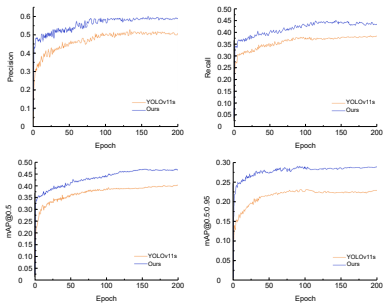
OEE | 光电工程 Opto-Electronic Engineering

Article

2026 年 2 月, 第 53 卷, 第 2 期

语义引导与多维特征感知的无人机航拍图像目标检测



王　光，李松泽*，李英硕

辽宁工程技术大学软件学院，辽宁 葫芦岛 125105

摘要：针对无人机航拍图像目标尺度差异大、空间分布不均衡，以及复杂背景下易出现的漏检、误检问题，本文提出一种语义引导与多维特征感知的无人机航拍图像目标检测方法。首先，嵌入自适应多维特征采样单元，通过多路径交互增强局部与全局特征协同融合，提高对多尺度特征的提取能力；其次，构建高效自适应池化层 (spatial pyramid adapt pool-fast, SPAPF) 池化层，保留特征细节与结构信息，并减少计算冗余，加快推理速度；然后，提出跨空间全局信息整合模块，利用跨空间多头自注意力机制增强对局部密集目标的全局感知能力，改善误检问题；最后，设计语义引导特征融合模块，实现浅层细节与深层语义特征高效融合，并集成一个针对小目标的附加检测头，提升定位和分类精度。在 VisDrone2019 和 VisDrone2021 数据集上进行的实验结果表明：本文算法的 mAP@0.5 达到 46.8% 和 44.7%，较基线模型提升了 7.2% 和 6.5%，优于其他对比算法，显著减少误检、漏检情况。

关键词：无人机航拍图像；小目标检测；自适应多维特征采样；跨空间全局信息整合；语义引导特征融合

The English title and abstract appear at the end of the paper.

DOI: 10.12086/oee.2026.250123 | CSTR: 32245.14.oee.2026.250123 | 中图分类号：TP391.4 | 文献标识码：A

王光，李松泽，李英硕. 语义引导与多维特征感知的无人机航拍图像目标检测 [J]. 光电工程，2026，**53**(2): 250123

Wang G, Li S Z, Li Y S. Semantic-guided and multi-dimensional feature perception for target detection in UAV aerial images[J]. *Opto-Electron Eng*, 2026, **53**(2): 250123

1　引　言

随着无人机平台在救援、农业监测等领域中的广泛应用，航拍图像数据量激增且质量提升，其高空、低空、多角度的拍摄优势，能够捕捉到丰富的空间信息和细粒度场景[1]。然而，传统的目标检测算法处理这类图像时面临诸多挑战，如拍摄距离远、目标尺寸小且分布密集，以及复杂背景的影响等，都阻碍了特征的准确提取，导致降低检测精度。此外，实时应用对检测速度的要求也对模型设计提出进一步的挑战。

传统机器学习方法依赖大量人工设计特征，难以捕捉图像中的高级特征[2]。为此，研究人员开发了基于深度学习的单阶段和两阶段检测算法。在两阶段算法中，模型如 R-CNN[3]、Faster R-CNN[4] 采用生成候选区域的策略，进行分类和边界框优化，以提高检测准确性。然而，两阶段算法虽提供了高检测精度，但复杂的网络结构和计算需求使推理速度减慢，不适合实时应用[5]。相比之下，YOLO 系列[6-9]、RetinaNet[10] 和 SSD[11] 等单阶段算法直接在图像上生成预测框，无需生成候选区域，显著缩短计算时间，更适合实时场景。其中，YOLO 系列因检测速度快、实现端到端受到广泛关注，在实时检测任务中表现出较大的优势。

收稿日期：2025-04-26；修回日期：2025-07-31；录用日期：2025-08-01；网络出版日期：2026-02-13

基金项目：国家自然科学基金资助项目 (62173171)

*通信作者：李松泽，2867314548@qq.com

版权所有©2026 中国科学院光电技术研究所

近年来，无人机航拍图像目标检测算法快速发展，众多高效算法被相继提出[12]。Sahin 和 Ozer[13] 在 YOLOv3 的基础上，提出 YOLODrone 算法，将原本 3 个不同尺度的输出层扩展到 5 个，获取更多位置信息，提高模型对小目标的定位性能。但在处理新数据时，过拟合风险增大，泛化能力降低。冒国韬等[14] 在 YOLOv5 基础上提出多尺度分割注意力单元和自适应加权特征融合模块，显著提升检测精度。然而模型的复杂度较大，忽视了精度和实时性之间的平衡。Chen 等[15] 将双层路由注意力机制引入 YOLOv7 中，持续捕获小目标关键特征，但模型参数量显著增加，降低了计算速度。Wang 等[16] 引入 BiFormer 机制到 YOLOv8 骨干网络中，并采用梯度分配及 WIoUv3 损失函数提高模型定位能力。但在高密度目标重叠时，易发生漏检、误检问题。Huang 等[17] 提出的 TCA-YOLOv5m，利用坐标注意力模块提升特征提取的准确性，但对多尺度目标捕捉不充分。Fang 等[18] 提出的 ODFC-YOLO，实现了图像去雾与目标检测的同步处理，然而增大了额外计算量和参数量。Betti 和 Tucci[19] 提出的 YOLO-S，利用 Reshape-passthrough 层促进跨网络的特征重用，并将低级位置信息与高级信息相融合，提高小目标检测精度。Qiu 等[20] 提出的 LD-YOLOv10，设计轻量级特征提取结构 RGELAN，利用重参数化卷积和 Conv-Tiny 作为计算结构，减轻计算负担，但检测精度还有待进一步加强。Dosovitskiy 等[21] 设计的 ViT 网络，通过分层窗口化捕获每个窗口内的长距离信息，并应用级联转换器提取不同窗口内的目标特征向量，但无法有效提取局部细节信息。

为解决上述挑战，本文基于 YOLOv11s 算法，提出语义引导与多维特征感知的无人机航拍图像目标检测算法。首先，在主干网络中嵌入自适应多维特征采样单元，对输入特征图沿通道分组，并采用并行路径进行多尺度特征提取。其次，在金字塔池化结构基础上设计 SPAPF 池化层，引入自适应特征池化机制，在不同尺度池化分支上采用加权特征融合策略，优化特征提取过程，并利用轻量化设计，提高推理速度。然后，构造跨空间全局信息整合模块，利用跨空间学习机制增强全局信息建模能力，实现多尺度目标的动态聚焦，降低误检率。最后，提出语义引导特征融合模块，结合跨层级特征对齐策略和亚像素卷积，实现浅层细节与深层语义信息的高效融合，增强对小目标纹理信息的捕捉。

2 语义引导与多维特征感知的无人机航拍图像目标检测

为解决无人机航拍图像目标检测中多尺度变化、小目标分布密集和背景干扰导致的漏检、误检问题，本文基于 YOLOv11s 算法，提出了语义引导与多维特征感知的无人机航拍图像目标检测算法，保证高检测精度的同时降低计算复杂度。

YOLOv11s 算法由三个主要部分构成：主干网络 (Backbone)、颈部网络 (Neck) 和检测头 (Head)。主干网络部分利用 5 个 CBS 卷积单元，以及 4 个 C3k2 模块对输入尺寸为 640×640 的图像进行特征采样，此阶段主要负责从原始图像中提取低、中、高层次的特征信息。为进一步整合和增强来自主干网络的多尺度特征，颈部网络设计了一系列模块，包括 1 个 SPPF 模块、1 个 C2PSA 模块、4 个拼接单元 (Concat)、4 个 C3k2 模块、2 个上采样操作单元 (Upsample) 和 2 个 CBS 卷积单元。通过对不同尺度特征的交互和融合，颈部网络可以捕获更丰富的上下文信息。最终将颈部网络输出的三组特征图送入检测头，进行分类与边界框回归任务，生成尺寸为 80×80×256、40×40×512 和 20×20×1024 的预测特征图，用于实现不同尺度目标的检测。

本文提出的改进算法总体结构如图 1 所示，主要做出以下四处改进：1) 设计了自适应多维特征采样 (adaptive multi-dimensional feature sampling, AMFS) 单元替换部分 CBS 卷积模块，利用多维度信息交互、动态优化特征采样过程，增强网络对小目标特征的提取能力。2) 构建高效自适应池化层 (spatial pyramid adapt pool-fast, SPAPF)：采用 Adapool[22] 池化机制，保持轻量化架构的同时，提升了对小目标特征的保留能力和跨尺度泛化性能。3) 提出 CS_C2PSA (cross spatial_cross stage partial with pyramid squeeze attention) 模块实现跨空间全局信息整合，结合阶段性残差连接结构、跨空间学习的多头自注意力机制和 AMFS，实现了全局上下文感知与局部细节增强的协同优化，同时抑制背景冗余干扰，更精准地捕捉目标特征。4) 加入语义引导特征融合模块 (semantic-guided feature fusion, SGFF)：通过高级语义引导低级特征融合，实现跨尺度的特征学习，并设计小目标专用检测头，有效提升对无人机航拍小目标的检测精度与效率。

2.1 自适应多维特征采样

针对无人机航拍图像中目标尺度极端变化差异大、冗余信息干扰显著等问题，本文设计了自适应多维特征采样单元 AMFS，通过多路径分支并行提取多维度特征，实现特征层级间的自适应协同融合，增强网络对局部细节和全局上下文信息的捕捉能力，从而弥补传统固定卷积在感受野上的不足，提高特征提取的丰富性和多样性；此外，采用自适应调整卷积核权重策略，动态优化特征提取过程，使模型聚焦于目标区域的关键语义表达，有效抑制背景冗余噪声干扰，提升对多尺度目标的感知能力，减少误检、漏检情况。其结构如图 2 所示。

