偏移具有更稳定的响应。尤其在配电塔这类小尺度目标中, 即使预测框与真实框的重叠面积发生明显变化, Wasserstein 距离仍能够从几何分布角度提供连续的优化信号。

为了将距离度量转换为可用于训练的相似度形式, 标准 NWD 对 Wasserstein 距离进行指数归一化处理, 得到归一化 Wasserstein 距离 (Nor-

malized Wasserstein Distance, NWD):

$$\text{NWD}(B_p, B_g) = \exp\left(-\frac{W_2(B_p, B_g)}{C}\right), \quad (15)$$

其中: C 为归一化常数, 用于控制 Wasserstein 距离经指数映射后的衰减速度。该参数与数据集中目标的典型空间尺度有关, 较小的 C 对边界框几何偏差更加敏感, 而较大的 C 可获得更加平滑的相似度变化。根据 SRSPTD 数据集的目标尺度分布, 并结合不同候选参数的设置, 本文所有实验统一取 C 为 64。需要指出的是, C 在训练过程中保持固定, 不随单个目标尺度变化。针对不同尺度目标的差异化定位监督由后续尺度感知权重 ω_s 实现。因此, 基于该相似度定义, 基础 NWD 损失可表示为:

$$\mathcal{L}_{\text{NWD}} = 1 - \text{NWD}(B_p, B_g). \quad (16)$$

该损失将边界框之间的几何距离转化为连续可优化的损失形式, 可以作为传统 IoU 类损失的补充, 从而增强小目标框回归过程的稳定性。

然而, 仅使用 NWD Loss 仍然对不同尺度目标采用相同的监督强度, 无法充分体现配电塔和输电塔之间的尺度差异。为进一步增强模型对小目标的定位关注, 本文在 NWD Loss 的基础上引入尺度感知权重。设真实框的等效尺度为:

$$s_g = \sqrt{w_g h_g}, \quad (17)$$

其中: s_g 可以反映目标框的整体尺寸。当目标越小时, s_g 越小; 当目标越大时, s_g 越大。根据该尺度定义, 尺度感知权重表示为:

$$\omega_s = \min\left(1 + \beta \cdot \exp\left(-\frac{s_g}{\tau}\right), \omega_{\max}\right), \quad (18)$$

其中: β 表示小目标权重增强系数, τ 控制尺度权重随目标尺寸增大而衰减的速度 $\omega_{\max}$ 用于限制最大权重, 避免极小目标在训练过程中产生过大的梯度影响。当目标尺寸较小时, 指数项取值较大, ω_s 随之增大, 使模型对小型配电塔施加更强的定位约束。当目标尺寸较大时, 指数项逐渐趋近于 0, ω_s 接近 1, 使大型输电塔基本保持常规定位监督强度。因此, 该权重并不改变 NWD 本身的几何距离定义, 而是在实例层面对其定位贡献进行尺度自适应调节。通过这种方式, 损失函数能够根据真实目标尺度自动调整回归权重, 从而实现对不同尺寸杆塔目标的差异化优化。

最终,尺度感知小目标定位损失定义为:

$$\mathcal{L}_{\text{SA-LM}} = \omega_s \cdot \mathcal{L}_{\text{NWD}}. \quad (19)$$

该损失同时包含几何分布距离和尺度自适应权重两部分。其中,归一化 Wasserstein 距离用于平滑刻画预测框与真实框之间的中心和宽高差异,尺度感知权重用于提升小目标样本在回归训练中的贡献。二者结合后,模型能够在保持大目标定位稳定性的同时,增强对小型配电塔边界框的优化能力。

整体边界框回归损失表示为:

$$\mathcal{L}_{\text{reg}} = \lambda_{\text{L1}} \mathcal{L}_{\text{L1}} + \lambda_{\text{GIoU}} \mathcal{L}_{\text{GIoU}} + \lambda_{\text{nwd}} \mathcal{L}_{\text{SA-LM}}, \quad (20)$$

其中: λ_{L1} , λ_{GIoU} 和 λ_{nwd} 分别表示三项损失的权重。L1 Loss 主要约束预测框坐标的直接回归误差, GIoU Loss 从区域重叠和空间包围关系角度优化边界框质量,而 SA-LM Loss 则从分布距离和尺度感知角度补充小目标定位监督。三者共同作用,使模型既能够保持原始 Grounding DINO 边界框回归的稳定性,又能够针对遥感图像中小型配电塔定位不稳定的问题进行专门优化。

SA-LM Loss 的引入主要提升了模型对小尺度目标的定位稳定性。对于小型配电塔,预测框的轻微偏移往往会导致 IoU 快速下降,而 SA-LM Loss 能够通过 Wasserstein 距离提供更加连续的几何约束,使模型在训练过程中获得更稳定的梯度反馈。同时,尺度感知权重能够提升小目标样本在损失计算中的相对贡献,促使模型更加关注配电塔这类难检测目标的边界框回归质量。对于大型输电塔,尺度权重逐渐接近常规值,不会对其定位过程造成过强干扰。因此,SA-LM Loss 在整体上实现了小目标强化与大目标稳定之间的平衡,有助于提高复杂遥感场景下不同尺度电力杆塔的检测性能。

