CN 51-1346/O4 ISSN 1003-501X (印刷版) ISSN 2094-4019 (网络版)

光电工程
Opto-Electronic Engineering

中文核心期刊 中国科技核心期刊
Scopus CSCD

自适应前景聚焦无人机航拍图像目标检测

肖振久，吴正伟，张杰浩，曲海成

引用本文:

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.

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.

https://doi.org/10.12086/oee.2024.240149

收稿日期: 2024-06-28; 修改日期: 2024-08-05; 录用日期: 2024-08-06

相关论文

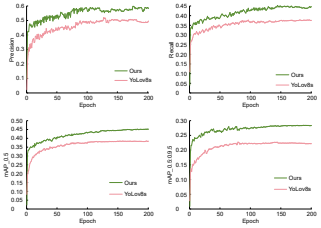
Opto-Electronic Engineering
光 电 工 程

Article

2024 年, 第 51 卷, 第 9 期

DOI: 10.12086/oee.2024.240149 CSTR: 32245.14.oee.2024.240149

自适应前景聚焦无人机航拍图像目标检测



肖振久，吴正伟*，张杰浩，曲海成

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前景目标尺度差异大、样本空间分布不均衡、背景冗余占比高所导致的漏检和误检问题，本文提出一种自适应前景聚焦无人机航拍图像目标检测算法。首先，构建全景特征细化分类层，通过重参数空间像素方差法及混洗操作，增强算法聚焦能力，提高前景样本特征的表示质量。其次，采用分离-学习-融合策略设计自适应双维特征采样单元，加强对前景焦点特征提取能力和背景细节信息保留能力，改善误检情况，加快推理速度。然后，结合多分支结构和广播自注意力机制构造多路径信息整合模块，解决下采样引起的歧义映射问题，优化特征的交互与整合，提高算法对多尺度目标的识别、定位能力，降低模型计算量。最终，引入自适应前景聚焦检测头，运用动态聚焦机制，增强前景目标检测精度，抑制背景干扰。在公开数据集 VisDrone2019 和 VisDrone2021 上进行相关实验，实验结果表明，该方法 mAP@0.5 数值达到了 45.1% 和 43.1%，较基线模型分别提升 6.6% 和 5.7%，且优于其他对比算法，表明该算法显著提升了检测精度，具备良好的普适性与实时性。

关键词：无人机航拍图像；全景特征细化分类；自适应双维特征采样；多路径信息整合；多尺度目标；动态聚焦

中图分类号：TP391.4 **文献标志码：**A

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

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

Adaptive foreground focusing for target detection in UAV aerial images

Xiao Zhenjiu, Wu Zhengwei*, Zhang Jiehao, Qu Haicheng

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

Abstract: To address the issues of missed and false detections caused by significant scale differences of foreground targets, uneven sample spatial distribution, and high background redundancy in UAV aerial images, an adaptive foreground-focused UAV aerial image target detection algorithm is proposed. A panoramic feature refinement classification layer is constructed to enhance the algorithm's focusing capability and improve the representation quality of foreground sample features through the re-parameterization spatial pixel variance method and shuffling operation. An adaptive dual-dimensional feature sampling unit is designed using a separate-learn-merge strategy to strengthen the algorithm's ability to extract foreground focus features and retain background detail information, thereby improving false detection situations and accelerating inference speed. A multi-path information integration module is constructed by combining a multi-branch structure and a broadcast self-attention

收稿日期：2024-06-28；修回日期：2024-08-05；录用日期：2024-08-06

基金项目：辽宁省高等学校基本科研项目 (LJKMZ20220699)；辽宁工程技术大学学科创新团队项目 (LNTU20TD-23)

*通信作者：吴正伟，1525545769@qq.com。

mechanism to solve the ambiguity mapping problem caused by downsampling, optimize feature interaction and integration, enhance the algorithm's ability to recognize and locate multi-scale targets, and reduce model computational load. An adaptive foreground-focused detection head is introduced, which employs a dynamic focusing mechanism to enhance foreground target detection accuracy and suppress background interference. Experiments on the public datasets VisDrone2019 and VisDrone2021 show that the proposed method achieves mAP@0.5 values of 45.1% and 43.1%, respectively, improving by 6.6% and 5.7% compared to the baseline model, and outperforming other comparison algorithms. These results demonstrate that the proposed algorithm significantly improves detection accuracy and possesses good generalizability and real-time performance.

Keywords: UAV aerial images; panoramic feature refinement classification; adaptive two-dimensional feature sampling; multi-path information integration; multi-scale objective; dynamic focusing

1 引　言

无人机航拍图像目标检测在违章建筑监测、线路巡查、军事侦察等多个领域中发挥着重要作用[1-2]，是遥感图像处理和目标检测的一个重要分支，其核心内容是从广域背景环境中准确识别出航拍目标的类别信息与位置信息。由于航拍图像的目标视角和尺度变动性大，同时受到光照和阴影、背景噪声等因素的影响，导致图像表征能力差，检测效果不佳。因此，提高无人机航拍图像检测的精度、加快检测速度成为了当下研究的重点与难点。

传统目标检测算法主要基于特征算子、颜色直方图以及浅层机器学习实现[3-4]，因其主观判断性高、检测精度低、鲁棒性匮乏、泛化能力差等问题，现已被深度学习检测算法所取代。基于卷积神经网络 (convolutional neural network，CNN) 的深度学习目标检测算法，实现了自动特征学习及多尺度特征表示，不再需要人工设计特征，解决了传统算法特征提取不充分的问题。现阶段，基于深度学习的目标检测算法依据是否需要区域建议可分为两类：单阶段目标检测模型、双阶段目标检测模型。