首先，将输入的特征映射$\boldsymbol{X}\in\mathbf{R}^{C\times H\times W}$沿通道维度划

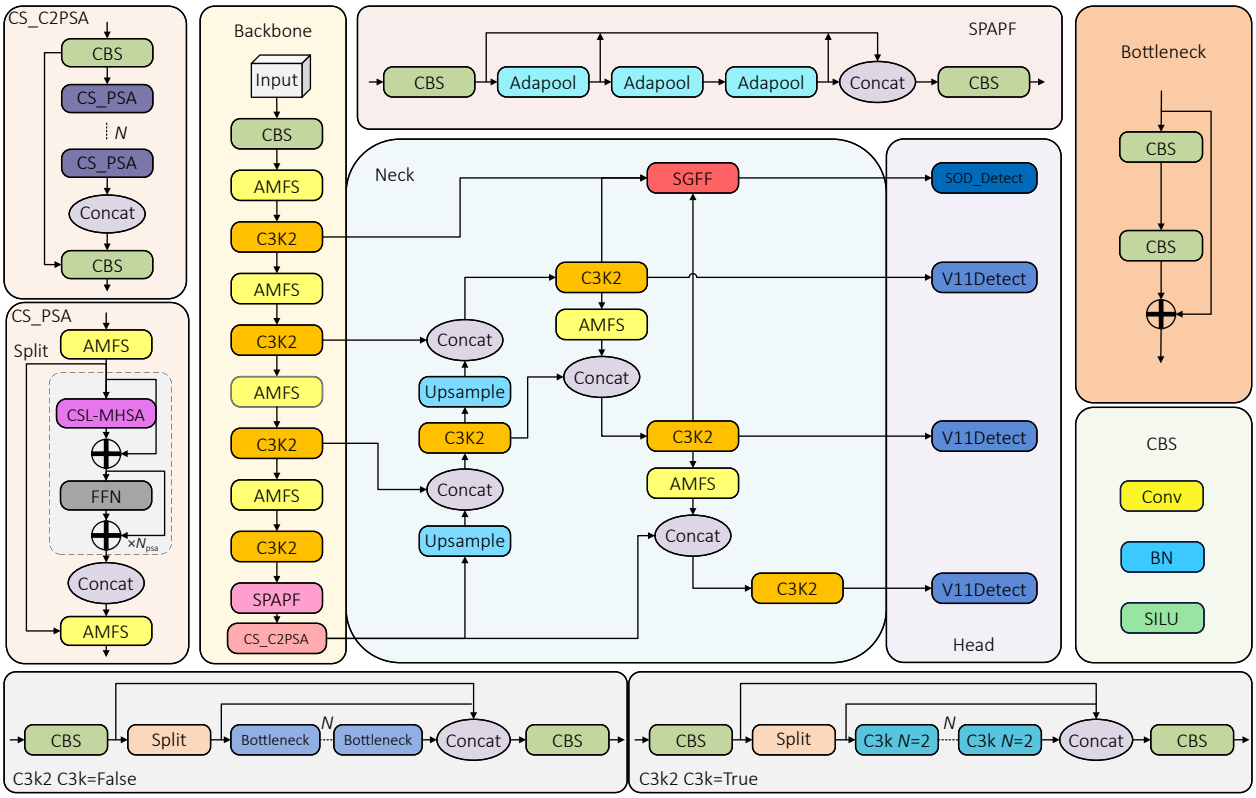


图 1 模型总体结构

Fig. 1 Overall model structure

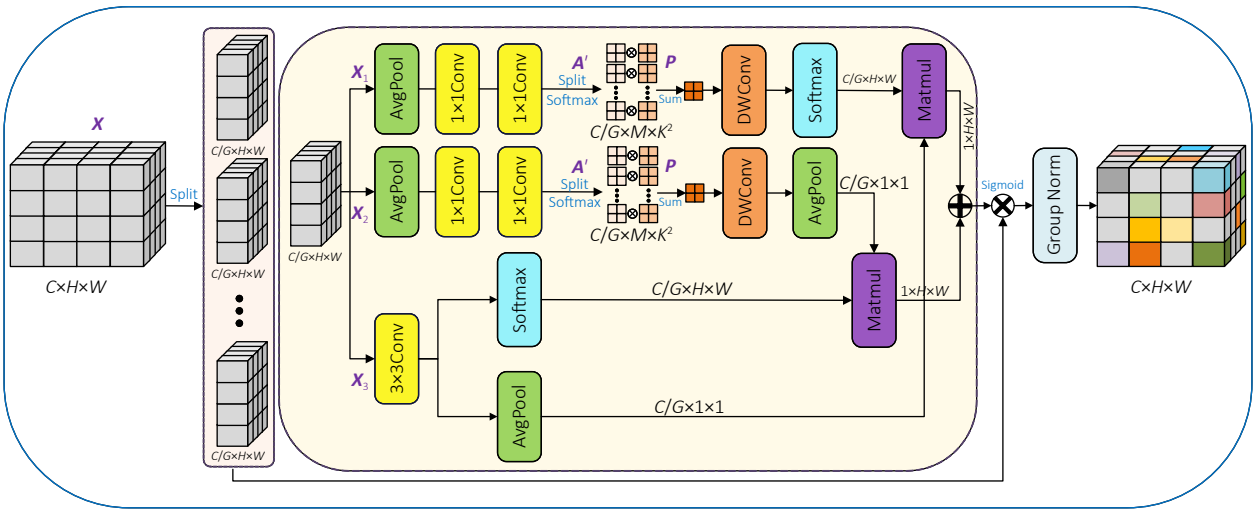


图 2 自适应多维特征采样单元 (AMFS) 模块结构

Fig. 2 Adaptive multi-dimensional feature sampling unit (AMFS) module structure

分为 G 个子特征组，每组尺寸为$\boldsymbol{X}\in\mathbf{R}^{C/G\times H\times W}$。其中：$C$ 表示特征映射的总通道数，H 表示特征映射的高度，W 表示特征映射的宽度。接着，使用三个并发路径从分组特征映射中提取注意力权重描述符，收集局部多尺度的空间信息。

然后，对前两条路径的子特征映射 $\boldsymbol{X}_1$ 和 $\boldsymbol{X}_2$，先通过自适应平均池化 (adaptive average pooling, AvgPool)-将空间维度压缩至 K^2，有效捕捉全局上下文信息；再利用两个连续的 1×1 卷积生成注意力图$\boldsymbol{A}'$，格式为$\mathbf{R}^{C/G\times M\times K^2}$，动态聚焦输入特征图中的关键区域。接着在通道维度上应用 Softmax 函数创建注意力权重$\boldsymbol{A}\in\mathbf{R}^{\frac{C}{G}\times M\times K^2}$($M$ 表示注意力权重描述符的数量），并将 $\boldsymbol{A}$ 与一组可学习参数$\boldsymbol{P}\in\mathbf{R}^{C/G\times K^2}$逐元素相乘，以动态调整输入特征属性，抑制背景噪声影响，提高网络对小目标的聚焦和检测能力；将这两个特征映射的结果进行深度卷积 (depthwise convolution, DWC)[23] 处理，更细致地捕捉局部特征，增强对小目标的感知能力。

为提升空间特征分析能力，先对子特征映射 $\boldsymbol{X}_3$ 进行 3×3 卷积处理，然后分为两条路径进行，第一条路径经过 Softmax 处理，输出格式为$C/G\times H\times W$，使网络更好地聚焦目标物体的关键空间特征，提高对小目标的定位能力；第二条路径通过 AvgPool 转换为$C/G\times 1\times 1$，用于捕捉全局上下文信息，进而有效区分局部相似目标，大幅降低误检率。

为了进一步增强多尺度特征的表达能力，将 $\boldsymbol{X}_1$ 经 DWC 处理后的输出应用 Softmax 函数，并与 $\boldsymbol{X}_3$ 第二条分支路径的输出通过矩阵乘法 Matmul 相融合，转换为 $1\times H\times W$；将 $\boldsymbol{X}_2$ 经 DWC 处理后的输出应用 AvgPool 转换为$C/G\times 1\times 1$后，与 $\boldsymbol{X}_3$ 第一条分支路径的输出通过 Matmul 相融合，同样转换为$1\times H\times W$。然后，将这两个结果相加以增强信息整合能力，并进行 Sigmoid 函数[24] 调制后，与初始的其他特征组相乘，通过残差连接保留原始特征的空间细节，避免信息丢失。最后，经过 Group Norm 归一化[25] 得到最终输出，提高了模型对多尺度目标的感知能力。

2.2 SPAPF

为解决传统空间金字塔池化层在特征下采样过程中计算开销大、关键信息易丢失等问题，本文提出一种 SPAPF 高效自适应池化层。通过 Adapool 池化机制，充分保留局部细节与结构化语义信息，从而增强对多尺度目标的感知能力，进一步优化特征层的表达与压缩，显著减少冗余计算，加快推理速度，提高模型整体性能。池化原理如图 3 所示。