4 实验结果与分析

在本节中将通过实验展示 SA-DINO 在卫星遥感图像电力杆塔检测中的性能。先介绍实验环境和参数,随后描述所使用的数据集,以及评估指标的选择。接下来,将 SA-DINO 与其他检测方法进行比较。最后,进行消融实验以分析各个模块的影响,以展示本文提出方法的有效性。

4.1 实验设置

实验使用 MMDetection 3. x 框架中的 MM-

Grounding DINO^[34]作为基线模型,并在此基础上对其进行改进。模型采用 Swin Transformer-B 作为视觉主干网络,使用在大规模目标检测数据集上预训练的 Grounding DINO 权重进行初始化,以提高模型的特征表达能力和训练稳定性。设置检测类别为配电塔(Distribution tower)和输电塔(Transmission tower)(下文简称 DT 和 TT),并将其作为文本提示输入文本编码器。解码器查询数量设置为 900,边界框回归分支采用 L1 Loss 和 GIoU Loss 作为基础定位损失。训练过程中,训练 batch_size 设置为 4,训练轮次设置为 40 epochs,同时采用 AdamW 优化器,初始学习率设置为 0.000 1,其中视觉主干网络学习率倍率设置为 0.1,文本编码器参数保持冻结,以充分利用预训练语言表征并降低训练成本。所有实验均在相同数据划分,训练策略和评价指标下进行,以保证不同方法之间的公平比较。详细的实验运行环境如表 1 所示。

表 1 实验运行环境

Tab. 1 Experimental runtime environment

参数	配置
CPU	AMD EPYC 7453
GPU	NVIDIA GeForce RTX 4090 24 G
操作系统	Ubuntu 20.04.4
RAM	251 G
CUDA 版本	12.0
Python 版本	Python 3.8.20
深度学习框架	PyTorch 1.10.1

4.2 数据集

本文中的实验采用了 SRSPTD 数据集来验证所提方法的有效性。SRSPTD 数据集源自美国杜克大学所采集的输配电基础设施影像资料。该数据集的数据源覆盖广泛,选取了包括美国的亚利桑那、康涅狄格、堪萨斯、北卡罗来纳以及奥地利的陶朗加和达尼丁在内的 6 个典型地区。在地理与环境特征上,该数据集极具多样性,包括了沙漠、平原、森林及沿海 4 种地貌类型。并且跨越了城市、郊区与农村 3 种不同的人口密度区域。经由筛选、裁剪及重新标注等预处理工序,最终构建了包含 2 740 张样本的卫星遥感电力杆塔数据集。数据集中包含配电塔

(DT)和输电塔(TT)两个类别,并按照8:1:1的比例进行划分,其中训练集包含2 194张图像,验证集包含272张图像,测试集包含274张图像。为进一步分析数据集中电力杆塔的尺度特征,对全部标注目标的边界框面积进行了统计,结果如图4所示。

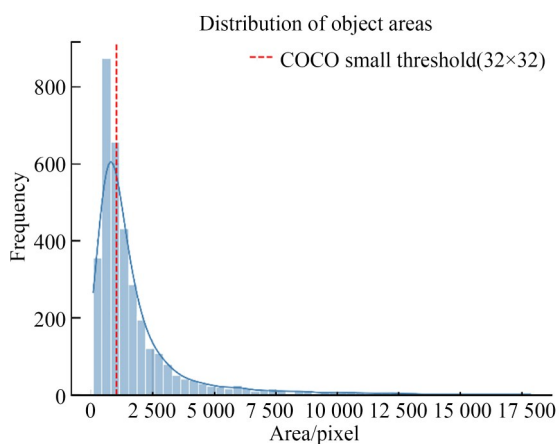


图4 SRSPTD数据集目标区域分布

Fig. 4 Distribution of target regions in the SRSPTD dataset

数据集中共包含3 590个目标实例,按照COCO目标尺度划分标准,边界框面积小于 32×32 像素的目标定义为小目标,面积介于 32×32 和 96×96 之间的目标定义为中等目标,大于 96×96 像素的目标定义为大目标。其中,小目标共1 693个,占全部目标的47.16%;中等目标共1 746个,占48.64%;大目标仅151个,占4.21%,目标面积分布呈明显的长尾特征,大量目标集中于 32×32 像素附近及以下。上述统计结果表明,SRSPTD数据集中小尺度电力杆塔占比较高,且多数目标的空间尺寸较小,这进一步说明了在该数据集上开展小目标特征增强、尺度感知查询选择和定位优化的必要性。