以 Faster R-CNN[5] 为代表的双阶段目标检测模型利用区域提议网络 (region proposal network，RPN) 生成候选目标框，随后采用检测网络对候选框进行检测和分类。双阶段检测模型提高了检测精度，但受结构、参数冗余的影响，其难以实现快速检测。代表性的单阶段算法包括 YOLO (you only look once) 系列[6-10]、SSD (single shot multibox detector)[11] 等，通过简单架构一次性关注所有空间区域完成目标的检测和分类，简化了检测流程，有效地缩短了训练和推理时间，在实时无人机目标检测领域具有更多优势。

近年来，无人机航拍图像目标检测算法得到了快速发展，众多高效检测算法被相继提出。Zhang 等[12] 提出了 FocSDet (focusing on small objects detector in aerial images) 算法，通过构建小目标特征聚合网络，提升了大目标以及密集遮挡小目标检测的精度，但模型泛化能力较差。Li 等[13] 在 YOLOv5 的基础上增加了小目标检测层，引入双向特征金字塔 (BiFPN) 融合不同尺度的特征信息，增强了小目标的特征提取能力，但算法推理速度较慢。Li 等[14] 在 YOLOv7 基础上，引入 SPPCSPC 模块和 Coordinate Attention 注意力机制，提高特征保留和动态卷积处理能力，提升了小目标检测精度。但在复杂背景下，仍存在漏检问题。Chen 等[15] 在 YOLOv8 模型中，集成 BiFormer 注意力机制和 FFNB (focal fasternet block) 特征处理模块，使浅层特征和深层特征得到充分融合。但该算法对小目标检测效果不佳。Zhu 等[16] 在基于 RT-DETR 的多尺度融合无人机目标检测中，提出了 GCD-DETR (gathering cascaded dilated DETR) 模型，使用 Dilated Re-param Block 和 Gather-and-Distribute 机制改进多尺度特征融合，并引入 Cascaded Group Attention 机制，提高了复杂场景中的检测性能。但在高动态场景中，模型的稳定性和鲁棒性不佳。Shao 等[17] 提出了轻量级 Aero-YOLO 目标检测算法，使用 GSConv 提高主干模型计算效率，并利用全局注意力机制增强特征提取能力。但在光照条件变化时，存在误检问题。Zhan 等[18] 在 YOLOv5 基础上集成了 Swin Transformer V2，通过使用 Focal-EIOU 改进 K-means 算法，提高了长距离小目标检测的准确性和效率。但模型计算复杂度较高，不利于在边缘设备上部署。陈朋磊等[19] 提出的 SFANet 网络及多元协同特征交互模板 MCFIM，有效抑制了小目标信息在网络中的流失，

并实现了相邻特征层之间的信息交互，丰富了小目标在网络中的特征表示。Sui 等[20] 在 YOLOv8 基础上提出 BDH-YOLO 算法，结合多头注意力机制提出动态检测头 DyHead (dynamic head)，显著提高了模型的检测能力，但在处理复杂背景时，稳定性较差，同时计算复杂度较高。

针对上述问题，本文基于 YOLOv8s 算法，提出了一种自适应前景聚焦的无人机航拍图像目标检测算法。首先，在主干网络中添加全景特征细化分类层，对前景和背景进行初步划分，加强网络对前景关键特征和背景细节特征的关注与判别能力。其次，采用分离-学习-融合策略，提出自适应双维特征采样单元。该单元使用可分离并行卷积 (separable parallel convolution，SPC) 和 SK (selective kernel) 软注意力机制[21]，加强网络对前景特征图中重要焦点特征的提取能力和小目标信息的保留能力，并降低模型计算量。然后，利用多分支结构和广播自注意力机制，充分融合多尺度特征信息与全局上下文信息，进一步提高模型的检测精度和推理速度，改善误检和漏检问题。最后，引入自适应前景聚焦检测头，实现对前景特征的智能聚焦和智能覆盖，进一步提高算法对前景多尺度目标的检测性能。

2 自适应前景聚焦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的漏检和误检问题，提高算法对多尺度样本的检测精度，本文以 YOLOv8s 为基线，提出了自适应前景聚焦无人机航拍图像目标检测算法。

基线模型 YOLOv8s 由主干网络 (Backbone)、颈部网络 (Neck)、检测头 (Head) 三部分构成。Backbone 采用 5 个卷积核为3×3，步长为 2 的 CBS 卷积单元与 4 个 c2f 模块对输入大小为640×640的彩色图像进行特征采样，输出三种不同尺度特征图。Neck 采用 1 个 SPPF 模块、4 个拼接单元 (Concat)、4 个 C2f 模块以及 2 个上采样操作单元 (Upsample) 与 2 个卷积核为3×3，步长为 2 的 CBS 卷积单元，对三组特征图进行进一步特征提取与融合，加强多尺度特征间的信息流动。Head 采用解耦头结构 (Decoupled-Head)[9]，对 Neck 输出的三组特征图进行单独分类与检测，得到输出大小为80×80×256、40×40×512、20×20×1024的预测特征图。

