基于多尺度特征融合和 SAM 引导的无人机小尺度目标检测*

钟嘉宇[1]，牛利玲[2]，任　超[1]
（1. 四川大学　电子信息学院，四川　成都　610065；
2. 四川航天电子设备研究所，四川　成都　610100）

摘　要：在无人机航拍中，因拍摄距离远、目标占比低，其线性尺度仅有十余像素且特征匮乏，导致检测性能显著下降。现有方法主要分为样本增强与多尺度感知，前者在航拍目标密集场景中易引入语义冲突，而后者在深层特征感知与全局建模上仍存在不足。为此，提出一种基于多尺度特征融合和 SAM 引导的小目标检测网络，通过设计包含小目标检测层的多尺度架构增强特征表达能力，融合空洞卷积与 Transformer 以扩大感受野并建模长程依赖，并引入 SAM 大模型的先验知识引导网络训练，从而提升对小目标特征的提取能力。实验表明，该方法在 VisDrone-DET2019 上显著提升了小目标检测精度。

关键词：目标检测；特征提取；深度学习
中图分类号：TP391. 4　　**文献标志码：**A　　**DOI：**10. 19358/j. issn. 2097 - 1788. 2026. 03. 004

中文引用格式：钟嘉宇，牛利玲，任超. 基于多尺度特征融合和 SAM 引导的无人机小尺度目标检测［J］. 网络安全与数据治理，2026，45(3)：24 - 32.
英文引用格式：Zhong Jiayu，Niu Liling，Ren Chao. UAV small-scale object detection based on multi-scale feature fusion and SAM guidance［J］. Cyber Security and Data Governance，2026，45(3)：24 - 32.

UAV small-scale object detection based on multi-scale feature fusion and SAM guidance

Zhong Jiayu[1]，Niu Liling[2]，Ren Chao[1]
（1. College of Electronics and Information Engineering，Sichuan University，Chengdu 610065，China；
2. Sichuan Aerospace Electronic Equipment Research Institute，Chengdu 610100，China）

Abstract：In UAV aerial images，the target objects to be detected are often only dozens of pixels in size due to long shooting distances and low target occupancy ratios，resulting in severe feature scarcity and a significant degradation in small object detection performance. Existing approaches primarily fall into two categories：sample augmentation and multi-scale perception. The former tends to introduce semantic conflicts in dense aerial scenarios，while the latter remains inadequate in deep feature perception and global modeling. To address these limitations，this paper proposes a small object detection network based on multi-scale feature fusion and SAM-guided learning. Specifically，we design a multi-scale architecture incorporating dedicated detection layers for small objects to enhance feature representation；integrate dilated convolutions with Transformers to enlarge the receptive field and model long-range dependencies；and leverage the prior knowledge of the Segment Anything Model（SAM）foundation model to guide network training，thereby improving the extraction of discriminative features for small objects. Experimental results demonstrate that our method significantly improves small object detection accuracy on the VisDrone-DET2019 benchmark.
Key words：object detection；feature extraction；deep learning

0　引言

近年来，深度学习技术的快速发展显著推动了计算机视觉与智能感知领域的进步，为遥感图像的自动化理解提供了强大的方法支撑。在此背景下，结合飞行控制与高分辨率成像技术的持续突破，无人机在遥感监测中的应用广度与深度不断拓展。依托其广域覆盖、灵活机动与近地观测等优势，无人机系统已广泛服务于农作物长势动态监测[1]、车辆检测与轨迹分析[2]以及灾情快速勘察[3]等多样化场景，将逐步发展

*基金项目：国家自然科学基金（62171304）；四川大学能力提升计划基金（2024SCUQJTX025）

为多行业实现高效、实时区域感知的核心技术手段。然而，在实际应用中，尤其是在执行低空飞行任务时，无人机所获取的图像常面临显著的目标检测挑战：待识别目标（如行人、车辆等）在图像中通常呈现为小尺度目标，其边界框尺度往往仅为数十像素，导致目标特征信息不足、信噪比较低，显著增加了检测与识别的难度。针对这一挑战，研究者们从数据与模型两个核心维度展开了系统性探索，主要形成了两大技术路径：样本导向的方法与多尺度感知的方法。

基于小目标在图像中占比小、与锚点重叠度低的问题，研究者们探索样本导向的方法，即通过人工合成或变换，在现有图像中增加小目标实例，解决小目标稀缺问题。RRNet[4]引入自适应重采样数据增强策略，利用先验分割图来引导小目标粘贴位置。DS-GAN[5]提出一种基于生成对抗网络的小目标检测数据增强方法，其通过大目标生成高质量合成小目标，并结合分割等方法合理选择位置粘贴。尽管上述样本导向的数据增强方法在小目标稀疏的训练数据中表现良好，但当原始训练集本身已包含密集分布的小目标时，其基于复制粘贴的增强机制易加剧样本中的目标重叠、尺度失配与背景语义冲突，不仅难以提升数据多样性，反而可能引入分布偏移与伪影干扰，导致模型训练不稳定。

同时，目标尺度的显著差异普遍存在：同一图像中常同时出现远距离的小尺度目标与近距离的大尺度目标。这一特性对检测模型的尺度适应性提出了挑战，促使多尺度感知能力成为现代目标检测系统的核心设计要素。其中，FPN[6]首次系统性地将高层语义特征通过上采样与低层高分辨率特征融合，在保持定位精度的同时增强特征语义信息，从而显著缓解尺度变化带来的性能下降问题，成为特征提取器的核心组件。随后，一大批优秀的多尺度特征融合方法涌现出来。PA-Net[7]在 FPN 的自顶向下路径基础上增加了自底向上的路径，实现双向跨尺度特征融合。Bi-FPN[8]在 PANet 双向特征融合基础上，通过精简冗余连接和引入可学习的加权融合机制，实现了更低计算开销的特征融合。SSPNet[9]通过上下文注意力模块、尺度增强模块和尺度选择模块协同优化多尺度特征利用，并结合加权负采样策略，显著提升了微小行人检测性能。SCRDet[10]通过采样融合网络提升多尺度小目标感知能力，结合监督像素注意力与通道注意力机制抑制背景干扰、强化目标特征。FFCA-YOLO[11]通过特征增强、多尺度融合与空间上下文感知三大模块，在显著提升遥感小目标检测精度与鲁棒性的同时兼顾实时性。上述研究表明，合理融合多层特征，可有效缓解尺度变化带来的性能下降。尽管如此，如何在多尺度分布下实现较好的检测性能，仍需要进一步探索。