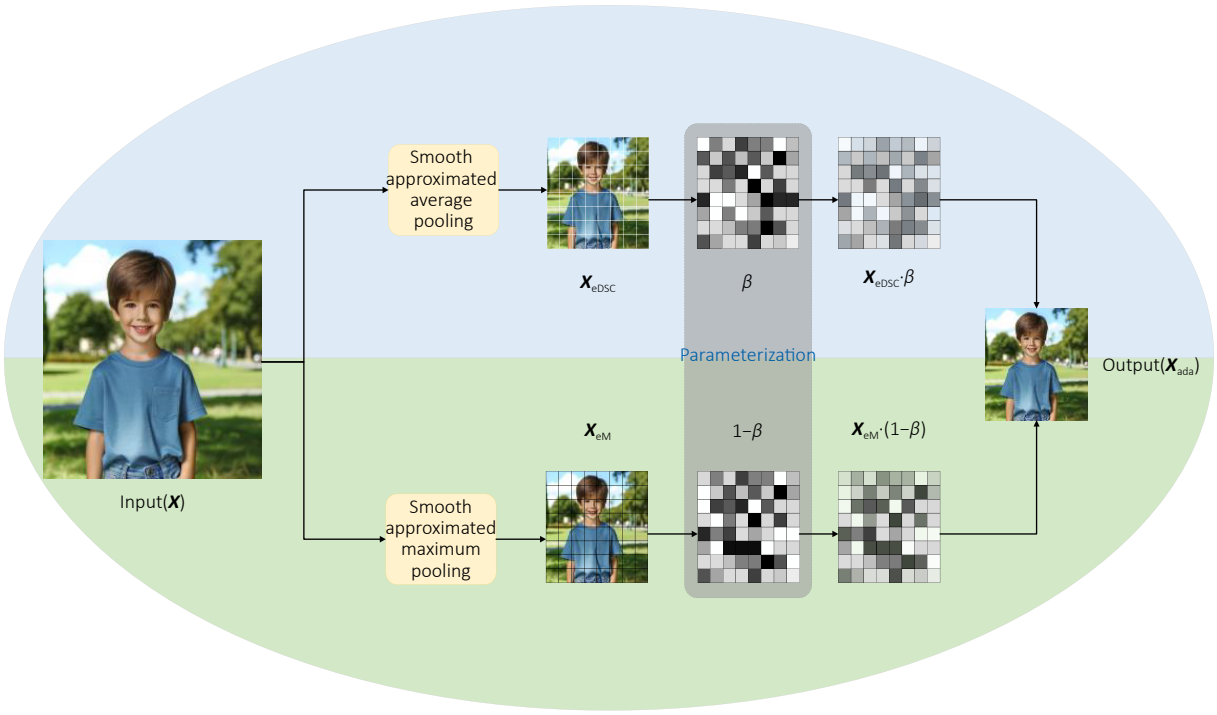


图 3 Adapool 池化原理

Fig. 3 Principle of Adapool pooling

Adapool 由平滑近似平均池化[26-27]和平滑近似最大池化[28]方法构成，利用核区域加权，在保留细节信息和本地图像结构之间提供了一种平衡，实现了两种池化方法的学习融合。

1) 平滑近似平均池化 (Smooth approximated average pooling)。Dice-sørensen coefficient 系数 (DSC) 是一种相似度度量，使用距离的指数作为权重$\boldsymbol{w}(\bar{\boldsymbol{a}},\boldsymbol{a}_i)$，获取核区域内激活向量$\boldsymbol{a}_i$和平均激活$\bar{\boldsymbol{a}}$之间的相似度，并通过相似度指数化操作保证池化方法在反向传播期间可微，并减少梯度消失问题的可能性，特征值至少具有预设的最小梯度。再根据指数化权重$\mathrm{e}^{\boldsymbol{w}(\bar{\boldsymbol{a}},\boldsymbol{a}_i)}$对核值进行指数加权，确保分配非零权重，从而提升激活特征的保留效果，降低离群值的负向影响，抑制背景干扰，强化有效目标特征，并输出池化结果$\tilde{\boldsymbol{a}}_{\text{eDSC}}$。该过程表示为

$$\begin{cases} \tilde{\boldsymbol{a}}_{\text{eDSC}} = \sum\limits_{i\in\mathbf{R}} \dfrac{\mathrm{e}^{\boldsymbol{w}(\bar{\boldsymbol{a}},\boldsymbol{a}_i)}\cdot\boldsymbol{a}_i}{\sum\limits_{j\in\mathbf{R}}\mathrm{e}^{\boldsymbol{w}(\bar{\boldsymbol{a}},\boldsymbol{a}_j)}} \\ \boldsymbol{w}(\bar{\boldsymbol{a}},\boldsymbol{a}_i) = \dfrac{2\boldsymbol{a}_i\cdot\bar{\boldsymbol{a}}}{\boldsymbol{a}_i+(\bar{\boldsymbol{a}})^2} \end{cases}. \tag{1}$$

2) 平滑近似最大池化 (Smooth approximated maximum pooling)。该方法使用类似 Softmax 的自然指数 e 计算核区域内特定特征值的权重参数，以确保较大的激活对输出产生更高的影响，最小的权重分配给最小的激活，从而突出目标的局部特征响应。权重计算方法为

$$\boldsymbol{w}(\boldsymbol{a}_i) = \frac{\mathrm{e}^{\boldsymbol{a}_i}}{\sum\limits_{j\in\mathbf{R}}\mathrm{e}^{\boldsymbol{a}_j}}, \tag{2}$$

式中：$\boldsymbol{w}(\boldsymbol{a}_i)$表示第 i 个核区域权重，其数值大小反映了该区域内激活值的重要性，$\boldsymbol{a}_i$表示当前核区域的激活值。

最终的池化输出由各区域的权重信息加权激活总和得到，方法为

$$\tilde{\boldsymbol{a}}_{\text{eM}} = \sum_{i\in\mathbf{R}} \boldsymbol{w}(\boldsymbol{a}_i)\cdot\boldsymbol{a}_i, \tag{3}$$

式中：$\tilde{\boldsymbol{a}}_{\text{eM}}$表示池化输出。

平滑近似最大池化中的权重是基于相应激活值的非线性变换。较高值的激活要比较低值的激活更具主导地位。此外，核区域内的所有激活对输出均有贡献，激活值越高，影响越大。

为进一步获取核区域内最具信息量的特征，通过可学习权重掩码β创建平滑近似平均池化和平滑近似最大池化的组合体，实现两者属性优势的互补，显著降低冗余计算，减少对关键特征的破坏，更好地保留纹理与边缘信息。其中β用来调整每个区域$\boldsymbol{R}$内两种方法的使用比例。组合过程为

$$\tilde{\boldsymbol{a}}_{\text{ada}} = \beta\cdot\tilde{\boldsymbol{a}}_{\text{eDSC}} + (1-\beta)\,\tilde{\boldsymbol{a}}_{\text{eM}}, \tag{4}$$

式中：$\tilde{\boldsymbol{a}}_{\text{ada}}$为自适应池化输出；$\beta\in\{0,\ldots,1\}$。

2.3 跨空间全局信息聚合

为提升复杂背景干扰下的检测精度，本文设计了 CS_C2PSA 模块，以进一步增强模型的全局感知能力，结构如图 4 所示。该模块利用跨空间学习的多头自注意力机制 (cross-spatial learning multi-head self-attention, CSL-MHSA) 与阶段性残差连接结构，加强对全局上下文信息的捕捉，提高模型在不同空间维度方向的信息聚合能力，解决了传统自注意力机制[29]在空间建模过程中易忽略相邻区域依赖关系，从而导致小目标因语义模糊与背景或相邻目标混淆的问题，有效改善误检现象。

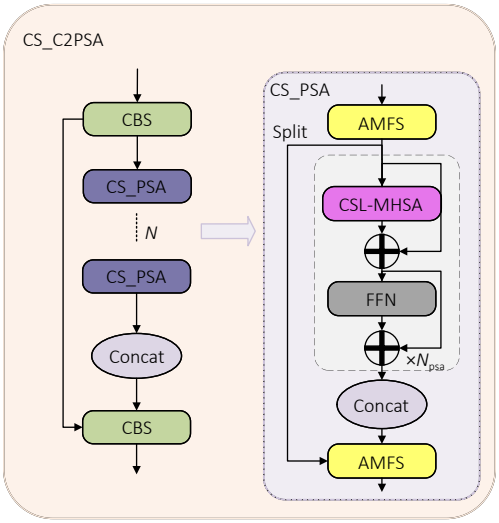


图 4 CS_C2PSA 结构

Fig. 4 Structure of CS_C2PSA

CSL-MHSA 由多个自注意力头并行组成，其中 $\boldsymbol{X}$ 为输入特征，$\boldsymbol{Z}_i(i=1,2,\ldots,n)$为每个头部的输出特征。首先，对 $\boldsymbol{X}$ 分别做三次线性变换 ($\boldsymbol{W}_Q$、$\boldsymbol{W}_K$和$\boldsymbol{W}_V$) 计算出查询 ($\boldsymbol{Q}$)、键 ($\boldsymbol{K}$) 和值 ($\boldsymbol{V}$) 矩阵，用于后续的注意力计算。其中$\boldsymbol{W}_Q$和$\boldsymbol{W}_K$使用 1×1 卷积以关联局部感受野，$\boldsymbol{W}_V$使用 3×3 卷积以关联全局多尺度感受野。此外，为了增强空间感知能力，引入反映垂直方向 (高度) 的空间位置信息 ($\boldsymbol{R}_{\text{h}}$) 和反映水平方向 (宽度) 的空间位置信息 ($\boldsymbol{R}_{\text{w}}$)。然后依据预定义的头部数量对其进行分层操作。

在每个注意力头中，输出结果由两部分组成，一是构建空间增强的主注意力分支，二是实现对空间结构感知的残差补偿。结构如图 5 所示。

首先，计算 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 的点积，得到注意力分数$\boldsymbol{QK}^{\mathrm{T}}$，然后与空间位置信息$\boldsymbol{Q}(\boldsymbol{R}_{\mathrm{h}}+\boldsymbol{R}_{\mathrm{w}})$相加，构建增强型注意力矩阵，接着对其进行 Softmax 函数归一化，并与全局感受野 $\boldsymbol{V}$ 做矩阵乘法 (Matmul)，完成主注意力信息的提取，实现全局特征融合。公式为

$$\boldsymbol{Z}_{i1}=Softmax\left(\boldsymbol{QK}^{\mathrm{T}}+\boldsymbol{Q}(\boldsymbol{R}_{\mathrm{h}}+\boldsymbol{R}_{\mathrm{w}})\right)\boldsymbol{V}. \tag{5}$$