本文算法的网络结构如图 1 所示，较基线模型 YOLOv8s 做出四项改进。1) 添加全景特征细化分类层 (panoramic feature refinement classification，PFRC)：利用重参数空间像素方差法和特征混洗操作，增强网络聚焦能力与特征表示质量。2) 设计自适应双维特征采样单元 (adaptive two-dimensional feature sampling，ATFS) 替换部分 CBS 卷积模块：采用 SPC 卷积与 SK 注意力机制，进一步提高网络对关键特征提取能力及细节信息保留能力。3) 设计多路径信息整合 (multi-path full-text information integration module，MPFT) 模块替换部分 C2f 单元：结合多分支结构、广播自注意力机制、PFRC 与 ATFS，优化多尺度特征间的整合与交互，加强算法对关键样本的定位与识别能力。4) 设计自适应前景聚焦检测头 (adaptive foreground focusdetection head，AFF_Detect) 替换原模型检测头：通过动态聚焦机制，增强算法对焦点特征的检测能力，提高检测精度，抑制背景冗余。

2.1 全景特征细化分类

针对无人机航拍图像前景特征分布不均衡、背景冗余特征占比过高导致的误检、漏检等问题。本文构建全景特征分类层 PFRC，通过重参数空间像素方差法 (reparameterized spatial pixel variance method，RSPVM)[22]，动态评估全景特征图中不同分布特征的重要性，并依据所得评估分数重新加权特征图，以增强特征表示的质量和稳定性，完成对全景特征图前后背景的初步划分，使网络自适应聚焦于前景重要特征，减少处理背景冗余特征所需的额外资源消耗。考虑到背景中可能隐藏着某些关键的小目标映射信息，本文采用特征混洗操作对评估后的前后背景特征图进行随机匹配，以生成包含前景多尺度特征和背景细粒度特征的焦点特征图。PFRC 结构如图 2 所示，由重参数空间像素方差法和特征混洗操作构成。

1) 重参数空间像素方差法。首先对输入全景特征 $\boldsymbol{X}\in\mathbf{R}^{C\times H\times W}$ (其中 C 为特征通道数，H 和 W 为特征的高度与宽度) 进行组归一化 (group normalization，GN)[23] 处理，得到对应映射信息的权重系数。并根据权重系数，完成对特征映射 $\boldsymbol{X}$ 中不同特征信息内容的初步评估。评估过程如下所示：

$$GN(\boldsymbol{X})=\frac{\boldsymbol{X}-\mu}{\sqrt{\sigma^2+\varepsilon}}\gamma+\beta, \tag{1}$$