近年来，以 SAM[12]、DINO[13]为代表的视觉大模型，因其在海量数据上训练获得的强大通用视觉表征与零样本泛化能力，为解决上述问题提供了新的思路。其内在的丰富语义信息，可为提升目标检测，尤其是小目标的检测性能，提供宝贵的先验知识。为了进一步提升模型对小尺度目标的检测能力，更有效地应对航拍遥感图像中因成像距离远、目标占比低导致的特征弱化与易漏检问题，本文提出了一种基于多尺度特征融合与 SAM 视觉特征引导的小目标检测网络（MSG-YOLO），其主要贡献如下：

（1）针对小目标检测中多层次特征融合不足的问题，提出基于特征金字塔的多尺度检测架构，通过引入小目标检测层，强化多尺度特征表达能力。

（2）针对深层特征感受野受限和全局上下文建模不足的问题，融合空洞融合模块（Dilated Fuse Block，DFB）与 Transformer 模块（Transformer Block，TFB），通过多分支空洞卷积扩大感受野，并利用自注意力机制捕获长程依赖，提升复杂场景下小目标的表征能力。

（3）针对网络对小目标特征提取不稳健的问题，引入 SAM 视觉大模型的视觉先验知识作为引导信息，指导检测网络的训练过程。

1 基于多尺度特征融合与 SAM 增强的目标检测网络

1.1 整体结构

目标检测模型主要由特征提取主干网络、多尺度特征融合网络和检测头三部分构成，整体结构如图 1 所示。小尺度目标在图像中占比小，经深度卷积网络逐层处理后易因下采样导致细节信息丢失。因此，需要提取浅层至深层的多尺度特征，其中高分辨率的浅层特征对保留目标细节尤为关键。同时，扩大深层特征感受野，引入 CBAM 注意力模块[14]和 DFB 模块以增强上下文建模能力，从而更有效地捕获小目标的语义信息。此外，考虑到特征提取网络的浅层特征虽富含细节，但噪声较强、干扰较多，故利用 SAM 模型的视觉先验来引导特征提取过程，获得更加稳健的浅层特征和深层特征。为了进一步提升特征的稳健性，使用自顶向下和自底向上的多尺度融合过程，并引入 TFB 模块增强长程建模能力。最后，不同层级的特征输入各层级的检测头，得到目标检测结果。