4.3 评价指标

为了客观评价改进模型对遥感图像中电力杆塔的检测性能,采用平均精确率(mean Average Precision, mAP)作为指标来评估检测的精度,并利用参数量(Params)来评估所提出模型的复杂度。而mAP的计算与精确率(Precision, P)和召回率(Recall, R)有关,精确率表示模型对于电力杆塔的检测精度,召回率表示模型是否能够

检测到所有电力杆塔的能力,计算公式如下:

$$P = \frac{TP}{TP + FP}, \quad (21)$$

$$R = \frac{TP}{TP + FN}, \quad (22)$$

其中:TP代表正样本被检测识别为正样本的数量,FP代表负样本被检测识别为正样本的数量,FN代表正样本被检测识别为负样本的数量。AP则用于评估某一个类别的检测精度,其计算公式如下:

$$AP = \int_0^1 P(R) dR. \quad (23)$$

最后,mAP则代表了所有标签类别的平均精度,mAP的数值越大则代表了模型的检测性能越好,其计算公式如下:

$$mAP = \frac{1}{N} \sum_{i=1}^N AP_i, \quad (24)$$

其中N代表数据集中的类别数。在本文中将同时使用mAP50和mAP50-95共同评估模型的性能。

4.4 对比实验

为验证本文所提出方法的有效性,选取多种具有代表性的目标检测模型进行对比实验,包括两阶段检测算法Faster R-CNN,一阶段YOLO系列检测算法YOLOv8s, YOLOv9s, YOLOv11s, YOLOv12s和YOLOv13s, Transformer系列算法Deformable-DETR和RT-DETR,以及视觉-语言检测模型GLIP, Grounding DINO和MM-Grounding DINO。为进一步验证本文方法在遥感小目标检测方面的性能,补充选取近期提出的遥感小目标检测模型SKYDE^[35]进行对比。上述方法涵盖传统卷积检测器、实时检测模型、Transformer检测器、视觉-语言模型以及面向遥感小目标的专用检测模型,可以较全面地评估本文方法在卫星遥感图像电力杆塔检测任务中的性能表现,实验结果如表2所示。其中,MM-Grounding DINO为本文方法的基线模型,本文方法在其基础上引入小目标细节保持分支、尺度感知语言引导Query选择模块和尺度感知小目标定位损失,并命名为SA-DINO。

从实验结果可以看出,不同检测模型在配电塔和输电塔检测任务上表现出明显差异。Faster R-CNN的mAP50为64.7%,mAP50-95为

表 2 SRSPTD数据集上不同模型的检测结果

Tab. 2 Detection results of different models on SRSPTD dataset

模 型	DT/%	TT/%	mAP50/%	mAP50-95/%	Params/M	FLOPs/G
Faster-RCNN	57.6	71.8	64.7	29.2	28.3	473.2
YOLOv8s	71.1	74.2	72.6	35.5	11.1	28.6
YOLOv9s	69.7	80.3	75.0	37.8	9.8	39.6
YOLOv11s	67.4	78.6	73.0	37.0	9.4	21.5
YOLOv12s	70.3	73.6	71.9	36.0	9.1	19.6
YOLOv13s	66.8	82.3	74.5	37.7	9.0	21.0
Deformable-DETR	64.2	68.9	66.6	27.8	40.2	173.3
RT-DETR	58.2	70.8	64.5	19.9	32.8	108.0
GLIP	71.9	80.8	76.4	39.0	232	434
Grounding DINO	75.0	79.5	77.3	37.9	233.0	428.2
MM-Grounding DINO	76.0	78.7	77.4	40.5	233.0	428.2
SKYDET	73.2	85.8	79.5	40.1	127.1	351.7
SA-DINO	77.5	84.4	80.9	42.2	235.2	450.2

29.2%,说明传统两阶段检测器虽然具有较稳定的候选框生成机制,但在高分辨率遥感图像中对小尺度和复杂背景目标的适应能力有限。YOLO系列模型整体表现优于Faster R-CNN,其中YOLOv9s的mAP50达到75.0%,YOLOv13s在输电塔上的AP达到82.3%,表明轻量化一阶段检测器在电力杆塔检测任务中具有较好的检测效率和一定的精度优势。然而,YOLO系列模型在配电塔上的检测性能仍存在波动,例如YOLOv13s的DT检测精度仅为66.8%,说明其对小型配电塔的细粒度特征表达仍不够稳定。

对于Transformer系列检测模型,Deformable-DETR和RT-DETR的mAP50分别为66.6%和64.5%,整体性能低于YOLO系列和多模态检测模型。这表明在当前数据规模和目标特征条件下,仅依赖通用Transformer检测框架难以充分适应电力杆塔遥感检测任务。相比之下,GLIP,Grounding DINO和MM-Grounding DINO等视觉语言检测模型取得了更好的检测结果,其中GLIP的mAP50为76.4%,Grounding DINO和MM-Grounding DINO的mAP50分别为77.3%和77.4%。这说明引入文本语义信息后,模型能够利用类别名称对目标区域进行更有效的语义引导,从而提升电力杆塔目标的识别能力。