式中：μ为 $\boldsymbol{X}$ 的均值，σ为 $\boldsymbol{X}$ 的标准差，ε是一个很小的常数，防止除法为零，λ和β是可学习的仿射映

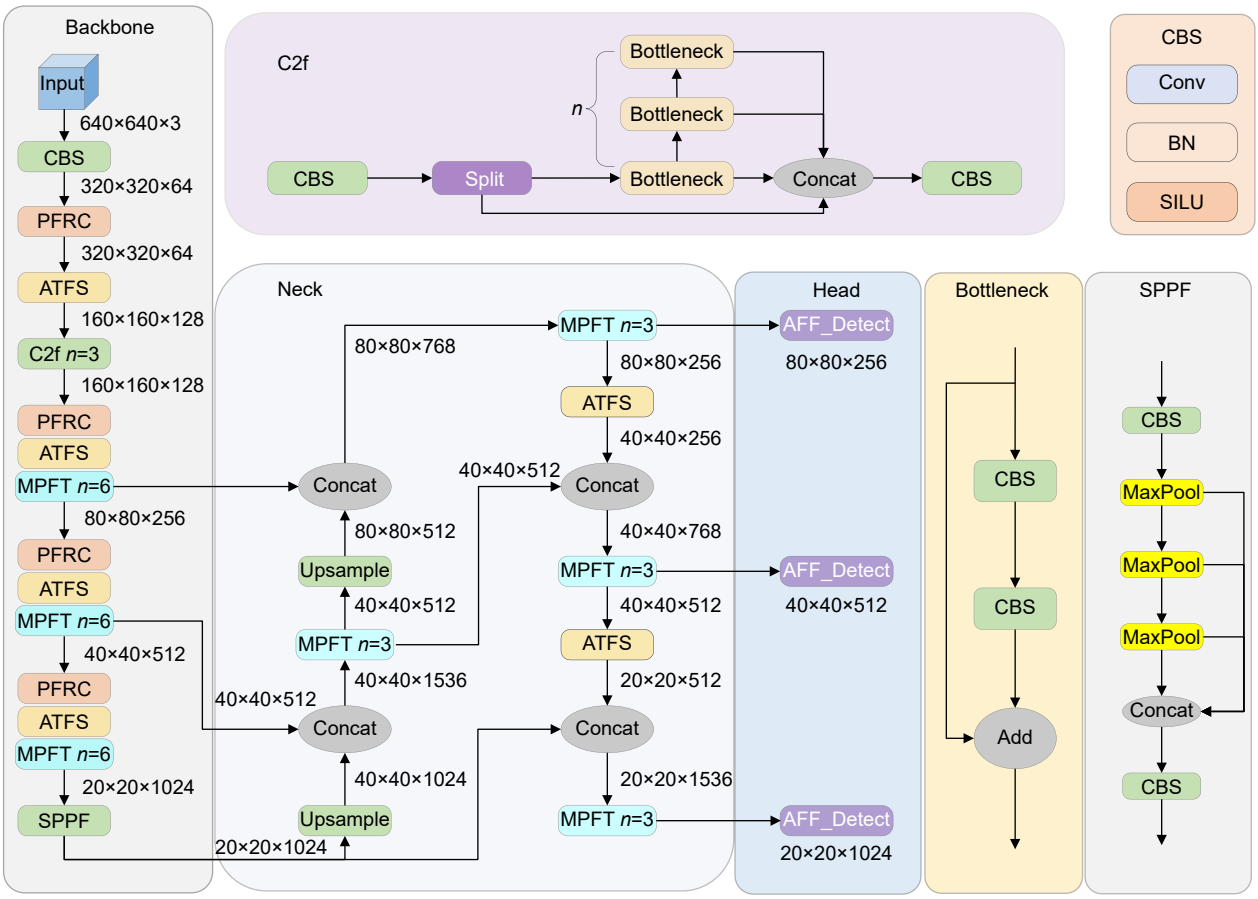


图 1　模型总体结构

Fig. 1　Overall model structure

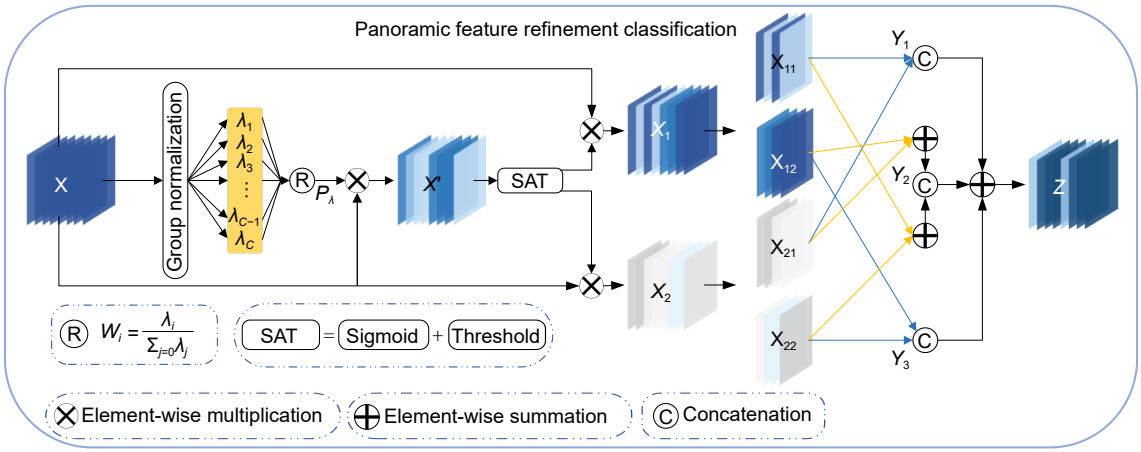


图 2　全景特征分类层 (PFRC) 结构

Fig. 2　Panoramic feature classification layer (PFRC) structure

射。本文将 GN 层中的标度因子λ作为衡量每个批次和通道空间像素方差的方法。更高的权重系数反映了更丰富的信息内容与空间像素的更多变化，从而导致更大的λ。随后对λ进行重参数化处理，得到表示不同映射信息重要性的相关权重$\boldsymbol{P}_\lambda \in \mathbf{R}^C$，公式如下:

$$\boldsymbol{P}_\lambda = \{W_i\} = \frac{\lambda_i}{\sum_{j=1}^{C}\lambda_j}, i, j = 1, 2, \cdots, C . \tag{2}$$

再利用 Sigmoid 激活函数[24]，将经重新加权后的特征映射$\boldsymbol{X}' \in \mathbf{R}^{C\times H\times W}$的权值映射到 (0, 1) 范围内，并

通过引入阈值进行门控[25]，以获取高于阈值 (阈值设置为：0.5) 的前景权重$\boldsymbol{P}_1$和低于阈值的背景权重$\boldsymbol{P}_2$。获取$\boldsymbol{P}$的过程可表示为

$$\boldsymbol{P} = Gate(Sigmoid(\boldsymbol{P}_\lambda(\boldsymbol{X}))), \tag{3}$$

$$\boldsymbol{X}_i = \boldsymbol{P}_i \times \boldsymbol{X}, i = 1,2. \tag{4}$$

最后将输入特征$\boldsymbol{X}$乘$\boldsymbol{P}_1$和$\boldsymbol{P}_2$，得到前景加权特征$\boldsymbol{X}_1$与背景加权特征$\boldsymbol{X}_2$。其中$\boldsymbol{X}_1$为具有高代表性、高表现性且信息丰富的前景空间内容，$\boldsymbol{X}_2$为信息贫乏的背景冗余内容。

2) 特征混洗。将细致划分后的加权特征$\boldsymbol{X}_1$、$\boldsymbol{X}_2$沿通道维度进行随机分割，得到特征$\boldsymbol{X}_{11} \in \mathbf{R}^{\alpha C \times H \times W}$，$\boldsymbol{X}_{12} \in \mathbf{R}^{(1-\alpha)C \times H \times W}$和 $\boldsymbol{X}_{21} \in \mathbf{R}^{(1-\alpha)C \times H \times W}$， $\boldsymbol{X}_{22} \in \mathbf{R}^{\alpha C \times H \times W}$，其中$\alpha(0 < \alpha < 1)$为分割因子，本文$\alpha$取值为 0.5。随后利用混洗操作，将两组特征中相同通道维度的特征进行相加，不同通道维度的特征进行拼接，以增强前后背景特征的信息交互，保留更多的背景关键信息。最后将分组混洗后的特征映射$\boldsymbol{Y}_1$、$\boldsymbol{Y}_2$、$\boldsymbol{Y}_3$进行融合，生成多尺度的细化前景焦点特征图$\boldsymbol{Z} \in \mathbf{R}^{C \times H \times W}$。该过程可表示为：

$$\begin{cases} \boldsymbol{X}_{11} \cup \boldsymbol{X}_{21} = \boldsymbol{Y}_1 \\ \boldsymbol{X}_{12} \cup \boldsymbol{X}_{22} = \boldsymbol{Y}_2 \\ \boldsymbol{X}_{11} \oplus \boldsymbol{X}_{22} \cup \boldsymbol{X}_{12} \oplus \boldsymbol{X}_{21} = \boldsymbol{Y}_3 \\ \boldsymbol{Y}_1 \oplus \boldsymbol{Y}_2 \oplus \boldsymbol{Y}_3 = \boldsymbol{Z} \end{cases}, \tag{5}$$