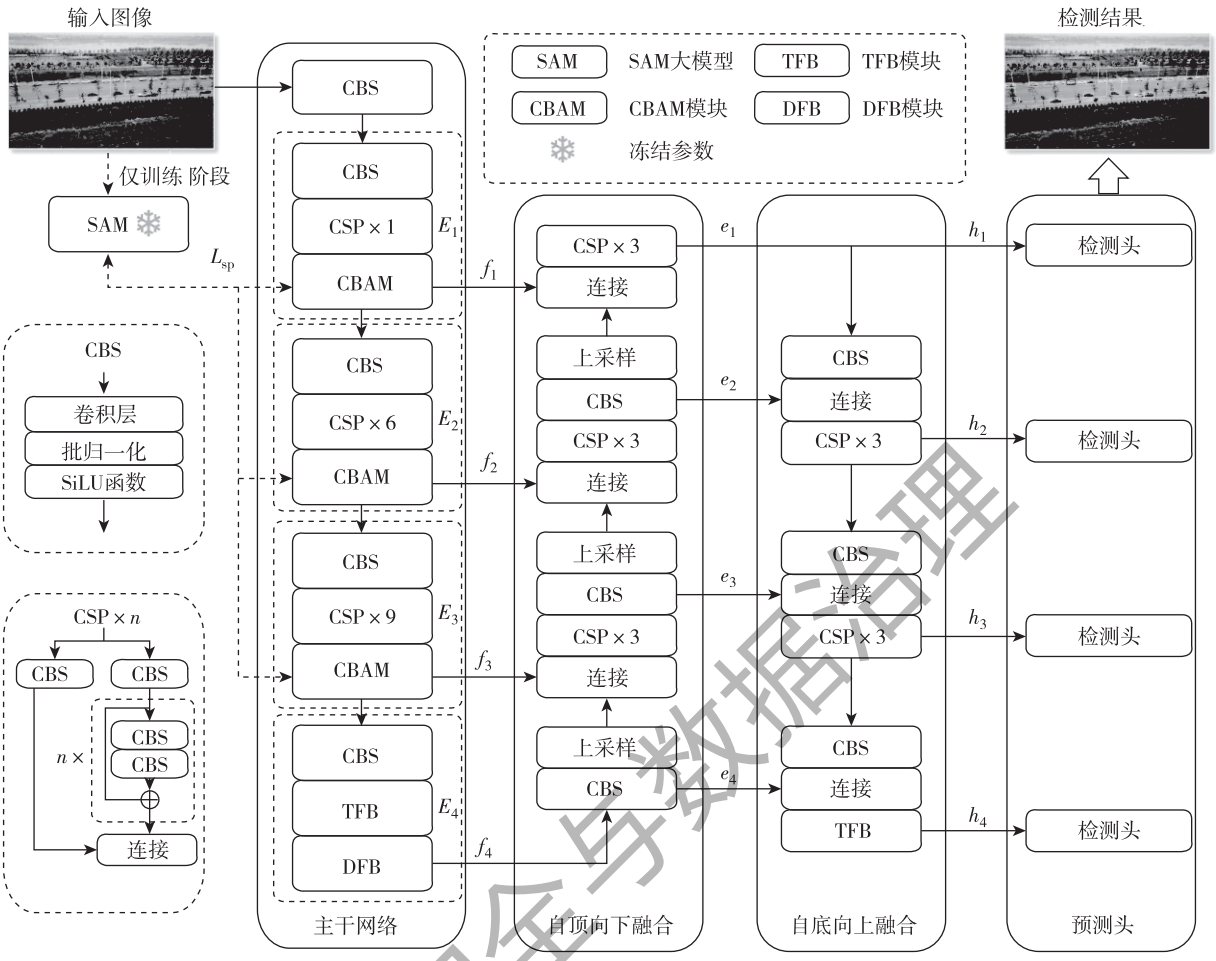


图 1　MSG-YOLO 的整体网络结构图

1.2　特征提取主干网络

为了有效提取图像的特征，采用业内常用的 CBS（Conv-BatchNorm-SiLU）模块作为网络基础模块，并通过残差连接组成 CSP 模块，其具体组成如图 1 所示。将整个特征提取网络分为四个层级，分别命名为 E_1，E_2，E_3，E_4。在每一层的最后，引入了 CBAM 注意力模块对每层特征进一步细化。对于输入图像 $I \in \mathbf{R}^{H\times W\times C}$，首先通过一层单独的 CBS 得到特征 $f_0 \in \mathbf{R}^{H\times W\times C}$：

$$f_0 = \mathrm{SiLU}(\mathrm{BatchNorm}(\mathrm{Conv}(I))) \tag{1}$$

随后，将其输入 E_1，得到浅层特征 f_1，其运算过程如下：

$$f_1 = \mathrm{CBAM}(\mathrm{CSP}_1(\mathrm{CBS}(f_0))) \tag{2}$$

其中，CSP_n 为 CBS 构成的卷积网络模块，其内部构成已在图 1 中展示。CBAM 注意力模块由通道注意力和空间注意力两部分构成，通过串联上述两个注意力模块，自适应地细化输入特征，对于输入特征 $x \in \mathbf{R}^{H\times W\times C}$，其核心计算公式如下[14]：

$$x^n = \mathrm{CBAM}(x) = M_s \odot (M_c \odot x) \tag{3}$$

$$M_c(x) = \sigma(\mathrm{MLP}(A(x)) + \mathrm{MLP}(G(x))) \tag{4}$$

$$M_s(x) = \sigma(\mathrm{Conv}^{7\times 7}(\mathrm{Concat}(A(x), G(x)))) \tag{5}$$

其中，M_c 表示通道注意力，M_s 表示空间注意力，MLP 表示多层感知机，$\odot$ 表示逐元素矩阵乘法，A 表示全局平均池化，G 表示全局最大池化，σ 表示 Sigmoid 激活函数，Concat 表示特征连接。类似地，将第一层输出 f_1 送入第二级模块 E_2，得到第二层输出；再将第二层输出 f_2 送入第三级模块 E_3，得到第三层输出 f_3：

$$f_2 = \mathrm{CBAM}(\mathrm{CSP}_6(\mathrm{CBS}(f_1))) \tag{6}$$

$$f_3 = \mathrm{CBAM}(\mathrm{CSP}_9(\mathrm{CBS}(f_2))) \tag{7}$$

为适配小目标检测任务对定位精度的更高要求，特征提取网络的最后一层 E_4 摒弃了传统的 CSP 结构。首先，仍将前序模块得到的 f_3 输入 CBS 模块进行下采样，得到 f_{3c}。如图 2 所示，为了提升上下文感知能力，将其输入由多头注意力构成的 TFB 模块中，以突破局

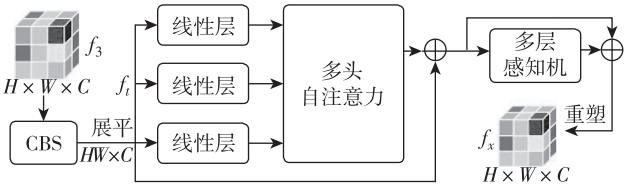


图 2　TFB 模块的网络结构图

部感受野限制、动态建模长程依赖。对于输入特征 $f_{3c}\in\mathbf{R}^{H\times W\times C}$，将其展平后输入线性层生成查询、键和值，计算公式如下：

$$f_t=\mathrm{Flat}\ (f_{3c})\in\mathbf{R}^{HW\times C} \tag{8}$$

$$f_q^i=\boldsymbol{W}_q^i f_t,\ f_k^i=\boldsymbol{W}_k^i f_t,\ f_v^i=\boldsymbol{W}_v^i f_t \tag{9}$$

其中，$\boldsymbol{W}_q^i$，$\boldsymbol{W}_k^i$，$\boldsymbol{W}_v^i$ 为第 i 个头线性层的投影矩阵。之后，将获得的特征输入多头自注意力，以实现每个位置的注意力权重随输入特征自适应调整：

$$f_{\mathrm{ma}}=\mathrm{Concat}_{i=1}^{h}\left[\mathrm{softmax}\left(\frac{f_q^i f_k^{i\mathrm{T}}}{\sqrt{d_k}}\right)f_v^i\right]+f_t \tag{10}$$

其中，d_k 为特征维度，Concat 为沿特征通道连接。上述公式中引入了残差连接，以稳定模型训练。之后，将获得的 f_{ma} 输入多层感知机，将特征维度重塑后得到 TFB 模块的输出 f_x：

$$f_x=\mathrm{Reshape}(\mathrm{MLP}\ (f_{\mathrm{ma}})+f_{\mathrm{ma}}) \tag{11}$$

为进一步扩大有效感知域并进一步捕获上下文信息，本文在该结构基础上加入 DFB 模块，其由多分支空洞卷积构成。在保持特征图尺寸不变的前提下，通过多尺度空洞采样增强残余空间结构建模能力，缓解小目标在深层的特征退化问题。如图 3 所示，DFB 采取了多个分支，除最后一个 CBS 外，其余所有 CBS 保持空间分辨率，仅对特征进行下采样。其计算过程如下：

$$f_{d1}=\mathrm{CBS}(f_x) \tag{12}$$

$$f_{d2}=\mathrm{DCBS}_{d=1}(\mathrm{CBS}(f_x)) \tag{13}$$

$$f_{d3}=\mathrm{DCBS}_{d=3}(\mathrm{CBS}(f_x)) \tag{14}$$

$$f_{d4}=\mathrm{DCBS}_{d=5}(\mathrm{CBS}(f_x)) \tag{15}$$

其中，$\mathrm{DCBS}_{d=n}$ 表示带有空洞卷积的 CBS 模块，n 表示膨胀率。随后，各分支的输入沿特征通道连接起来，经过卷积后得到 E_4 层级的最终输出 f_4。