Softmax 函数用于将查询、键、高度和宽度的相似度以概率形式归一化，公式为

$$Softmax(\boldsymbol{f}_y)=p(y\mid\boldsymbol{x})=\frac{\exp\boldsymbol{f}_y}{\sum_{c=1}^{C}\exp\boldsymbol{f}_c}, \tag{6}$$

其中：$\boldsymbol{x}$ 表示输入的特征图；y 表示特征图中某一通道 (或某一空间位置) 的索引；$p(y\mid\boldsymbol{x})$ 是该通道 (或位置) 的注意力概率，概率越高，模型对该部分特征的关注度越强，从而聚焦小目标、抑制背景噪声。

为了进一步强化结构信息感知，将$\boldsymbol{QK}^{\mathrm{T}}$进行 Softmax 获得注意力权重。随后，将融入位置关系的注意力矩阵$(\boldsymbol{QK}^{\mathrm{T}}+\boldsymbol{Q}(\boldsymbol{R}_{\mathrm{h}}+\boldsymbol{R}_{\mathrm{w}}))$与上述注意力权重进行 Matmul，最终得到结构感知补偿输出。公式为

$$\boldsymbol{Z}_{i2}=Softmax\left(\boldsymbol{QK}^{\mathrm{T}}\right)\left(\boldsymbol{QK}^{\mathrm{T}}+\boldsymbol{Q}(\boldsymbol{R}_{\mathrm{h}}+\boldsymbol{R}_{\mathrm{w}})\right). \tag{7}$$

将两个运算结果相加，生成该注意力头的最终输出，其公式为

$$\begin{aligned}\boldsymbol{Z}_i=&Softmax\left(\boldsymbol{QK}^{\mathrm{T}}+\boldsymbol{Q}(\boldsymbol{R}_{\mathrm{h}}+\boldsymbol{R}_{\mathrm{w}})\right)\boldsymbol{V}\\&+Softmax\left(\boldsymbol{QK}^{\mathrm{T}}\right)\left(\boldsymbol{QK}^{\mathrm{T}}+\boldsymbol{Q}(\boldsymbol{R}_{\mathrm{h}}+\boldsymbol{R}_{\mathrm{w}})\right).\end{aligned} \tag{8}$$

最后，将所有头的输出加和，得到 CSL-MHSA 的总输出。

2.4 语义引导特征融合

为在检测阶段更好地表达细粒度语义并增强对小目标的感知能力，本文设计了语义引导特征融合模块 SGFF，并依此优化检测头，提出专门用于小目标检测的 SOD_Detect，有效缓解密集小目标场景下的定位不准与分类模糊问题。该模块利用跨层级特征对齐策略，实现浅层细节与深层语义信息的协同融合，增强语义表达能力的同时，保持目标边缘与定位信息的完整性，从而生成语义丰富的高分辨率特征图，作为 SOD_Detect 的输入。此外，为保留和恢复小目标图像的更多细节信息，引入亚像素卷积[30] 至上采样阶段，利用高维通道特征的空间重排，提升小目标的判别语义，显著降低漏检率。结构如图 6 所示。

1) 跨层级特征对齐策略。$\boldsymbol{F}_1$~$\boldsymbol{F}_4$ 为四个检测头的输入，其中 $\boldsymbol{F}_1$ 对应新增检测头 SOD_Detect，$\boldsymbol{F}_2$、$\boldsymbol{F}_3$、$\boldsymbol{F}_4$ 为路径聚合特征金字塔网络 (PAFPN) 提取的深层三层特征。首先，对 $\boldsymbol{F}_3$ 进行通道扩张、语义特征提取和亚像素重排操作以获得$\boldsymbol{F}'_3$。在语义特征提取阶段，本文利用 SE 通道注意机制[31]，自适应学习每个通道的重要性，有效抑制背景噪声和冗余信息干扰，有助于生成更好的

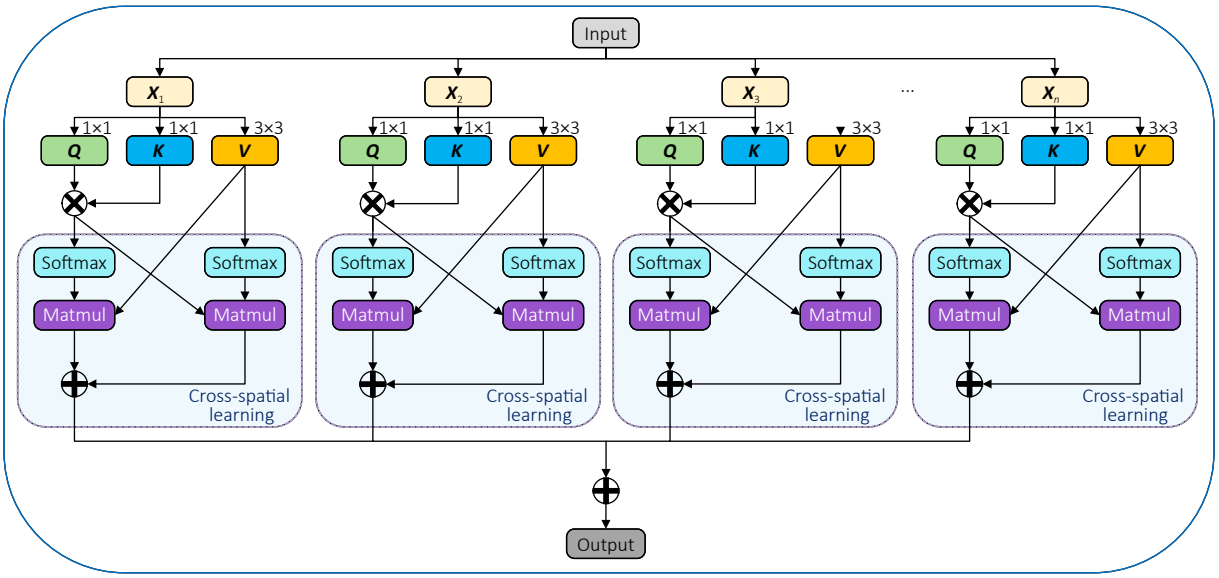


图 5 CSL-MHSA 结构

Fig. 5 Structure of CS-MHSA

高分辨率特征图。首先使用全局平均池化 (global average pooling, GAP)[32] 对特征映射进行压缩，然后使用一组完全连通的层来探索不同通道之间的重要关系，再使用 Sigmoid 激活函数为不同通道生成权重，最后对原始特征图进行通道加权，以增强通道维度上的特征表达，使得小目标的辨识更加明显。

为保证细节信息的丰富性，将$\boldsymbol{F}_3'$与 $\boldsymbol{F}_2$ 沿通道维度进行连接。然后，将其输入细节特征提取模块，学习局部关键纹理特征，最终生成具有更强语义特征和局部细节特征的高分辨率特征图 $\boldsymbol{F}_{23}$，表示为

$$\boldsymbol{F}_{23} = \boldsymbol{F}_3' + EF_{\mathrm{d}}(Cat(\boldsymbol{F}_2, \boldsymbol{F}_3')) , \tag{9}$$

式中：$EF_{\mathrm{d}}(\cdot)$表示细节特征提取模块；$Cat(\cdot)$表示特征拼接操作。

最后，通过上采样将 $\boldsymbol{F}_{23}$ 与骨干网络提取的浅层特征$\boldsymbol{P}_1'$融合得到 $\boldsymbol{F}_1$。与原始特征图$\boldsymbol{P}_1'$相比，$\boldsymbol{F}_1$ 具有同样大的尺寸，但包含更丰富的语义和局部细节信息。$\boldsymbol{F}_1$ 的生成过程表示为

$$\boldsymbol{F}_1 = \boldsymbol{P}_1' + Up_{2\times}(\boldsymbol{F}_{23}) , \tag{10}$$

式中：$Up_{2\times}(\cdot)$表示采用倍数为 2 的上采样。

2) 亚像素卷积。与传统上采样方法相比，亚像素卷积不仅减少冗余计算和内存消耗，还显著增强了特征表达对空间结构的适应性，对小目标的精细感知尤为有效，特别是在密集小目标场景中表现出明显优势。首先，通过常规卷积将低分辨率特征图转换为具有较高通道数的特征图，然后利用像素重排 (pixel shuffle) 操作将通道信息重新分配到空间维度，实现语义特征从通道域到空间域的转换，完成高效上采样。操作过程如图 7 所示。通过 pixel shuffle，尺寸为$H\times W\times u^2C$的特征图$\boldsymbol{F}$变换为尺