式中：∪为拼接，⊕为逐元相加。

2.2 自适应双维特征采样

为充分提取前景特征图$\boldsymbol{Z}$中多尺度目标及小目标的特征信息，本文提出了自适应双维特征采样单元 ATFS。该单元利用分离-采样-融合策略对卷积层进行重构，旨在减少特征学习过程中关键信息的流失，降低资源的消耗。ATFS 结构如图 3 所示。

1) 分离。首先对$\boldsymbol{Z}$其进行分组压缩，以提升算法运算效率。分组方法与上文一致，采用随机分割因子，得到特征$\boldsymbol{X}_1' \in \mathbf{R}^{\alpha C \times H \times W}$、$\boldsymbol{X}_2' \in \mathbf{R}^{(1-\alpha)C \times H \times W}$。对分组后的特征，利用核值为1×1卷积进行通道维度的压缩操作，并引入压缩因子L (L=2) 来控制通道数，以平衡算法的计算成本与检测速率，得到压缩特征$\boldsymbol{X}_1'' \in \mathbf{R}^{\frac{\alpha C}{L} \times H \times W}$、$\boldsymbol{X}_2'' \in \mathbf{R}^{\frac{(1-\alpha)C}{L} \times H \times W}$。

2) 学习。本文利用深度卷积 DWC (depthwise convolution)[26-27] 和点卷积 PWC (pointwise convilution)[28] 提出了可分离并行卷积 (separable parallel convolution，SPC)，结构如图 3 所示。SPC 使用 DWC 的单通道卷积核来提取和融合$\boldsymbol{X}_1''$各个通道的局部空间特征与边缘细节信息，提高模型对小目标的识别、定位能力。PWC 通过非线性组合$\boldsymbol{X}_1''$不同通道之间的特征，弥补 DWC 在捕捉全局信息方面的不足，增强多尺度目标的语义特征表示。并置连接的布局方式，能够保留原始特征细节信息，提高特征提取的丰富性和多样性，实现跨尺度特征学习和融合。此过程可表示为

$$\boldsymbol{Y}_1 = \boldsymbol{E}^{\mathrm{DWC}} \times \boldsymbol{X}_1'' + \boldsymbol{E}^{\mathrm{PWC}_1} \times \boldsymbol{X}_1'', \tag{6}$$

式中：$\boldsymbol{E}^{\mathrm{DWC}} \in \mathbf{R}^{\frac{\alpha C}{L} \times k \times k \times C}$、$\boldsymbol{E}_1^{\mathrm{PWC}} \in \mathbf{R}^{\frac{\alpha C}{L} \times 1 \times 1 \times C}$为 DWC 与 PWC_1的可学习权矩阵。

为丰富特征$\boldsymbol{Y}_1 \in \mathbf{R}^{C \times H \times W}$的表示，本文重用$\boldsymbol{X}_2''$与 PWC 操作生成的浅层隐藏映射信息进行拼接，作为$\boldsymbol{Y}_1$详细信息补充。该过程可表示为：

$$\boldsymbol{Y}_2 = \boldsymbol{E}^{\mathrm{PWC}_2} \times \boldsymbol{X}_2'' \cup \boldsymbol{X}_2'', \tag{7}$$

式中：$\boldsymbol{E}^{\mathrm{PWC}_2} \in \mathbf{R}^{\frac{(1-\alpha)C}{L} \times 1 \times 1 \times \left(1-\frac{1-\alpha}{L}\right)C}$为$\mathrm{PWC}_2$的可学习权矩阵，$\boldsymbol{Y}_2 \in \mathbf{R}^{C \times H \times W}$。

3) 融合。为自适应融合并输出$\boldsymbol{Y}_1$、$\boldsymbol{Y}_2$间跨尺度映射信息，本文利用简化 SK 注意力机制，通过全局