1.3　基于 SAM 视觉先验引导的特征提取

尽管特征提取网络的浅层特征 f_1 富含高分辨率空间细节，对小目标定位非常重要，但其语义判别性弱、易受背景干扰，导致特征鲁棒性不足，容易对检测产生负面影响。同时，深层特征对小尺度目标的响应仍不够充分，有待进一步增强。因此，需要抑制干扰、

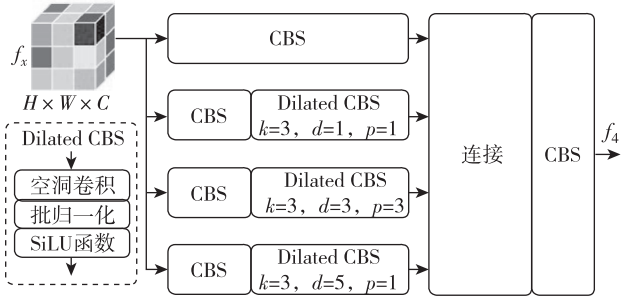


图 3　DFB 模块网络结构图

增强小目标相关特征响应的稳健性。考虑到近期 SAM 等视觉大模型展现出强大的零样本分割能力，故可借鉴 SAM 编码器特征在几何建模上的强鲁棒性，将其作为视觉先验，迫使浅层特征 f_1 在保留高分辨率细节的同时，与 SAM 提取的目标内在结构对齐，从而有效抑制小目标噪声干扰，并提升深层特征的响应能力。对于输入图像 $I\in\mathbf{R}^{H\times W\times C}$，首先使用 SAM 的视觉编码器提取视觉特征得到其稳健的浅层、中层、深层特征 $f_{s1}\in\mathbf{R}^{H\times W\times C}$，$f_{s2}\in\mathbf{R}^{H\times W\times C}$，$f_{s3}\in\mathbf{R}^{H\times W\times C}$。考虑到 f_{s1}，f_{s2}，f_{s3} 和 f_1，f_2，f_3 的空间尺度差距较大，故首先对 f_{s1}，f_{s2}，f_{s3} 进行双线性下采样得到 f_{d1}，f_{d2}，f_{d3} 以适应 f_1，f_2，f_3 的空间尺度。其次，SAM 特征和检测特征的特征尺度也存在间隙，故分层引入 1×1 卷积将 f_1，f_2，f_3 采样到适应的特征维度得到 f_{c1}，f_{c2}，f_{c3}，确保来自不同模型的特征在同等空间维度下进行比较。整个过程如图 4 所示。

为了确保 SAM 视觉先验对特征提取的有效引导，需要仔细挑选特征匹配的度量策略。L1 范数约束是特征对齐中常见的损失函数，其侧重于像素级或特征值的绝对差异[15]。然而，小尺度目标在特征图上的响应区域小、响应强度弱，容易受背景干扰或因下采样而

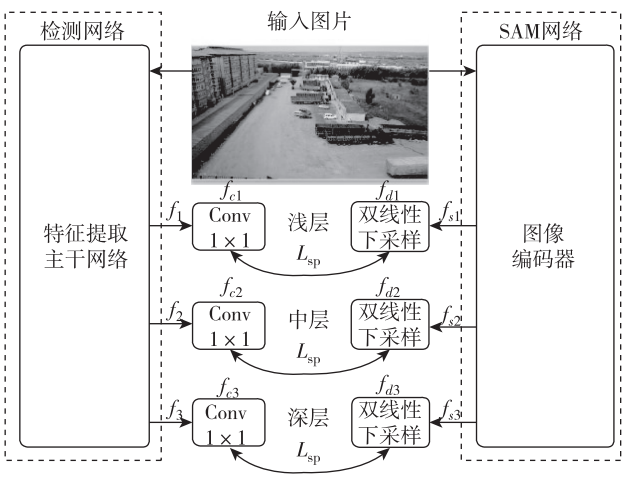


图 4　SAM 视觉先验引导示意图

丢失细节。此时，若直接对齐特征图的 L1 距离，模型容易过度拟合 SAM 视觉先验的绝对响应强度，忽略整体特征结构的一致性。相反，余弦相似度是特征对齐中常用的相似性度量，其侧重于特征方向的一致性，在高维语义空间中已被广泛验证能更有效地衡量语义相似性[16]。故采用余弦相似度损失对特征提取网络的特征进行约束，其公式如下所示：

$$L_{sp}=\sum_{i=1}^{3}\left(1-\frac{f_{ci}^{T}f_{di}}{\|f_{ci}\|_2\|f_{di}\|_2}\right) \tag{16}$$

通过该损失函数，可使 SAM 视觉先验有效引导检测网络的多层特征，从而进一步提升小目标检测性能。

1.4 多尺度特征融合与检测网络

如图 1 所示，多尺度特征融合网络和检测网络参照业内常用做法，即在特征金字塔基础上同时引入自顶向下的语义信息传递路径与自底向上的空间细节增强路径，通过上采样、下采样配合通道连接实现跨尺度特征的双向融合。自顶向下融合公式如下所示：

$$e_4=\mathrm{CBS}(f_4) \tag{17}$$

$$e_3=\mathrm{CSP}_3(\mathrm{Concat}(\mathrm{Up}(\mathrm{CBS}(e_4)),f_3)) \tag{18}$$

$$e_2=\mathrm{CSP}_3(\mathrm{Concat}(\mathrm{Up}(\mathrm{CBS}(e_3)),f_2)) \tag{19}$$

$$e_1=\mathrm{CSP}_3(\mathrm{Concat}(\mathrm{Up}(\mathrm{CBS}(e_2)),f_1)) \tag{20}$$