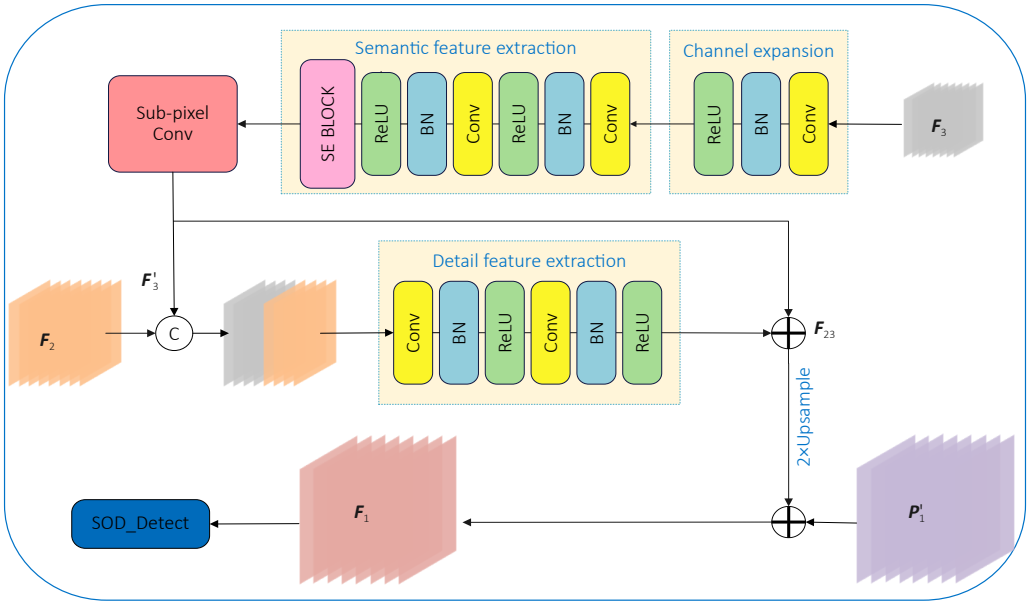


图 6 语义引导特征融合 (SGFF) 结构

Fig. 6 Semantic-guided feature fusion (SGFF) structure

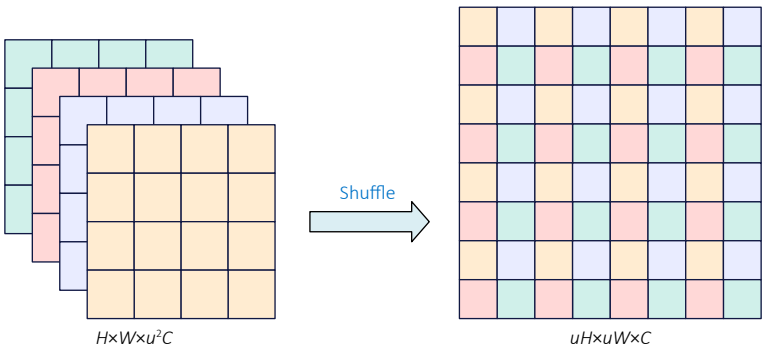


图 7 亚像素重排操作

Fig. 7 Subpixel shuffle operation

寸为$uH\times uW\times C$的特征图$\boldsymbol{F}'$。公式如下：

$$PS(\boldsymbol{F})_{x,y,c}=\boldsymbol{F}_{\left\lfloor\frac{x}{u}\right\rfloor,\left\lfloor\frac{y}{u}\right\rfloor,C*u*\mathrm{mod}(y,u)+C\times\mathrm{mod}(x,u)+c}, \tag{11}$$

式中：$PS(\boldsymbol{F})_{x,y,c}$表示在像素重排操作$PS(\cdot)$之后，$\boldsymbol{F}'$上位置$(x,y,c)$的像素值；$u$ 为上采样因子 (u=2)；$\lfloor\cdot\rfloor$表示底面函数；mod(a,b)表示 a 除以 b 的余数。

3 实验结果与分析

3.1 数据集及实验环境

VisDrone2019[33] 数据集是由天津大学机器学习与数据挖掘实验室 AISKYEYE 团队构建的无人机视觉基准数据集，旨在推动无人机平台下复杂场景的目标检测与分析研究。该数据集包含 10209 张静态图像，其中 6471 张用于训练集，548 张用于验证集，3190 张用于测试集，支持模型的分阶段训练与评估。其覆盖了中国 14 个城市的多样化场景，包括城市与乡村环境、密集与稀疏目标分布，以及雨雾、强光、低光照等极端天气条件。提供了超过 260 万个边界框标注，涵盖行人、汽车、自行车和公交车等 10 类核心目标。VisDrone2021[34] 数据集在 VisDrone2019 基础上扩展了场景多样性，提升了标注精细度。

本实验使用的操作系统为 Windows 10，GPU 为 NVIDIA GeForce RTX 3090，深度学习框架为 Pytorch 1.10.1＋CUDA 11.1，编程语言为 Python 3.8，设置 batch_size 为 16，epochs 为 200。

3.2 评价指标

本实验采用准确率 (Precision)、召回率 (Recall)、平均精度均值 (AP)、每秒检测帧率 (FPS)、模型计算量 (GFLOPs) 来衡量模型性能。对应计算公式为

$$Precision=\frac{TP}{TP+FP}, \tag{12}$$

$$Recall=\frac{TP}{TP+FN}, \tag{13}$$

$$AP=\int_0^1 Precision(Recall)\mathrm{d}(Recall), \tag{14}$$

$$mAP=\frac{1}{n}\sum_{i=1}^{n}\int_0^1 Precision(Recall)\mathrm{d}(Recall), \tag{15}$$

$$FPS=\frac{Framenum}{ElapsedTime}, \tag{16}$$

式中：TP 为正确检测目标数目；FP 为误检目标数目；FN 为漏检目标数目；$Framenum$ 表示测试阶段处理的图像帧总数；$ElapsedTime$ 表示处理 N 帧图像的总时间。

3.3 消融实验

为评估各项改进策略在 YOLOv11s 基础上的实际效果，本文在 VisDrone2019 数据集上进行了系统的消融实验，在实验过程中保持所有参数设置与环境配置完全一致，确保对比的公平性。实验结果见表 1 所示，其中“√”表示引入相应模块，“×”表示未引入。

表 1 实验结果表明，单独引入 AMFS 单元，Precision 提高了 3.5%，Recall 提高 1.5%，mAP@0.5 提高 3.2%，说明 AMFS 增强了特征提取能力，有效降低误检率。用 SPAPF 替换传统 SPPF，Recall 提高 2.2%，GFLOPs 降低 4.4，保留更多关键细节，加快模型推理速度。改进 CS_C2PSA 模块后，Precision 提高 2.3%，Recall 提高 1.9%，mAP@0.5 提高 2.0%，说明该模块提高了获取全局上下文信息的能力，抑制冗余信息干扰，减少类别混淆。SGFF 模块利用深层语义信息在浅层特征中的充分表达，使 Precision 提高 3.9%，mAP@0.5 提高 3.4%，显著提升了小目标的定位精度，增强整体检测

表 1 改进算法在 VisDrone2019 数据集的消融实验

Table 1 Ablation experiments of the improved algorithm on the VisDrone2019 dataset

Number	YOLOv11s	AMFS	SPAPF	CS_C2PSA	SGFF	Precision/%	Recall/%	mAP@0.5/%	GFLOPs
1	√	×	×	×	×	50.2	38.4	39.6	30.7
2	√	√	×	×	×	53.7	39.9	42.8	32.5
3	√	×	√	×	×	52.3	40.6	41.2	26.3
4	√	×	×	√	×	52.5	40.3	41.6	33.8
5	√	×	×	×	√	54.1	39.5	43.0	31.2
6	√	√	√	×	×	54.4	42.0	43.4	27.4
7	√	√	√	√	×	55.0	42.6	43.6	29.7
8	√	√	√	√	√	58.3	43.1	46.8	30.8

性能。

将 AMFS 和 SPAPF 同时加入基线模型，mAP@0.5 提升了 3.8%，GFLOPs 降低 3.3，说明 SPAPF 对整体模型起到轻量化作用，加快模型训练速度。将 AMFS、SPAPF 和 CS_C2PSA 引入原模型后，Recall 提高了 4.2%，mAP@0.5 相比仅添加 AMFS 和 SPAPF 又提高了 0.2%，增强模型对关键特征的提取能力。将 AMFS、SPAPF、CS_C2PSA 和 SGFF 四个模块集成到基线模型 YOLOv11s 中，与原模型相比，Precision 提升了 8.1%，Recall 提高 4.7%，mAP@0.5 增长了 7.2%。实验结果表明，本文提出的改进算法显著增强小目标检测性能，改善了背景冗余干扰和小目标密集分布下易出现的误检、漏检现象，进一步验证了模型结构优化的合理性与实用性。

为更好地展示改进算法在各类目标上的检测效果，本文在 VisDrone2019 数据集上，通过消融实验对比各目标类别的 mAP@0.5，见表 2。结果表明，各类别目标的 mAP@0.5 均有提升，说明改进算法在多尺度目标检测任务中具有显著优势，具备良好的泛化能力，尤其适用于无人机航拍场景下的目标检测任务。

3.4 对比试验

为进一步验证改进算法的检测效果，本文算法分别与 SSD、Faster R-CNN、RetinaNet、YOLOv5s、Improved YOLOv7[35]、YOLOv8s、YOLOv10s[36]、YOLOv11s、PP-YOLOE[37]、ESO-DETR[38]、MIS-YOLOv8[39] 在准确率、召回率、平均精度等方面进行对比，实验结果如表 3 所示。

实验表明，与经典算法 Faster-RCNN 和 RetinaNet 相比，本文算法的 Precision 分别提升了 9.8% 和 20.9%，mAP@0.5 提升 10.5% 和 16.9%，GFLOPs 分别减少 12.1 和 54.8，FPS 分别提高了 5 倍和 4 倍，减少冗余计算，充分展现出极高的效率优势。与 YOLOv5s 相比，本文算法的 Precision 提高了 9.9%，Recall 提高 6.9%，mAP@0.5

表 2 消融实验各类别精度对比 (单位：%)

Table 2 Comparison of accuracy for each category in the ablation experiments (unit: %)

Number	Pedestrian	People	Bicycle	Motor	Tricycle	Awningtricycle	Car	Van	Bus	Truck	mAP@0.5
1	40.2	30.0	16.8	41.2	25.2	22.8	77.6	45.0	56.1	40.8	39.6
2	43.9	33.8	20.7	44.3	28.6	26.2	80.5	48.1	58.9	43.7	42.8
3	42.5	31.6	18.6	42.8	27.0	24.4	78.9	46.6	57.4	42.2	41.2
4	42.2	32.2	18.8	43.2	27.4	24.9	79.4	47.1	57.8	42.8	41.6
5	43.9	33.1	20.3	44.7	28.9	26.1	81.0	48.5	58.9	43.9	43.0
6	44.4	33.8	21.1	45.0	29.4	26.6	81.2	48.8	59.6	**46.3**	43.7
7	44.6	34.0	21.8	45.2	29.6	26.8	81.6	49.1	59.7	44.8	43.6
8	**49.4**	**39.2**	**23.0**	**47.2**	**32.4**	**29.4**	**86.8**	**51.7**	**63.2**	45.9	**46.8**