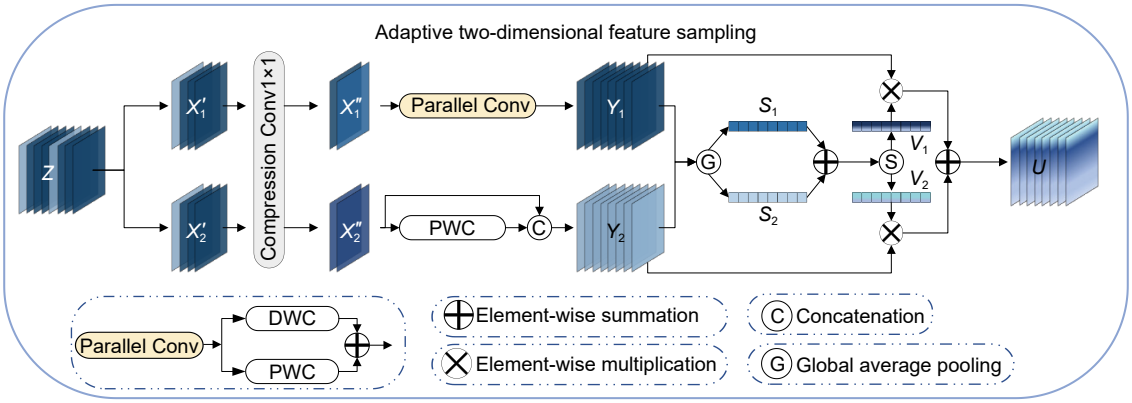


图 3 自适应双维特征采样 (ATFS) 结构

Fig. 3 Adaptive two-dimensional feature sampling (ATFS) structure

信息抽取和注意力分配来动态调整不同尺度特征的权重，以提高网络对前景目标聚焦和检测能力。首先，应用全局平均池化 (global average pooling, GAP)[29] 统计$\boldsymbol{Y}_1$、$\boldsymbol{Y}_2$各个通道的全局空间信息$\boldsymbol{S}_m \in \mathbf{R}^{C\times1\times1}$。计算过程如下：

$$\boldsymbol{S}_m = Pooling(\boldsymbol{Y}_m) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{H}\boldsymbol{Y}_c(i,j), m=1,2\,. \quad (8)$$

随后，使用 softmax 对$\boldsymbol{S}_1$、$\boldsymbol{S}_2$逐元相加后的特征向量进行归一化处理，得到对应通道注意力信息$\boldsymbol{V}_1, \boldsymbol{V}_2 \in \mathbf{R}^C$。并在$\boldsymbol{V}_1$、$\boldsymbol{V}_2$的指导下，将$\boldsymbol{Y}_1$、$\boldsymbol{Y}_2$沿通道进行合并，得到通道精细特征$\boldsymbol{U} \in \mathbf{R}^{C\times H\times W}$。

$$\boldsymbol{V}_1 = \frac{\mathrm{e}^{s_1}}{\mathrm{e}^{s_1}+\mathrm{e}^{s_2}}, \boldsymbol{V}_2 = \frac{\mathrm{e}^{s_2}}{\mathrm{e}^{s_1}+\mathrm{e}^{s_2}}, \boldsymbol{V}_1+\boldsymbol{V}_2 = 1\,, \quad (9)$$

$$\boldsymbol{U} = \boldsymbol{V}_1\boldsymbol{Y}_1 + \boldsymbol{V}_2\boldsymbol{Y}_2\,. \quad (10)$$

2.3 多路径全文信息整合

为捕捉前景特征间的长程依赖关系并保留更多的全局上下文信息，本文提出了一种高效的多路径全文信息整合模块 MPFT。该模块利用多路径、残差连接及 Bottleneck 瓶颈单元，对全局特征映射的关键特征进行逐尺度跨层交互与融合，进一步提炼代表性前景特征信息，同时缓解因特征采样产生的歧义映射问题。此外，通过在 Bottleneck 单元中添加轻量级广播自注意力机制 BSA (broadcast self-attention)[30] 对背景所含的全局上下文信息进行整合，以稳定前景布局并丰富前景焦点特征，有效改善因缺失上下文信息导致的漏检问题。MPFT 结构如图 4 所示。

为获取背景上下文特征信息，Transformer 的多头自注意力机制 (multihead self-attention，MSA)[31] 需利用复杂运算遍历多维像素。与前者不同，BSA 只需对 Patch Embeddings[32] 的位置编码 Position Encoding 进行一维遍历，并利用广播运算在全局信息传输中的优势，将计算复杂度由$O(N^2)$降至$O(N)$。BSA 结构如图 5 所示。

具体而言，BSA 接受 K 个 N 维 Patch Embeddings 组成的输入$\boldsymbol{X} \in \mathbf{R}^{K\times N}$，并计算出对应的位置分数$\boldsymbol{\tau}$、键矩阵 $\boldsymbol{k}$ 和值矩阵 $\boldsymbol{v}$。计算如下：

$$\begin{aligned}&\boldsymbol{\tau} = SoftMax(\boldsymbol{W}^{\tau}\times\boldsymbol{X}), \boldsymbol{k} = \boldsymbol{W}^{k}\times\boldsymbol{X},\\ &\boldsymbol{v} = ReLU(\boldsymbol{W}^{v}\times\boldsymbol{X}), \boldsymbol{\tau}\in\mathbf{R}^{K\times1}, \boldsymbol{k}\in\mathbf{R}^{K\times N},\\ &\boldsymbol{v}\in\mathbf{R}^{K\times N}, \boldsymbol{W}^{\tau}\in\mathbf{R}^{N\times1}, \boldsymbol{W}^{k}\in\mathbf{R}^{N\times N}, \boldsymbol{W}^{v}\in\mathbf{R}^{N}\,, \end{aligned} \quad (11)$$