其中，Up 表示空间上采样，Concat 表示特征连接。接着，对获得的特征进行自底向上的融合过程：

$$h_1=e_1 \tag{21}$$

$$h_2=\mathrm{CSP}_3(\mathrm{Concat}(\mathrm{CBS}(h_1),e_2)) \tag{22}$$

$$h_3=\mathrm{CSP}_3(\mathrm{Concat}(\mathrm{CBS}(h_2),e_3)) \tag{23}$$

$$h_4=\mathrm{TFB}(\mathrm{Concat}(\mathrm{CBS}(h_3),e_4)) \tag{24}$$

得到融合后的特征 h_1，h_2，h_3，h_4 之后，将其输入检测头以输出预测结果。检测网络的损失函数沿用 YOLOv5 的官方设置，记作 L_{yolo}[17]。此外，引入了前述余弦相似度损失作为 SAM 视觉引导约束，故最终损失函数如下：

$$L_{total}=L_{yolo}+\lambda\cdot L_{sp} \tag{25}$$

其中，λ 为 L_{sp}的权重超参数。

2 实验结果与分析

2.1 数据集

为确保不同方法进行公平、客观且具有专业性的比较，本文选用广泛认可的公开无人机航拍基准数据集作为实验数据。这类数据集在采集设备、场景多样性、目标尺度分布及标注规范性等方面均具有代表性，能够全面反映算法在真实无人机视角下的检测性能与鲁棒性。VisDrone-DET2019[18] 是一个面向无人机视角下目标检测任务的大规模图像数据集，全部图像由无人机在不同时间、不同高度及多样化城市与城郊场景采集，共计 8 599 张高分辨率图像。其中训练集 6 471 张、验证集 548 张、测试集 1 580 张。数据集共标注超过 54 万个目标实例，涵盖 10 类常见交通场景对象：行人（pedestrian）、人（person）、小汽车（car）、厢式货车（van）、公交车（bus）、卡车（truck）、摩托车（motor）、自行车（bicycle）、带篷三轮车（awning-tricycle）和三轮车（tricycle）。此外，如图 5 所示，其小尺度目标占比较高，适合作为小目标检测的基准数据集。

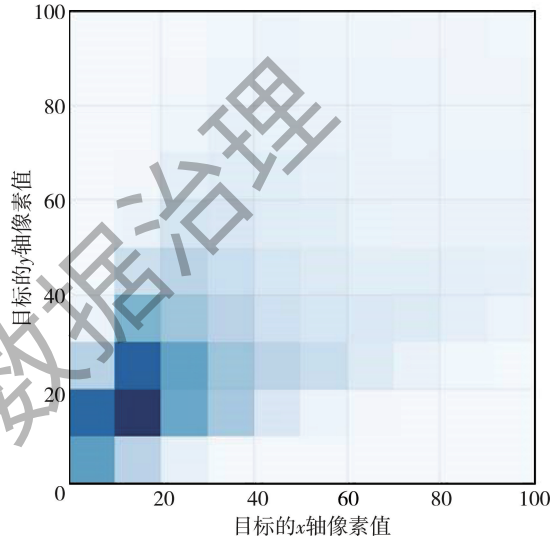


图 5 VisDrone-DET2019 的目标分布

注：区域颜色越深代表此区域范围的目标数量越多

2.2 实验细节

实验中使用的操作系统为 Ubuntu 24.04，实验设备为 NVIDIA RTX 4090 GPU，实验基于 YOLOv5 和 PyTorch[19] 实现。在训练过程中，采用 SGD 优化器[20] 以及余弦退火策略动态调整学习率（初始学习率为 0.01），batch size 设置为 10，超参数 λ 设置为 0.025。在训练过程中，与官方 YOLOv5 数据增强操作相同，启用 Mosaic[21]、随机水平翻转、随机尺度缩放作为数据增强策略。

2.3 实验结果

为全面评估所提出 MSG-YOLO 模型的检测性能，采用平均精度均值（mean Average Precision，mAP）作为核心量化指标，mAP_{50}表示在交并比阈值为 0.5 的条件下计算的各类别 mAP，mAP_{50-95}表示交并比阈值从 0.5 到 0.95（步长 0.05）共 10 个水平下的平均 mAP。同时，受 COCO[22] 评估体系的启发，进一步针对无人机航拍场景中小尺度差异显著的特点，将目标按尺度划分为小目标（像素值小于 32 × 32），并计算其对应

的 mAP，以更细粒度地分析不同模型在小尺度目标上的检测能力。在此基础上，MSG-YOLO 在 VisDrone-DET2019 验证集和测试集上与当前主流方法（YOLOv5，FFCA-YOLO，TPH-YOLO[23]，FBRT-YOLO[24]）进行了对比，其可视化结果如图 6 所示。