另外,专门针对遥感小目标检测设计的

SKYDET取得了较强的检测性能,其DT和TT类别AP分别为73.2%和85.8%,mAP50和mAP50-95分别达到79.5%和40.1%,明显优于多数通用目标检测模型,说明针对遥感小目标进行专门的多尺度特征建模具有重要作用。值得注意的是,SKYDET在输电塔类别上取得了85.8%的最高AP,但其DT检测精度为73.2%。相比之下,SA-DINO的DT检测精度达到77.5%,较SKYDET提高了4.3%,同时mAP50和mAP50-95较SKYDET进一步提高了1.4%和2.1%。这表明,本文方法在保持不同尺度目标综合检测性能的同时,对小尺度配电塔具有更好的适应能力。

与所有对比方法相比,本文提出的SA-DINO取得了最高的mAP50和mAP50-95,分别达到了80.9%和42.2%。其中,DT和TT的AP分别为77.5%和84.4%。尽管SKYDET在TT类别以及模型复杂度方面具有一定优势,但SA-DINO在DT,mAP50和mAP50-95 3个指标上表现更优。相较于基线模型MM-Grounding DINO,本文方法在mAP50和mAP50-95上分别提升了3.5%和1.7%。该结果表明,本文提出的三个改进模块能够增强Grounding DINO在电力杆塔遥感图像中的检测能力。此外,从参数量可以看出,本文方法的参数量为235.2 M,仅比MM-Grounding DINO增加2.2 M,但mAP50提

升了 3.5%, mAP50-95 提升了 1.7%。这说明本文方法在只引入少量额外参数的情况下,能够显著提升模型的检测精度。虽然 SKYDET 的参数量和计算量分别为 127.1 M 和 351.7 G, 低于 SA-DINO 的 235.2 M 和 450.2 G, 表明 SKY-DET 在模型轻量化方面具有一定优势。SA-DINO 由于建立在 Grounding DINO 视觉-语言框架之上, 整体计算量相对较高, 但其在 DT, mAP50 及更严格的 mAP50-95 指标上取得了更高的检测精度。因此, 本文方法主要体现为在现有 Grounding DINO 框架下以较小的额外参数开销提升多尺度电力塔尤其是小目标的检测性能, 而模型进一步轻量化仍是后续研究方向。综合来

看, SA-DINO 充分利用了 Grounding DINO 的视觉语言建模能力, 并进一步通过尺度感知特征增强、query 选择和定位监督提高了模型对复杂遥感场景中电力杆塔的适应性, 验证了本文方法的有效性。

为直观验证本文方法的检测效果, 选取不同复杂场景下的卫星遥感图像进行可视化对比, 如图 5 所示, 其中第一行为原始图像, 第二行为基线模型 MM-Grounding DINO 的检测结果, 第三行为本文改进模型 SA-DINO 的检测结果。模型漏检处在图中已用红色圆圈标出(彩图见期刊电子版)。

为进一步直观比较不同检测方法在复杂遥



图 5 其他检测模型与本文提出方法的检测可视化结果

Fig. 5 Detection visualization results for other detection models and proposed method in this paper

感场景下的检测效果,选取具有代表性的测试图像进行可视化分析,如图 5 所示。图中从上至下依次为原始图像、基线模型 MM-Grounding DINO, YOLOv13, SKYDET 以及本文提出的 SA-DINO 的检测结果。所选图像涵盖道路、居民区、荒漠以及阴影干扰等不同背景,并包含不同尺度的电力塔目标,以比较各方法在复杂背景、小目标识别以及边界框定位方面的差异。

从可视化结果中可以看出,不同模型在复杂遥感场景下表现出一定差异。MM-Grounding DINO 和 YOLOv13 虽然能够检测出多数尺度较大、结构特征较为显著的电力塔目标,但在道路交叉区域、建筑物密集区域、荒漠背景以及阴影干扰较强的场景中,仍容易受到周围地物纹理和背景结构的影响。当目标尺度较小、有效像素数量有限或目标外观与背景相似时,两种模型均出现不同程度的漏检现象。基线模型甚至误将阴影识别为杆塔目标。相比之下,SKYDET 针对遥感小目标进行了专门优化,对部分小尺度目标表现出较好的检测能力,但在部分复杂场景中仍存在冗余检测或者错检漏检的情况。