式中：$ReLU$用于特征的非线性映射激活函数；$\boldsymbol{W}^{\tau}$、$\boldsymbol{W}^{k}$、$\boldsymbol{W}^{v}$为1×1卷积的权重值。$\boldsymbol{W}^{k}$和$\boldsymbol{W}^{v}$用于特征映射过程；$\boldsymbol{W}^{\tau}$用于拟合目标的特征分布，输出每个位置信息的重要性，并通过 SoftMax 归一化获得位置得分$\boldsymbol{\tau}$。随后利用广播操作实现τ和键矩阵 $\boldsymbol{k}$ 的信息传输，得到详细的注意力权重$\boldsymbol{\gamma}\in\mathbf{R}^{1\times N}$，再将$\boldsymbol{\gamma}$传输到值矩阵 $\boldsymbol{v}$ 中，获取输入图像的背景上下文信息。该过程表示为

$$\boldsymbol{\gamma} = \sum_{i=1}^{K}(\boldsymbol{\tau}_i \otimes \boldsymbol{k}_i)\,, \quad (12)$$

$$\boldsymbol{Y} = \sum_{j=1}^{N}(\boldsymbol{\gamma}_j \otimes \boldsymbol{V}_j)\times\boldsymbol{W}^{\mathrm{T}}\,, \quad (13)$$

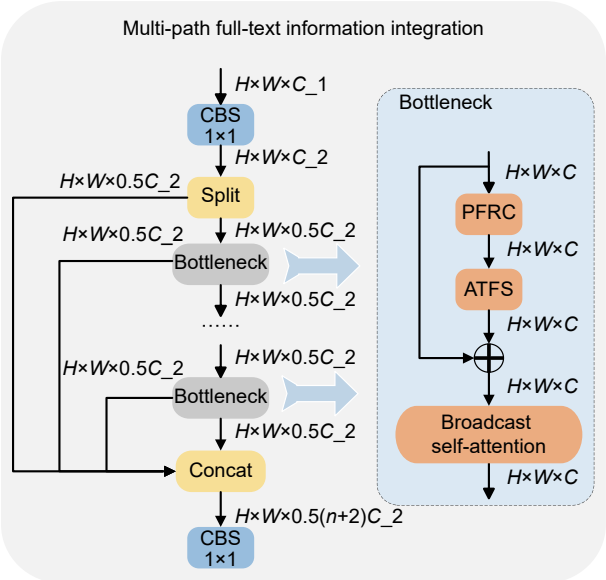


图 4 多路径全文信息整合 (MPFT) 结构

Fig. 4 Structure of multi-path full-text information integration (MPFT)

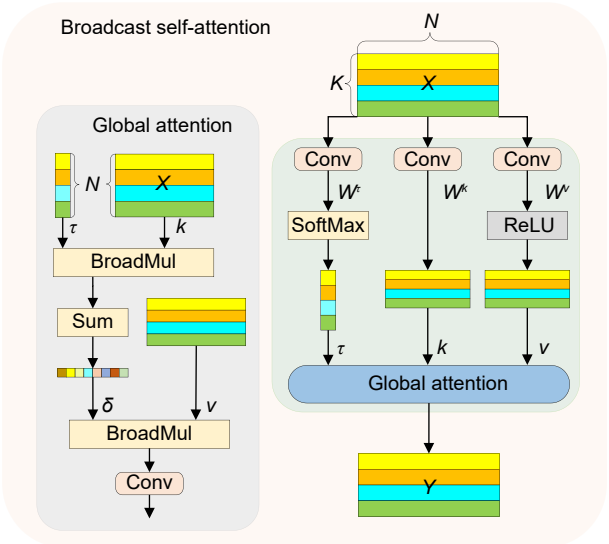


图 5 广播自注意力机制 (BSA) 结构

Fig. 5 Broadcast self-attention (BSA) mechanism structure

式中：⊗为广播操作；$\boldsymbol{W}^{\mathrm{T}} \in \mathbf{R}^{N\times N}$为卷积权值；$\boldsymbol{Y} \in \mathbf{R}^{K\times N}$为输出特征。

2.4 自适应前景聚焦检测头

稀疏卷积 (Sparse convolution, SC)[33] 根据可学习掩码对前景区域内焦点特征进行选择采样，以降低计算成本并加快推理速度，这对无人机高分辨率图像的检测任务非常有利。但固定的掩码比无法满足对前景区域多尺度特征的精准聚焦，同时由于缺少对背景中关键上下文信息的学习，使得检测结果在前景区域存在较大波动。因此本文对 SC 进行两部分优化，提出 SC_FCI (adaptive sparse convolution module for fusing contextual information) 模块，结构如图 6 所示；并利用该模块进一步优化检测头，所得自适应前景聚焦检测头 AFF_Detect 如图 7 所示。

对于不同尺度的特征图，本文通过构建 AM (adaptive masking) 模块以完成对前景关键特征的动态聚焦及智能覆盖，以此提高检测精度与推理效率。

具体而言，AM 利用标签分配技术[34] 对不同 PAN 层输出特征所含的真值标签 ground-truth 进行分类，得到第 i 层真实的分类结果$\boldsymbol{C}_i \in \mathbf{R}^{h_i\times w_i\times c}$，其中 c 表示背景内所有类别的数量，h_i和w_i分别表示特征图的高度和宽度。再利用分类结果评估出最佳掩码比P_i。

$$P_i = \frac{Pos(C_i)}{Numel(C_i)}, i = 1,2,3\,, \tag{14}$$

式中：$Pos(C_i)$表示前景像素量，$Numel(C_i)$表示所有像素数量。再利用焦点损失函数迫使掩码矩阵$\boldsymbol{H}_i$自适应生成与P_i相同的最佳掩码比。损失为

$$\mathcal{L}_{AM} = \frac{1}{L}\sum_i \left(\frac{Pos(\boldsymbol{H}_i)}{Numel(\boldsymbol{H}_i)} - P_i\right)^2, \tag{15}$$