表 3 VisDrone2019 数据集对比实验

Table 3 Comparison experiments on the VisDrone2019 dataset

Model	Precision/%	Recall/%	mAP@0.5/%	mAP@0.5:0.95/%	GFLOPs	FPS
SSD	45.1	33.2	32.4	15.7	20.5	110
Faster-RCNN	48.5	35.9	36.3	20.5	42.9	25
RetinaNet	37.4	36.0	29.9	18.1	85.6	31
YOLOv5s	48.4	36.2	38.4	19.7	17.5	122
Improved YOLOv7	56.2	**45.4**	44.6	26.6	120.1	89
YOLOv8s	48.5	37.8	38.2	21.3	28.4	126
YOLOv10s	49.4	38.7	38.8	22.4	29.1	133
YOLOv11s	50.2	38.4	39.6	22.9	29.3	138
PP-YOLOE	54.0	41.1	40.2	24.8	**15.5**	121
ESO-DETR	56.9	42.2	44.7	27.1	58.9	128
MIS-YOLOv8	56.5	41.3	45.9	27.8	37.3	99
Ours	**58.3**	43.1	**46.8**	**28.9**	30.8	**142**

提高 8.4%，mAP@0.5:0.95 提高 9.2%，FPS 高出 20，处理速度更快。相较于最新 YOLO 系列中的 YOLOv8s、YOLOv10s 与 YOLOv11s，本文算法的 mAP@0.5 分别高出 8.6%、8.0% 和 7.2%，且在 Precision、Recall 及 mAP@0.5:0.95 方面也保持稳定领先优势，应对小目标复杂场景时具有更强的检测表现。虽然 Improved YOLOv7 的 Recall 较本文算法更高，mAP@0.5 相近，但本文算法的 GFLOPs 仅为其 1/4，FPS 高出 53，表现出更优的轻量化特性。与轻量化模型 PP-YOLOE 相比，所提算法的 Precision 提高了 4.3%，mAP@0.5 提升 6.6%，FPS 提高了 21，兼顾高效率和高精度。与无人机专用的检测算法 ESO-DETR 相比，mAP@0.5 提升 2.1%，Precision 和 Recall 也均有提高，对小目标检测效果显著。与最近公开发表的改进模型 MIS-YOLOv8 相比，Precision 提升了 1.8%，FPS 高出 43，在检测精度和运算效率方面均有优势。

此外，图 8 表示 YOLOv11s 和改进模型在相同实验条件和参数下的训练过程中各项指标变化对比情况。实验结果表明，与 YOLOv11s 相比，改进模型在检测性能上有显著提升，各项指标均优于基线模型。

综上所述，本文提出的改进算法在 VisDrone2019 数据集上的检测性能明显优于其他热门目标检测算法，在提高检测精度的同时，有效兼顾了计算效率和实时性。展现出在多尺度目标检测，尤其是小目标检测中的显著优势，有效改善漏检、误检问题，提供更好的性能保障。

本文从 VisDrone2019 数据集中选取“密集小目标”、“俯拍多尺度”和“暗光遮挡”三种具有挑战性的检测场景，将代表性的 4 种算法 (Faster-RCNN、YOLOv11s、ESO-DETR、MIS-YOLOv8) 与本文算法的测试结果进行可视化展示，结果如图 9 所示。由此可见，与 YOLOv11s 相比，改进算法未产生与其相同的误检情况，且检测出了更多的 Pedestrain 和 Car 类别；MIS-YOLOv8 算法在密集条件下对小目标的检测效果不佳，易产生漏检问题，而本文算法有效解决了此问题；与其他算法相比，本文算法在三种场景下的综合检测性能均表现最佳，出色完

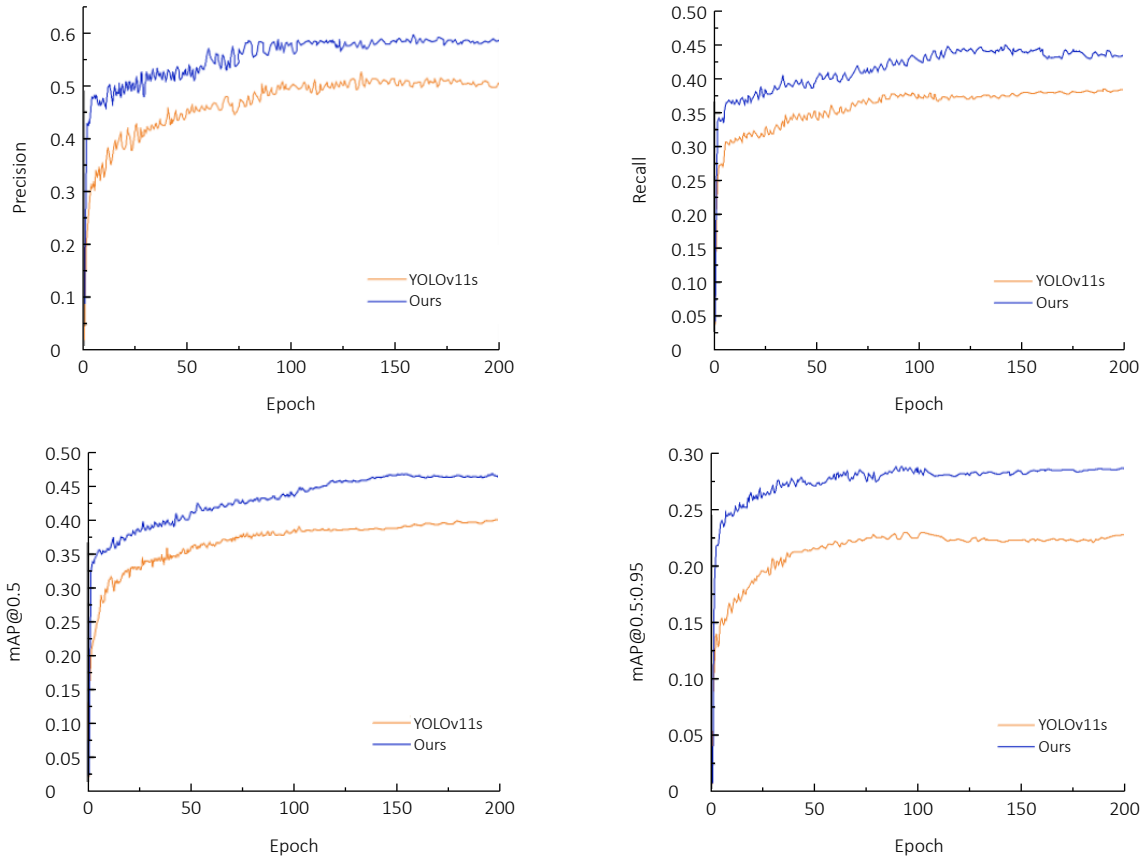


图 8 改进模型与 YOLOv11s 评估指标对比

Fig. 8 Comparison of evaluation indicators between the improved model and YOLOv11s

成了检测任务，能够在各种场景下准确识别和确定目标位置，显著减少误检和漏检情况。然而，在密集场景中对 Pedestrain 和 People 两类的检测效果易产生混淆，有待进一步加强。

为进一步检验改进算法在不同数据环境下的泛化能力，以及全面评估算法应对复杂情况的鲁棒性，本文在 VisDrone2021 数据集上开展了多算法对比实验。实验结果见表 4 所示。结果表明，本文提出的改进算法检测性