相比之下,SA-DINO 在多个场景中表现出更稳定的检测能力。对于道路旁、居民区附近以及沙漠背景中的小型配电塔目标,本文方法能够

更准确地给出检测框,并有效减少基线模型中的漏检和错检现象。另外,从图中也可以看出 SA-DINO 也避免了将杆塔阴影错检为电力杆塔的情况。在一些目标尺度较小且周围存在阴影或复杂纹理干扰的图像中,SA-DINO 仍能够保持较好的目标响应,说明小目标细节保持分支增强了高分辨率特征中的局部边缘和纹理信息,使小型配电塔在后续编码和解码过程中具有更清晰的视觉表征。此外,从检测框的定位效果来看,SA-DINO 的预测框整体更加贴合目标区域,尤其是在目标尺度较小或背景干扰较强的场景中,其边界框偏移现象相较基线模型有所减轻。综合可视化结果可以看出,本文方法不仅提升了模型对弱纹理小型配电塔的召回能力,也改善了复杂背景下的定位稳定性。与 MM-Grounding DINO 相比,SA-DINO 在漏检抑制、小目标识别和边界框定位方面均表现出更好的效果,进一步验证了改进后模型的有效性。

4.5 消融实验

为验证各改进模块对模型检测性能的影响,本文分别对本文提出的 3 个改进模块进行了消融实验,其结果如表 3 所示。消融实验中的参数设置与之前实验的设置相同,表 3 中“√”代表模型中加入了该模块,“×”代表模型中没有加入该模块。

表 3 消融实验结果
Tab. 3 Ablation experiment results

DPSF	SA-LGQS	SA-LM	DT/%	TT/%	mAP50/%	Params/M	FLOPs/G	FPS
×	×	×	76.0	78.7	77.4	233.0	428.2	23.4
√	×	×	77.2	80.1	78.7	235.2	450.2	22.4
×	√	×	78.4	79.1	78.7	233.0	428.2	22.2
×	×	√	77.3	79.7	78.5	233.0	428.2	23.4
√	√	×	77.4	82.1	79.7	235.2	450.2	21.8
√	×	√	78.1	80.8	79.5	235.2	450.2	22.4
×	√	√	78.9	79.4	79.1	233.0	428.2	22.2
√	√	√	77.5	84.4	80.9	235.2	450.2	21.8

由表中结果可知,基线模型的 DT, TT 和 mAP50 分别为 76.0%, 78.7% 和 77.4%。单独引入 DPSF, SA-LGQS 和 SA-LM 后, mAP50 分别提升至 78.7%, 78.7% 和 78.5%, 说明 3 种模块都能够改善检测性能, 其中 SA-LGQS 将 DT 提升至 78.4%, 对小型配电塔的改善较为明显。

不同模块组合进一步体现出互补性, DPSF 与 SA-LGQS, DPSF 与 SA-LM 以及 SA-LGQS 与 SA-LM 组合时, mAP50 分别达到 79.7%, 79.5% 和 79.1%, 其中 SA-LGQS 与 SA-LM 组合取得最高的 DT 精度 78.9%。当 3 个模块同时引入时, 模型获得最佳总体性能, DT, TT 和

mAP50 分别达到 77.5%, 84.4% 和 80.9%, 较基线提升了 1.5%, 5.7% 和 3.5%, 表明完整模型能够在不同尺度电力杆塔之间取得更好的性能平衡。

复杂度方面, 新增计算开销主要来自 DPSF, 使参数量由 233.0 M 增加至 235.2 M, FLOPs 由 428.2 GFLOPs 增加至 450.2 GFLOPs。SA-LGQS 不增加可学习参数和理论 FLOPs, 但由于增加了 Query 筛选操作, 推理速度略有下降。SA-LM 仅作用于训练阶段, 不增加推理阶段的参数量和计算量。最终完整模型 FPS 为 21.8, 相比基线有所下降, 但在提升检测精度的同时整体计算开销仍处于可接受范围内, 同时具备一定

的在线图像处理能力。

在相同测试条件下, MM-Grounding DINO 和 SA-DINO 的显存占用分别为 15.5 GB 和 18.6 GB。虽然改进模型在复杂度、推理速度和显存占用方面产生了一定额外开销, 但其主要来源于 DPSF 对高分辨率特征的增强, 而 SA-LGQS 和 SA-LM 对模型复杂度影响较小。考虑到本文并未针对边缘设备进行专门的轻量化设计, 当前模型在高帧率实时部署方面仍存在一定局限, 后续将进一步研究模型压缩和高效推理方法。