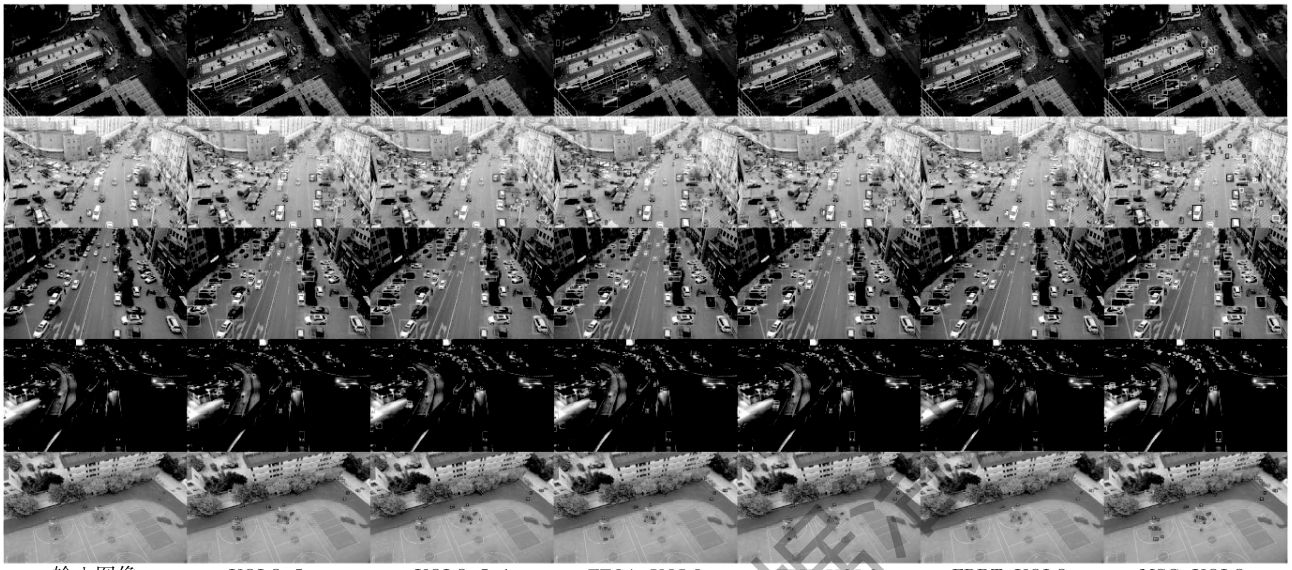


图 6　不同方法的检测结果可视化

2.3.1　VisDrone-DET2019 验证集

如表 1 所示，在 VisDrone-DET2019 验证集上的实验结果表明，MSG-YOLO 在检测精度方面优于所列对比方法。具体而言，该方法在所有目标上的 mAP_{50} 高于 FFCA-YOLO、FBRT-YOLO 和 TPH-YOLO。与此同时，在小目标上，MSG-YOLO 的 mAP_{50} 较 FFCA-YOLO 和 FBRT-YOLO 分别高出 4.0% 和 1.2%，相较 TPH-YOLO 提升更为明显。YOLOv5-n 与 YOLOv5-l 的整体与小目标性能均低于上述小目标检测方法，表明通用检测模型在该场景下仍存在一定局限。上述结果说明，MSG-YOLO 在不引入过多计算复杂度的同时，展现出了更优的检测性能，尤其在小尺度目标上具有相对优势。

表 1　基于 VisDrone-DET2019 验证集的性能比较

方法	所有目标		小目标		性能参数	
	mAP_{50} /%	mAP_{50-95} /%	mAP_{50} /%	mAP_{50-95} /%	GFLOPs	参数量/M
YOLOv5-n	43.4	26.9	24.6	10.9	4.5	1.9
YOLOv5-l	48.9	29.7	38.0	19.2	109.1	46.5
FFCA-YOLO	51.2*	30.9*	41.2*	21.4*	51.4	7.14
TPH-YOLO	51.7	32.9	38.5	20.3	145.7	60.4
FBRT-YOLO	54.7*	34.5*	44.0*	23.7*	119.2	14.6
MSG-YOLO	55.8	35.2	45.2	24.8	122.3	47.3

注：“*”表示基于官方开源代码复现的结果

2.3.2　VisDrone-DET2019 测试集

如表 2 所示，在 VisDrone-DET2019 测试集上的结果进一步表明，MSG-YOLO 在各项检测指标上均优于对比方法。其整体性能高于 FFCA-YOLO 与 TPH-YOLO。在小目标上，MSG-YOLO 的 mAP_{50} 较 FFCA-YOLO 和 FBRT-YOLO 分别提升 3.3% 和 2.4%，优势显著。值得注意的是，各方法在测试集上的指标普遍低于验证集，其中 MSG-YOLO 在测试集所有目标上的表现相对更优，mAP_{50} 相较于 FFCA-YOLO、TPH-YOLO 和 FBRT-YOLO 分别提高了 5.2%、2.7% 和 2.4%，表明 MSG-YOLO 在跨子集评估中具备相对更稳定的性能表现。

表 2　基于 VisDrone-DET2019 测试集的性能比较

方法	所有目标		小目标		性能参数	
	mAP_{50} /%	mAP_{50-95} /%	mAP_{50} /%	mAP_{50-95} /%	GFLOPs	参数量/M
YOLOv5-n	34.8	19.9	17.6	7.2	4.5	1.9
YOLOv5-l	40.5	23.5	27.1	12.5	109.1	46.5
FFCA-YOLO	40.7*	23.2*	28.9*	13.4*	51.4	7.14
TPH-YOLO	43.2	26.0	28.2	13.1	145.7	60.4
FBRT-YOLO	43.5*	26.1*	29.8*	14.4*	119.2	14.6
MSG-YOLO	45.9	27.2	32.2	15.3	122.3	47.3

注：“*”表示基于官方开源代码复现的结果

2.4 消融实验

2.4.1 各组件对检测结果的影响

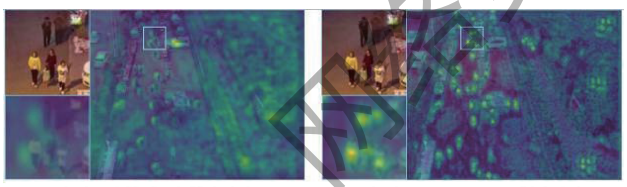
表3展示了逐步引入所提模块对检测性能的影响。消融实验结果表明，各组件均对模型性能产生了积极贡献。在基线模型基础上，引入CBAM、DFB和TFB模块后，模型性能获得首次显著提升，小目标mAP_{50}提升12.4%。这表明上述模块能有效增强模型对多尺度目标的特征表征能力。

在此基础上进一步引入SAM视觉引导损失后，模型性能实现第二次跃升。完整模型在所有目标上达到最优性能。在小目标检测上，mAP_{50}相比上述添加模块后的模型再提升5.7%。该结果表明，SAM引导损失通过引入高层语义监督，有效引导网络关注小目标的关键特征区域，与前述模块形成了良好的协同增强效应。如图7所示，MSG-YOLO相较于基线方法显著增强了对小尺度目标的特征响应，使模型更加注意小尺度目标区域，从而有效提升了小尺度目标的检测精度。

表3 各组件对检测性能的影响比较 （%）

方法	所有目标		小目标	
	mAP_{50}	mAP_{50-95}	mAP_{50}	mAP_{50-95}
基线	48.9	29.7	27.1	12.5
基线+M1	50.2	31.6	39.5	21.1
基线+M1+M2	55.8	35.2	45.2	24.8

注：1）M1指添加了CBAM、DFB和TFB等模块；
2）M2指添加了SAM视觉引导损失



（a）基线方法特征图 （b）MSG-YOLO特征图

图7 基线方法和MSG-YOLO的中间特征图比较

注：颜色越明亮表明模型更关注这个区域

2.4.2 SAM引导系数对检测结果的影响

为进一步探究超参数对SAM视觉引导检测结果的影响，需要对λ进行消融实验。如表4所示，实验结果表明超参数对检测结果存在一定影响。从整体趋势看，当系数为0.025时，模型在所有目标尺度上取得了相对最优的检测性能。系数向两侧变化时，多数评估指标呈现轻微的下降趋势。这表明，损失引导系数的选择存在一个相对适宜的中间范围，其取值可能需要在不同目标尺度的检测需求之间进行平衡。