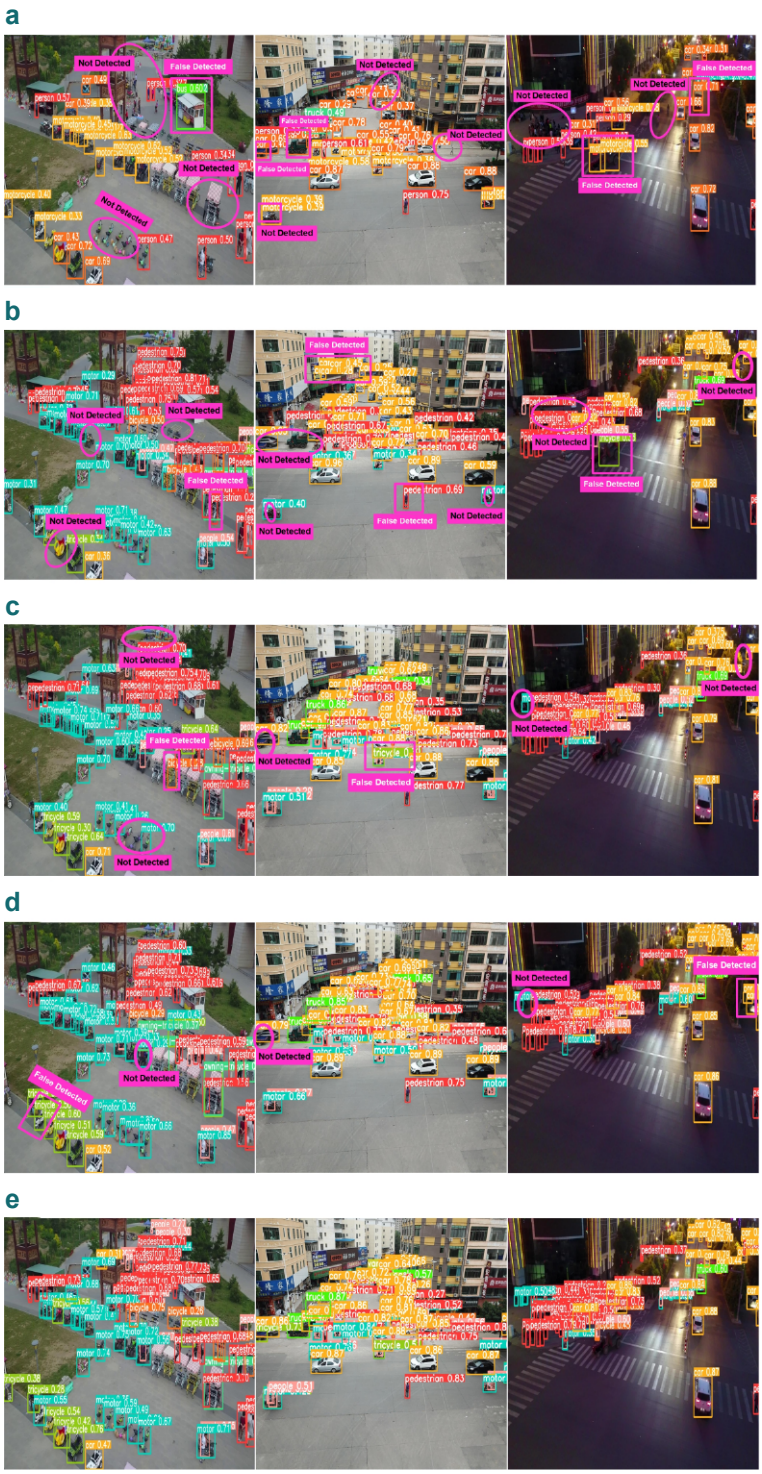


图 9 对比试验可视化结果。(a) Faster-RCNN; (b) YOLOv11s; (c) ESO-DETR; (d) MIS-YOLOv8; (e) Ours

Fig. 9 Comparison test visualisation results. (a) Faster-RCNN; (b) YOLOv11s; (c) ESO-DETR; (d) MIS-YOLOv8; (e) Ours

表 4 VisDrone2021 数据集对比实验

Table 4 Comparison experiments on the VisDrone2021 dataset

Model	Precision/%	Recall/%	mAP@0.5/%	mAP@0.5:0.95/%	GFLOPs	FPS
SSD	43.5	32.2	30.6	13.9	20.5	105
Faster-RCNN	46.1	33.5	34.7	18.8	42.9	23
RetinaNet	45.8	33.7	33.6	15.9	31.7	31
YOLOv5s	46.2	34.3	36.1	17.5	**15.5**	119
Improved YOLOv7	53.8	**43.6**	41.4	23.3	16.9	98
YOLOv8s	47.4	36.8	35.9	18.7	28.4	121
YOLOv10s	48.1	35.5	36.4	20.3	29.1	128
YOLOv11s	48.8	36.1	37.9	20.5	29.3	132
PP-YOLOE	52.1	38.7	38.3	22.4	18.5	120
ESO-DETR	54.2	39.0	41.5	25.6	59.0	124
MIS-YOLOv8	53.6	37.9	42.9	26.9	41.1	92
Ours	**55.9**	40.8	**44.7**	**28.4**	31.4	**139**

能明显优于其他对比算法，证明了改进算法的有效性和实时性。

综上所述，本文提出的改进模型在解决无人机航拍图像目标检测任务方面表现出卓越的检测精度和实时性能。该模型针对无人机平台的资源限制进行了优化，确保高效的计算能力和较低的延迟，使其能够在嵌入式设备上顺利部署，满足实际应用需求。

4 结 论

针对无人机航拍图像在复杂背景下对小目标的检测精度过低以及多尺度变化导致的漏检、误检问题，本文提出语义引导与多维特征感知的无人机航拍图像目标检测算法。通过设计自适应多维特征采样单元，对各组子特征进行多路径交互并行处理，增强网络的特征提取能力；提出 SPAPF 池化层，利用 Adapool 机制，优化传统空间金字塔池化层 SPPF，保持图像细节的同时减少冗余信息，降低计算量；构建跨空间全局信息整合模块，对全局上下文进行建模，减少小目标与背景或其他目标的混淆，改善误检情况；利用语义引导特征融合模块，实现深层语义对浅层特征的引导，利用高维通道信息高效恢复空间细节，使融合后的特征图在分辨率和语义表达上均得到改善，极大提升小目标的检测精度。在公开数据集 VisDrone2019 和 VisDrone2021 上开展实验，结果表明，本文算法相较于基线模型，在检测精度与速度方面均有显著提升，对复杂背景下的密集小目标检测效果较好，有效减少漏检、误检问题。未来研究中将进一步探索和优化该算法，以应对更加复杂的目标检测场景，获得更好的检测性能。

作者贡献

王光：提出核心方法，完成模型设计、实验开展与论文初稿撰写；李松泽：指导研究思路，把控学术方向，负责经费与论文定稿；李英硕：协助完成实验复现与论文修改工作。

利益冲突

所有作者声明无利益冲突

参考文献

1. Liang L M, Chen K Q, Wang C B, et al. Remote sensing image detection algorithm integrating visual center mechanism and parallel patch perception[J]. *Opto-Electron Eng*, 2024, **51**(7): 240099.
梁礼明, 陈康泉, 王成斌, 等. 融合视觉中心机制和并行补丁感知的遥感图像检测算法[J]. 光电工程, 2024, **51**(7): 240099.
2. Xiao Z J, Lin B H, Qu H C. SAR ship detection with multi-mechanism fusion[J]. *J Image Graphics*, 2024, **29**(2): 545−558.
肖振久, 林渤翰, 曲海成. 融合多重机制的 SAR 舰船检测[J]. 中国图象图形学报, 2024, **29**(2): 545−558.
3. Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]// *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014: 580–587.
https://doi.org/10.1109/CVPR.2014.81.
4. Ren S Q, He K M, Girshick R, et al. Faster R-CNN: towards real-

time object detection with region proposal networks[J]. *IEEE Trans Pattern Anal Mach Intell*, 2017, **39**(6): 1137−1149.

5. Zhang R M, Xiao Y F, Jia Z N, et al. Improved YOLOv7 algorithm for target detection in complex environments from UAV perspective[J]. *Opto-Electron Eng*, 2024, **51**(5): 240051.
张润梅, 肖钰霏, 贾振楠, 等. 改进 YOLOv7 的无人机视角下复杂环境目标检测算法[J]. 光电工程, 2024, **51**(5): 240051.
6. Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 779–788. https://doi.org/10.1109/CVPR.2016.91.
7. Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C]//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017: 6517–6525. https://doi.org/10.1109/CVPR.2017.690.
8. Redmon J, Farhadi A. YOLOv3: an incremental improvement[Z]. arXiv: 1804.02767, 2018. https://doi.org/10.48550/arXiv.1804.02767.
9. Ge Z, Liu S T, Wang F, et al. YOLOX: exceeding YOLO series in 2021[Z]. arXiv: 2107.08430, 2021. https://doi.org/10.48550/arXiv.2107.08430.
10. Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//*Proceedings of the IEEE International Conference on Computer Vision*, 2017: 2999–3007. https://doi.org/10.1109/ICCV.2017.324.
11. Liu W, Anguelov D, Erhan D, et al. SSD: single shot MultiBox detector[C]//*14th European Conference on Computer Vision*, 2016: 21–37. https://doi.org/10.1007/978-3-319-46448-0_2.
12. Xiao Z J, Wu Z W, Zhang J H, et al. Adaptive foreground focusing for target detection in UAV aerial images[J]. *Opto-Electron Eng*, 2024, **51**(9): 240149.
肖振久, 吴正伟, 张杰浩, 等. 自适应前景聚焦无人机航拍图像目标检测[J]. 光电工程, 2024, **51**(9): 240149.
13. Sahin O, Ozer S. YOLODrone: improved YOLO architecture for object detection in drone images[C]//*2021 44th International Conference on Telecommunications and Signal Processing (TSP)*, 2021: 361–365. https://doi.org/10.1109/TSP52935.2021.9522653.
14. Mao G T, Deng T M, Yu N J. Object detection in UAV images based on multi-scale split attention[J]. *Acta Aeronaut Astronaut Sin*, 2023, **44**(5): 326738.
冒国韬, 邓天民, 于楠晶. 基于多尺度分割注意力的无人机航拍图像目标检测算法[J]. 航空学报, 2023, **44**(5): 326738.
15. Chen J G, Wen R H, Ma L L. Small object detection model for UAV aerial image based on YOLOv7[J]. *Signal, Image and Video Processing*, 2024, **18**(3): 2695−2707.
16. Wang G, Chen Y F, An P, et al. UAV-YOLOv8: a small-object-detection model based on improved YOLOv8 for UAV aerial photography scenarios[J]. *Sensors*, 2023, **23**(16): 7190.
17. Huang M, Zhang Y Y, Chen Y Z. Small target detection model in aerial images based on TCA-YOLOv5m[J]. *IEEE Access*, 2023, **11**: 3352−3366.
18. Fang W X, Zhang G Q, Zheng Y H, et al. Multi-task learning for UAV aerial object detection in foggy weather condition[J]. *Remote Sensing*, 2023, **15**(18): 4617.
19. Betti A, Tucci M. YOLO-S: a lightweight and accurate YOLO-like network for small target detection in aerial imagery[J]. *Sensors*, 2023, **23**(4): 1865.
20. Qiu X Y, Chen Y J, Cai W H, et al. LD-YOLOv10: a lightweight target detection algorithm for drone scenarios based on YOLOv10[J]. *Electronics*, 2024, **13**(16): 3269.
21. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: transformers for image recognition at scale[C]//*9th International Conference on Learning Representations*, 2021.
22. Stergiou A, Poppe R. AdaPool: exponential adaptive pooling for information-retaining downsampling[J]. *IEEE Trans Image Process*, 2023, **32**: 251−266.
23. Guo Y H, Li Y D, Wang L Q, et al. Depthwise convolution is all you need for learning multiple visual domains[C]//*Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, 2019: 8368–8375. https://doi.org/10.1609/aaai.v33i01.33018368.
24. Yin X Y, Goudriaan J, Lantinga E A, et al. A flexible sigmoid function of determinate growth[J]. *Ann Bot*, 2003, **91**(3): 361−371.
25. Wu Y X, He K M. Group normalization[C]//*Proceedings of the 15th European Conference on Computer Vision (ECCV)*, 2018: 3–19. https://doi.org/10.1007/978-3-030-01261-8_1.
26. Dice L R. Measures of the amount of ecologic association between species[J]. *Ecology*, 1945, **26**(3): 297−302.
27. Sorensen T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons[J]. *Biol. Skrifter*, 1948, **5**: 1–34.
28. Bieder F, Sandkühler R, Cattin P C. Comparison of methods generalizing max- and average-pooling[Z]. arXiv: 2103.01746, 2021. https://doi.org/10.48550/arXiv.2103.01746.
29. Tan H C, Liu X P, Yin B C, et al. MHSA-Net: Multihead self-attention network for occluded person re-identification[J]. *IEEE Trans Neural Networks Learn Syst*, 2023, **34**(11): 8210−8224.
30. Shi W Z, Caballero J, Huszár F, et al. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network[C]//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 1874–1883. https://doi.org/10.1109/CVPR.2016.207.
31. Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: 7132–7141. https://doi.org/10.1109/CVPR.2018.00745.
32. Lin M, Chen Q, Yan S C. Network in network[Z]. arXiv: 1312.4400, 2013. https://doi.org/10.48550/arXiv.1312.4400.
33. Du D W, Zhu P F, Wen L Y, 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 Workshop*, 2019: 213–226. https://doi.org/10.1109/ICCVW.2019.00030.
34. Cao Y R, He Z J, Wang L J, et al. VisDrone-DET2021: the vision meets drone object detection challenge results[C]//*Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021: 2847–2854. https://doi.org/10.1109/ICCVW54120.2021.00319.
35. Zhao H Y, Zhang H P, Zhao Y Y. YOLOv7-sea: Object detection of maritime UAV images based on improved YOLOv7[C]//*Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023: 233–238. https://doi.org/10.1109/WACVW58289.2023.00029.