本文通过 Eigen-CAM 分析了改进模型方法的可解释性。图 6 展示了基线模型和 SA-DINO 在 SRSPTD 数据集上的热力图可视化结果。

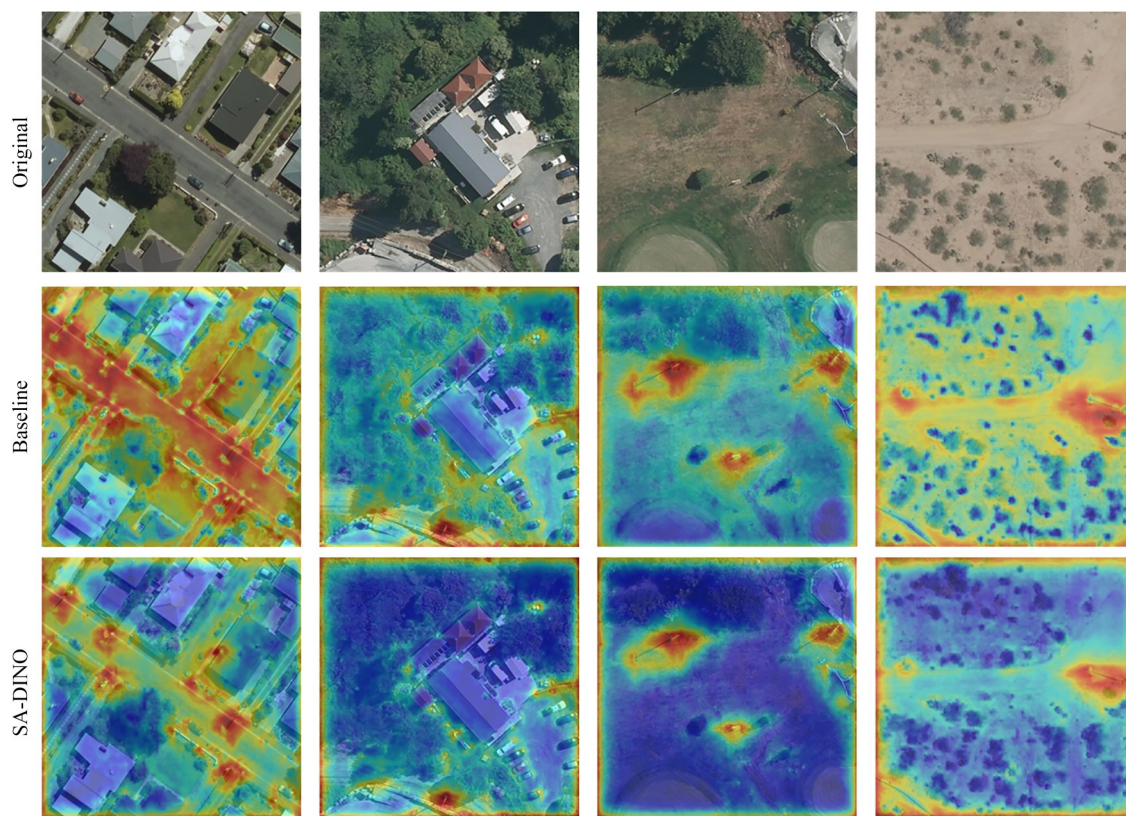


图 6 基线模型与 SA-DINO 的热力图可视化对比

Fig. 6 Visual comparison of heatmaps between baseline and SA-DINO

从图中可以看出, 基线模型 MM-Grounding DINO 虽然能够对图像中的部分目标区域产生响应, 但其注意力分布相对分散, 容易受到道路、建筑物边缘、沙漠和植被背景等无关区域的干扰, 说明原始模型在复杂遥感场景中对电力杆塔细粒度特征的关注仍不够稳定。相比之下, 本文改

进后的 SA-DINO 在不同场景下均表现出更加集中和清晰的激活区域, 热力响应能够更准确地覆盖潜在杆塔目标及其邻近结构, 背景区域的无效响应明显减少。尤其是在城市、道路交叉区域和荒漠背景中, SA-DINO 对小尺度目标区域的关注更强, 表明小目标细节保持分支增强了高分辨

率特征中的局部纹理和边缘信息,尺度感知语言引导 Query 选择模块使模型能够更加关注与文本语义相关的目标候选区域。整体来看,与基线模型相比,SA-DINO 的特征关注区域更加符合电力杆塔的真实空间分布,说明本文方法能够有效提升模型在复杂背景下的小目标感知能力和检测可靠性。

5 结 论

本文针对高分辨率卫星遥感图像中电力杆塔尺度差异显著、小型配电塔特征弱和定位困难等问题,提出了一种基于改进 Grounding DINO 的尺度感知电力杆塔检测方法 SA-DINO。该方法从特征增强、Query 选择和定位监督 3 个方面进行优化。通过小目标细节保持分支增强高分辨率特征中的边缘、纹理和局部结构信息,缓解小目标特征在深层传递中的衰减;通过尺度感知语言引导 Query 选择模块为高分辨率特征层分配更多查询,使小型配电塔候选区域获得更多进入解码器的机会;同时引入尺度感知小目标定位损失,利用归一化 Wasserstein 距离和尺度自适应权重提升小目标边界框回归的稳定性。实验结果表明,SA-DINO 在 SRSPTD 数据集上相比基

线模型 MM-Grounding DINO 的 mAP50 和 mAP50-95 分别提升了 3.5% 和 1.7%,且仅增加少量参数,说明本文方法能够有效提升复杂遥感场景下不同尺度电力杆塔的检测精度,尤其改善了小型配电塔的漏检和定位偏移问题。