表4 损失引导系数对检测结果的影响 （%）

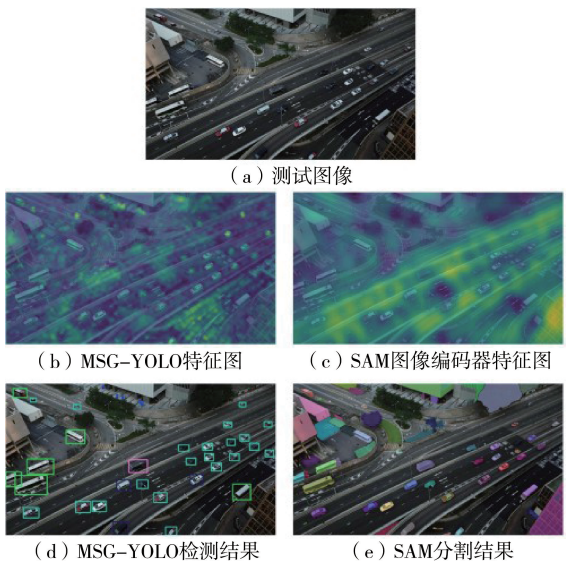
超参数λ	所有目标		小目标	
	mAP_{50}	mAP_{50-95}	mAP_{50}	mAP_{50-95}
0.015	55.2	35.1	45.0	24.7
0.025	55.8	35.2	45.2	24.8
0.035	54.9	34.9	44.7	24.7

2.4.3 引导约束方式对检测结果的影响

为进一步探究SAM特征与检测任务特征之间最佳的对齐机制，需要针对不同引导约束方式开展对比实验与分析。在已构建的SAM引导对齐框架基础上，本文分别采用L1损失与余弦相似度损失作为对齐准则，在相同实验设置下进行性能对比。实验结果如表5所示，采用余弦相似度约束的方法在所有目标和小目标上的mAP均高于采用L1损失的约束方式。具体而言，L1损失试图强制SAM特征与检测特征在数值分布上趋于一致。由于SAM与检测模型在训练目标、特征尺度与分布上存在固有差异，此类严格的数值约束可能会破坏检测网络自身已学习到的判别性特征结构。如图8所示，尽管检测特征与SAM特征在空间上关注相近区域，但两者在数值分布与语义聚焦层次上存在显著差异。因此，采用对绝对值不敏感、而专注于方向一致性的余弦相似度作为对齐准则更为有效。

表5 引导约束方式对检测结果的影响 （%）

引导方式	所有目标		小目标	
	mAP_{50}	mAP_{50-95}	mAP_{50}	mAP_{50-95}
余弦相似度	55.8	35.2	45.2	24.8
L1损失	54.7	34.9	44.3	24.5



（a）测试图像

（b）MSG-YOLO特征图 （c）SAM图像编码器特征图

（d）MSG-YOLO检测结果 （e）SAM分割结果

图8 MSG-YOLO和SAM的特征图及预测结果比较

注：颜色越明亮表明模型更关注这个区域

2.5 局限性分析

尽管 MSG-YOLO 在检测精度上取得了显著提升，但其参数量与计算复杂度仍处于较高水平，在客观上构成了其在资源受限的端侧设备上部署的局限。进一步的归因分析表明，这一复杂度主要源于为应对小目标检测挑战而设计的高容量特征提取主干网络。如表6所示，模型的主干网络部分贡献了约 58% 的参数与 55% 的 GFLOPs。其主要原因在于，为克服小目标分辨率低、特征弱的固有难题，模型必须构建强大的细粒度特征提取与上下文语义理解能力。为此，MSG-YOLO 采用了层次更丰富、容量更大的特征提取网络，旨在通过充足的参数建模能力满足高性能检测需求。为应对此局限，未来的轻量化研究将采取以下策略：首先通过知识蒸馏将当前模型的小目标表征能力迁移至轻量网络，大幅降低特征提取主干网络复杂度；随后实施结构化剪枝与量化，协同优化模型规模与计算效率。通过这一系列优化，力求在保持精度的同时，显著提升本方法在无人机等端侧平台的部署可行性。

表 6　模型复杂度参数分析

模型参数	主干网络	自顶向下融合	自底向上融合	检测头
GFLOPs	67.85	27.92	24.85	0.55
参数量/M	27.3	4.13	15.84	0.09

3 结论

本文针对无人机航拍中因成像距离远、目标占比小导致的小目标检测难题，提出一种基于多尺度特征融合和 SAM 视觉先验引导的小目标检测网络。该网络首先通过特征提取网络提取多尺度特征，并引入多分支空洞卷积和 Transformer 模块扩大模型的感知域并增强上下文建模能力。随后引入 SAM 大模型的稳健视觉先验，通过余弦相似度约束模型特征提取，增强模型特征表征能力。此后，通过多尺度特征自顶向下和自底向上的融合，得到多尺度融合的特征，并输入检测头得到检测结果。在 VisDrone-DET2019 上的实验结果表明，该方法在未降低所有目标检测精度的同时，显著提升了模型对小尺度目标的检测精度，达到业内较先进的水平。

参考文献

[1] ISTIAK A, SYEED M M M, HOSSAIN S, et al. Adoption of unmanned aerial vehicle (UAV) imagery in agricultural management: a systematic literature review [J]. Ecological Informatics, 2023, 78: 102305.

[2] 王泉，叶广飞，陈祺东. YOLO-SWR：无人机视角下轻量级交通车辆检测算法 [J]. 计算机工程与应用，2025，61(14)：112－122.

[3] 吕德利. 应急救援下无人机智能路径规划研究 [J]. 科技创新与应用，2023，13 (5)：97－99，103.

[4] CHEN C, ZHANG Y, LV Q, et al. RRNet: a hybrid detector for object detection in drone-captured images [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019: 100－108.