36. Wang A, Chen H, Liu L H, et al. YOLOv10: real-time end-to-end object detection[C]//*Proceedings of the 38th International Conference on Neural Information Processing Systems*, 2024: 3429.
37. Xu S L, Wang X X, Lv W Y, et al. PP-YOLOE: an evolved version of YOLO[Z]. arXiv: 2203.16250, 2022. https://doi.org/10.48550/arXiv.2203.16250.
38. Liu Y F, He M, Hui B. ESO-DETR: an improved real-time detection transformer model for enhanced small object detection in UAV imagery[J]. *Drones*, 2025, **9**(2): 143.
39. Sun T, Yang S Q, Liu H Y, et al. MIS-YOLOv8: an improved algorithm for detecting small objects in UAV aerial photography based on YOLOv8[J]. *IEEE Trans Instrum Meas*, 2025, **74**: 5020212.

作者简介



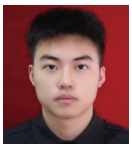
王光 (1979-)，男，山东邹城人，副教授，硕士研究生导师，主要从事图像与视觉信息计算、智能数据处理方面的研究。

E-mail：275469783@qq.com



【通信作者】李松泽 (2002-)，女，辽宁营口人，硕士研究生，主要从事图像与视觉信息计算方面的研究。

E-mail：2867314548@qq.com



李英硕 (2002-)，男，辽宁辽阳人，硕士研究生，主要从事医学图像、数据挖掘方面的研究。

E-mail：867472310@qq.com

Semantic-guided and multi-dimensional feature perception for target detection in UAV aerial images

Wang Guang, Li Songze*, Li Yingshuo

School of Software, Liaoning University of Engineering and Technology, Huludao, Liaoning 125105, China

Abstract

Objective UAV aerial images have been widely applied in military surveillance, emergency rescue, agricultural monitoring, and other fields due to their advantages of flexible shooting angles and rich spatial information acquisition. However, target detection in such images faces prominent challenges: extreme scale variations of targets, dense spatial distribution of small targets, and frequent false or missed detections caused by complex backgrounds. Additionally, real-time application scenarios impose strict requirements on the inference speed of detection models, which makes it difficult for existing algorithms to balance detection accuracy and computational efficiency. To address these issues, this study aims to propose a high-precision and efficient UAV aerial image target detection method, which enhances the model's ability to perceive multi-scale targets, especially small targets, reduces false and missed detections, and ensures real-time performance to meet practical application needs.

Methods Based on the YOLOv11s baseline model, this study designs a multi-module collaborative optimization framework, and the specific technical methods are as follows: First, an Adaptive Multi-dimensional Feature Sampling (AMFS) unit is embedded in the backbone network. The input feature maps are grouped along the channel dimension, and three parallel paths are used for multi-scale feature extraction. Adaptive average pooling, depthwise convolution, and spatial attention mechanisms are integrated to realize the adaptive synergy fusion of local details and global contextual information, which makes up for the limitation of fixed convolution receptive fields and enhances the feature extraction ability for small targets. Second, a Spatial Pyramid Adaptive Pooling Fusion (SPAPF) layer is constructed to replace the traditional SPPF module. This layer combines smoothed approximate average pooling and smoothed approximate maximum pooling through learnable weight masks. It not only retains structural details and texture information of targets but

Foundation item: National Natural Science Foundation of China (62173171)

*Correspondence: Li S Z, E-mail: 2867314548@qq.com

also reduces redundant computations, thus accelerating the model's inference speed while maintaining lightweight characteristics. Third, a Cross-Spatial Contextual Parallel Self-Attention (CS_C2PSA) module is proposed. It introduces cross-spatial multi-head self-attention (CSL-MHSA) and phased residual connection structures. By associating local and global receptive fields through 1×1 and 3×3 convolutions, and integrating vertical and horizontal spatial position information, the module strengthens the model's global contextual perception ability, alleviates the confusion between small targets and backgrounds or adjacent targets, and improves the false detection problem. Finally, a Semantic-guided Feature Fusion (SGFF) module and a dedicated Small Object Detection (SOD_Detect) head are designed. The SGFF module adopts a cross-layer feature alignment strategy and sub-pixel convolution technology. It uses the SE channel attention mechanism to enhance the expression of deep semantic information in shallow features, realizes the efficient fusion of shallow detail features and deep semantic features, and generates high-resolution feature maps with rich semantics for the SOD_Detect head, which significantly improves the positioning and classification accuracy of small targets. The model is trained and tested on VisDrone2019 and VisDrone2021 datasets, with batch_size set to 16 and epochs to 200. The experimental environment is Windows 10 operating system, NVIDIA GeForce RTX 3090 GPU, PyTorch 1.10.1 + CUDA 11.1 deep learning framework, and Python 3.8 programming language.

Results and Discussions Ablation experiments on the VisDrone2019 dataset show that each improved module plays a positive role in performance enhancement: The AMFS unit increases the mAP@0.5 by 3.2%, the SPAPF layer reduces the GFLOPs by 4.4 while improving the Recall by 2.2%, the CS_C2PSA module increases the Precision by 2.3% and the mAP@0.5 by 2.0%, and the SGFF module improves the Precision by 3.9% and the mAP@0.5 by 3.4%. When all four modules are integrated, the model's Precision reaches 58.3%, Recall 43.1%, and mAP@0.5 46.8%, which are 8.1%, 4.7%, and 7.2% higher than those of the baseline YOLOv11s, respectively, and the GFLOPs are maintained at 30.8, ensuring computational efficiency. Comparative experiments with 11 mainstream algorithms (including SSD, Faster-RCNN, YOLOv5s, YOLOv11s, etc.) on VisDrone2019 and VisDrone2021 datasets show that the proposed method achieves mAP@0.5 of 46.8% and 44.7% on the two datasets, respectively, which are 7.2% and 6.5% higher than the baseline model. Compared with the latest YOLOv10s and YOLOv8s, the mAP@0.5 is increased by 8.0% and 8.6% respectively on VisDrone2019 dataset. Compared with the UAV-specific detection algorithm ESO-DETR, the mAP@0.5 is improved by 2.1%, and the FPS is increased by 14. Even compared with the lightweight model PP-YOLOE, the proposed method still has advantages in both accuracy and speed, with Precision increased by 4.3% and FPS increased by 21. Visualization results of challenging scenarios such as dense small targets, aerial multi-scale, and low-light occlusion show that the proposed method effectively reduces false and missed detections, and has better detection performance for dense small targets than algorithms such as MIS-YOLOv8. However, there is still room for improvement in the distinction between Pedestrian and People categories in extremely dense scenarios, which is a direction for subsequent optimization.

Conclusions This study proposes a UAV aerial image target detection method based on semantic guidance and multi-dimensional feature perception. Through the collaborative optimization of AMFS, SPAPF, CS_C2PSA, and SGFF modules, the method effectively solves the key problems of large target scale variation, dense small target distribution, and high false and missed detection rates in complex backgrounds in UAV aerial image target detection. Experimental results on public datasets verify that the proposed method significantly improves the detection accuracy while maintaining efficient computational performance, and has excellent generalization ability and robustness. It provides a reliable technical solution for high-precision and real-time target detection in UAV aerial imagery. In future research, we will further optimize the module structure to enhance the model's ability to distinguish similar target categories in extremely dense scenarios, and explore the application of the method in more complex environments such as severe weather and heavy occlusion to expand its practical application scope.

Keywords: UAV aerial images; small object detection; adaptive multi-dimensional feature sampling; cross-spatial global information integration; semantic guided feature fusion

Wang G, Li S Z, Li Y S. Semantic-guided and multi-dimensional feature perception for target detection in UAV aerial images[J]. *Opto-Electron Eng*, 2026, **53**(2): 250123; DOI: 10.12086/oee.2026.250123

Opto-Electronic Journals Group
www.oejournal.org

扫描二维码，获取 PDF 全文