尽管本文方法取得了一定提升,但仍存在一些不足。由于 Grounding DINO 本身参数量较大,改进后的 SA-DINO 在实际大范围遥感图像快速部署中仍面临一定计算压力。本文实验主要基于 SRSPTD 数据集展开,模型在更多地区、更多成像条件和多源遥感数据下的泛化能力仍需进一步验证。未来研究将从模型轻量化、跨区域泛化、多源遥感数据融合以及复杂场景下更精细的杆塔检测等方面继续改进,以进一步提升卫星遥感电力基础设施智能巡检的应用能力。

作者贡献声明:

石超君:论文构思和审核,实验指导;
曹煜捷:实验设计与执行,初稿撰写;
李 喆:数据管理与验证;
张 珂:提供资源,实验指导;
王晓阳:实验数据整理,调查研究;
霍梦成:实验数据整理,可视化。

参考文献:

- [1] LIU J J, LIU C Y. Review of deep learning-based intelligent inspection research for transmission lines [J]. *Computers, Materials & Continua*, 2026, 87 (2): 1-10.
- [2] HAROUN F M E, DEROS S N M, DIN N M. Detection and monitoring of power line corridor from satellite imagery using RetinaNet and K-mean clustering [J]. *IEEE Access*, 2021, 9: 116720-116730.
- [3] HU B, SHAN C Y. Integration of InSAR and deterministic coefficient methods to study the disaster susceptibility of transmission tower slope [J]. *Geomatics, Natural Hazards and Risk*, 2025, 16 (1): 2458010.
- [4] GUI S X, SONG S, QIN R J, *et al.* Remote sensing object detection in the deep learning era: a review [J]. *Remote Sensing*, 2024, 16 (2): 327.
- [5] 梁海翔,唐艳慧,王宇庆,等. 高分辨率遥感图像的目标检测[J]. *液晶与显示*, 2024, 39 (10): 1350-1360.
- [6] LIANG H X, TANG Y H, WANG Y Q, *et al.* Object detection methods for high resolution remote sensing images [J]. *Chinese Journal of Liquid Crystals and Displays*, 2024, 39 (10): 1350-1360. (in Chinese)
- [7] ZHU H L, JIANG C L, ZHU X L. LMGANet: a multi-scale guided aggregation network for small-object detection in urban remote sensing [J]. *Remote Sensing*, 2026, 18 (10): 1578.
- [8] 杨潞霞,刘泽凯,张红瑞,等. 基于上下文协同感知的航拍小目标检测算法[J]. *液晶与显示*, 2025, 40 (9): 1308-1320.
- [9] YANG L X, LIU Z K, ZHANG H R, *et al.* Aerial small target detection algorithm based on context collaborative perception [J]. *Chinese Journal of Liquid Crystals and Displays*, 2025, 40 (9): 1308-1320. (in Chinese)
- [10] CARION N, MASSA F, SYNNAEVE G, *et al.*