[5] BOSQUET B, CORES D, SEIDENARI L, et al. A full data augmentation pipeline for small object detection based on generative adversarial networks [J]. Pattern Recognition, 2023, 133: 108998.

[6] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 2117－2125.

[7] LIU S, QI L, QIN H, et al. Path aggregation network for instance segmentation [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 8759－8768.

[8] TAN M, PANG R, LE Q V. Efficientdet: scalable and efficient object detection [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 10781－10790.

[9] HONG M, LI S, YANG Y, et al. SSPNet: scale selection pyramid network for tiny person detection from UAV images [J]. IEEE Geoscience and Remote Sensing Letters, 2021, 19: 1－5.

[10] YANG X, YANG J, YAN J, et al. SCRDet: towards more robust detection for small, cluttered and rotated objects [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 8232－8241.

[11] ZHANG Y, YE M, ZHU G, et al. FFCA-YOLO for small object detection in remote sensing images [J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 5611215.

[12] KIRILLOV A, MINTUN E, RAVI N, et al. Segment anything [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023: 4015－4026.

[13] OQUAB M, DARCET T, MOUTAKANNI T, et al. DINOv2: learning robust visual features without supervision [J]. arXiv preprint arXiv: 2304.07193, 2023.

[14] WOO S, PARK J, LEE J Y, et al. CBAM: convolutional block attention module [C]//Proceedings of the European Conference on Computer Vision, 2018: 3－19.

[15] TIBSHIRANI R. Regression shrinkage and selection via the lasso [J]. Journal of the Royal Statistical Society Series B: Statistical Methodology, 1996, 58 (1): 267－288.

[16] LIU C. The Bayes decision rule induced similarity measures [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29 (6): 1086－1090.

[17] JOCHER G. Ultralytics YOLOv5 [EB/OL]. (2025－11－16) [2025－12－10]. https://github.com/ultralytics/yolov5.

[18] DU D, ZHU P, WEN L, et al. VisDrone-DET2019: the vision meets drone object detection in image challenge results [C]//

Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019：213 – 226.

[19] PASZKE A, GROSS S, MASSA F, et al. PyTorch：an imperative style, high-performance deep learning library [J]. Advances in Neural Information Processing Systems, 2019, 32.

[20] SHAMIR O, ZHANG T. Stochastic gradient descent for non-smooth optimization：convergence results and optimal averaging schemes [C]//International Conference on Machine Learning. PMLR, 2013：71 – 79.

[21] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. Yolov4：optimal speed and accuracy of object detection [J]. arXiv preprint arXiv：2004. 10934, 2020.

[22] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO：common objects in context [C]//European Conference on Computer Vision. Cham：Springer International Publishing, 2014：740 – 755.

[23] ZHU X, LYU S, WANG X, et al. TPH-YOLOv5：improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2021：2778 – 2788.

[24] XIAO Y, XU T, XIN Y, et al. FBRT-YOLO：faster and better for real-time aerial image detection [C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2025, 39 (8)：8673 – 8681.

（收稿日期：2025 – 12 – 29）

作者简介：

钟嘉宇（2000 – ），男，硕士研究生，主要研究方向：目标检测、深度学习图像处理。

牛利玲（1982 – ），女，硕士，工程师，主要研究方向：目标检测、机器学习。

任超（1988 – ），通信作者，男，博士，副研究员，主要研究方向：图像/视频处理、计算机视觉。E-mail：chaoren@ scu. edu. cn。

“工业互联网安全技术研究”主题专栏征稿启事

工业互联网作为新一代信息技术与制造业深度融合的产物，是推动产业数字化转型、实现经济高质量发展的关键基础设施。然而，随着其广泛部署与深度应用，网络攻击手段不断演进，安全风险日益凸显。工业互联网安全不仅关乎生产系统的稳定运行，更影响着关键基础设施的安全底线与数字经济的健康发展。为集中展示我国在工业互联网安全领域的最新理论研究成果、核心技术突破与创新应用实践，推动构建自主、安全、可靠的工业互联网安全保障体系，本刊拟在2026年第6期推出“工业互联网安全技术研究”主题专栏，现面向国内外广大专家学者、科研人员及行业工程师公开征稿。

一、征文主题：工业互联网安全技术研究

包括但不限于以下学术方向：

1. 工业互联网安全参考架构与标准体系；
2. 工控协议深度分析与安全加固；
3. 工业入侵检测与威胁感知；
4. 数据安全与隐私保护；
5. 工业智能体安全；
6. 5G + MEC 环境下的工业安全；
7. 供应链安全（第三方组件、软件库、开源工具的安全管控）；
8. 人工智能/机器学习用于攻击检测与防御；
9. 区块链技术在工业数据完整性、溯源方面的应用。

二、投稿要求

1. 稿件请用 word 格式录入，并套用本刊投稿模板。模板下载网址：http：//files. chinaaet. com/files/Periodical/pcachina_Templates. doc

2. 投稿文章须未在其他期刊或者出版正式论文集的会议上刊登过，且不在其他刊物或会议的审稿过程中，不存在一稿多投现象。

3. 无抄袭、剽窃、侵权、虚假引用等不良学术行为，且不违反相关法律法规，不涉及国家、企业秘密，稿件文责自负。

4. 论文要求观点鲜明、逻辑严谨、论据充分、方法合理，字数在5000 ~ 8000字。

5. 请在官方投稿网站（http：//www. pcachina. com）注册、投稿。注册后请投稿在“主题专栏”栏目，“稿件标题”请填“工业互联网 + 文章题目”。稿件经评审合格录用后，在《网络安全与数据治理》2026年第6期（正刊）以主题专栏形式发表。

三、时间安排

截稿日期：2026年4月30日

审稿反馈日期：2026年5月15日

出版日期：2026年6月15日

《网络安全与数据治理》编辑部

2026年3月

版权声明

凡《网络安全与数据治理》录用的文章，如作者没有关于汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权等版权的特殊声明，即视作该文章署名作者同意将该文章的汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权授予本刊，本刊有权授权本刊合作数据库、合作媒体等合作伙伴使用。同时，本刊支付的稿酬已包含上述使用的费用，特此声明。

《网络安全与数据治理》编辑部