- End-to-end object detection with transformers [C]. *Computer Vision-ECCV* 2020. Cham: Springer, 2020: 213-229.
- [9] ZHANG H, LI F, LIU S L, *et al.* DINO: DETR with improved DeNoising anchor boxes for end-to-end object detection [EB/OL]. 2022: *arXiv*: 2203.03605. <https://arxiv.org/abs/2203.03605>.
- [10] LIU S L, ZENG Z Y, REN T H, *et al.* Grounding DINO: marrying DINO with Grounded pre-training for Open-set object detection [C]. *Computer Vision-ECCV* 2024. Cham: Springer, 2025: 38-55.
- [11] XU C, WANG J W, YANG W, *et al.* Detecting tiny objects in aerial images: a normalized Wasserstein distance and a new benchmark [J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2022, 190: 79-93.
- [12] REZATOFIGHI H, TSOI N, GWAK J, *et al.* Generalized intersection over union: a metric and a loss for bounding box regression [C]. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 15-20, 2019, Long Beach, CA, USA. IEEE, 2020: 658-666.
- [13] 李智杰, 惠爱婷, 李昌华, 等. 面向遥感图像道路提取的多尺度上下文感知网络 [J]. *光学精密工程*, 2025, 33(4): 610-623.
- LI ZH J, HUI A T, LI CH H, *et al.* Multi-scale context-aware network for road extraction in remote sensing images [J]. *Optics and Precision Engineering*, 2025, 33(4): 610-623. (in Chinese)
- [14] REN S Q, HE K M, GIRSHICK R, *et al.* Faster R-CNN: towards real-time object detection with region proposal networks [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, 39(6): 1137-1149.
- [15] WANG C Y, BOCHKOVSKIY A, LIAO H M. YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors [C]. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 17-24, 2023, Vancouver, BC, Canada. IEEE, 2023: 7464-7475.
- [16] TIAN Y J, YE Q X, DOERMANN D. YOLOv12: attention-centric real-time object detectors [EB/OL]. 2025: *arXiv*: 2502.12524. <https://arxiv.org/abs/2502.12524>.
- [17] LU Q, LIN Y J, YIN L F. Quantum-enhanced deep lightweight object detection model for power tower base occlusion detection [J]. *Engineering Applications of Artificial Intelligence*, 2025, 160: 111972.
- [18] HUANG S H, LIN C, PENG J S, *et al.* DLR-YOLO: Dynamic low-rank training for a lightweight power tower object detection network in multi-scenario remote sensing images [J]. *Advanced Engineering Informatics*, 2026, 72: 104463.
- [19] CHENG T H, SONG L, GE Y X, *et al.* YOLO-world: real-time open-vocabulary object detection [C]. 2024 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 16-22, 2024, Seattle, WA, USA. IEEE, 2024: 16901-16911.
- [20] CHENG G, YUAN X, YAO X W, *et al.* Towards large-scale small object detection: survey and benchmarks [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, 45(11): 13467-13488.
- [21] ZHA W T, HU L W, SUN Y L, *et al.* ENGD-BiFPN: a remote sensing object detection model based on grouped deformable convolution for power transmission towers [J]. *Multimedia Tools and Applications*, 2023, 82(29): 45585-45604.
- [22] FU K, CHANG Z H, ZHANG Y, *et al.* Rotation-aware and multi-scale convolutional neural network for object detection in remote sensing images [J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2020, 161: 294-308.
- [23] SHI C J, ZHENG X, ZHAO Z B, *et al.* LSKF-YOLO: large selective kernel feature fusion network for power tower detection in high-resolution satellite remote sensing images [J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2024, 62: 5620116.
- [24] SHI H W, YANG W Z, CHEN D N, *et al.* ASG-YOLOv5: Improved YOLOv5 unmanned aerial vehicle remote sensing aerial images scenario for small object detection based on attention and spatial gating [J]. *PLoS One*, 2024, 19(6): e0298698.
- [25] XIAO X, XUE X R, ZHAO Z Y, *et al.* A recursive prediction-based feature enhancement for small object detection [J]. *Sensors*, 2024, 24(12): 3856.
- [26] ZHANG H X, OU Z H, YAO S Y, *et al.* HS-

- FANet: hierarchical scale-sensitive feature aggregation network for small object detection in UAV aerial images[J]. *Drones*, 2025, 9(9): 659.
- [27] ZHU X Z, SU W J, LU L W, *et al.* Deformable DETR: deformable transformers for end-to-end object detection [EB/OL]. 2020: *arXiv*: 2010.04159. <https://arxiv.org/abs/2010.04159>.
- [28] WANG D, ZHANG Q M, XU Y F, *et al.* Advancing plain vision transformer toward remote sensing foundation model[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2023, 61: 5607315.
- [29] LI L H, ZHANG P C, ZHANG H T, *et al.* Grounded language-image pre-training [C]. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 18-24, 2022, New Orleans, LA, USA. IEEE, 2022: 10955-10965.
- [30] MINDERER M, GRITSENKO A, STONE A, *et al.* Simple open-vocabulary object detection [C]. *Computer Vision-ECCV 2022*. Cham: Springer Nature Switzerland, 2022: 728-755.
- [31] ZANG Y H, LI W, ZHOU K Y, *et al.* Open-vocabulary DETR with Conditional matching [C]. *Computer Vision-ECCV 2022*. Cham: Springer, 2022: 106-122.
- [32] HUANG X, CHEN Y B, TANG C, *et al.* Grounding DINO with text-guided deformable attention mechanism for power transmission inspection [C]. *The Proceedings of the International Council on Electrical Engineering Conference 2025 (ICEE 2025)*. Singapore: Springer, 2026: 539-551.
- [33] HU Z B, GAO K, ZHANG X D, *et al.* Efficient grounding DINO: efficient cross-modality fusion and efficient label assignment for visual grounding in remote sensing[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2025, 63: 5609414.
- [34] ZHAO X Y, CHEN Y C, XU S L, *et al.* An open and comprehensive pipeline for unified object grounding and detection [EB/OL]. 2024: *arXiv*: 2401.02361. <https://arxiv.org/abs/2401.02361>.
- [35] ZHANG Y, GUO W, XIE B X, *et al.* SKY-DET: an end-to-end multiscale attentive detection network from foundation models for small objects in remote sensing images[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2026, 64: 5634117.

作者简介:



石超君(1992—),男,河北保定人,博士,副教授,2021年于河北工业大学获得博士学位,主要从事计算机视觉等方面的研究。E-mail:scj@ncepu.edu.cn

通讯作者:



张珂(1980—),男,河北保定人,博士,教授,2012年于北京邮电大学获得博士学位,主要从事电力视觉等方面的研究。E-mail:zhangkeit@ncepu.edu.cn