式中：L 表示 PAN 总层数，$\boldsymbol{H}_i \in \{0,1\}^{B\times 1\times H\times W}$是由 Gumbel-Softmax[35] 对软特征$\boldsymbol{S}_i \in \mathbf{R}^{B\times 1\times H\times W}$计算得到的掩码矩阵。

$$\boldsymbol{H}_i = \begin{cases} \sigma(\boldsymbol{S}_i + g_1 - g_2) > 0.5, & \text{For training} \\ \boldsymbol{S}_i > 0, & \text{For inference} \end{cases}, \tag{16}$$

式中：$\sigma(\cdot)$表示 sigmoid 函数，$\boldsymbol{S}_i$由掩码网络$\boldsymbol{W}_{\mathrm{mask}} \in \mathbf{R}^{C\times 1\times 3\times 3}$对给定 PAN 第 i 层特征映射$\boldsymbol{X}_i \in \mathbf{R}^{B\times C\times H\times W}$卷积所得的软特征，$g_1$、$g_2$为随机 gumbel 噪声。

为弥补稀疏卷积造成的上下文信息流失，稳定前

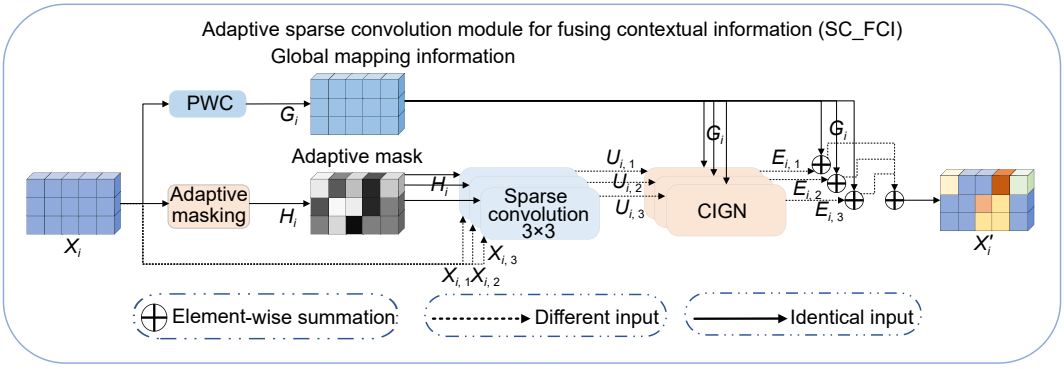


图 6 SC_FCI 结构

Fig. 6 SC_FCI structure

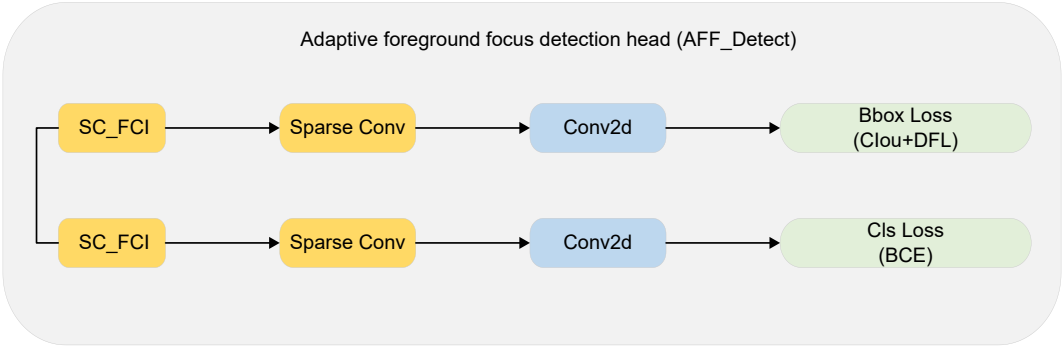


图 7 自适应前景聚焦检测头 (AFF_Detect) 结构

Fig. 7 Adaptive foreground focus detect head (AFF_Detect) structure

景布局，本文通过 PWC 对输入特征$\boldsymbol{X}_i$进行点卷积得到对应全局映射信息G_i，并构建 CIGN (context information group normalization) 归一化模块利用全局信息G_i与焦点特征$\boldsymbol{U}_{i,j}$进一步增强特征的表达能力。经 CIGN 增强后的特征$\boldsymbol{E}_{i,j}$如下：

$$\boldsymbol{E}_{i,j}=\omega\times\frac{\boldsymbol{U}_{i,j}-mean[G_i]}{std[G_i]}+b, j=1,2,3, \tag{17}$$

式中：$\boldsymbol{U}_{i,j}$由输入特征$\boldsymbol{X}_i$应用 SC 所得，$mean[\cdot]$表示均值，$std[\cdot]$表示标准差，ω和 b 为可学习参数。

SC_FCI 通过引入全局上下文损失与自适应掩码比，显著提高了前景目标的聚焦能力，降低了背景噪声的影响与计算量，使得多尺度密集分布目标的漏检、误检情况有所改善。

3 实验结果与分析

3.1 数据集及实验环境

本文选用 VisDrone2019[36] 公开数据集进行实验验证，该数据集被国际无人机视觉领域广泛使用，由天津大学 AISKYEYE 团队收集制作。该数据集在中国十个不同城市使用各种无人机多角度、多场景、多背景拍摄完成，涵盖不同方面的场景。数据集包含 10209 张静态图片，其中 6471 张用于模型训练，548 张用于验证，3190 张用于测试，共预测 10 种类别：行人 (pedestrian)、人 (people)、自行车 (bicycle)、汽车 (car)、面包车 (van)、卡车 (truck)、三轮车 (tricycle)、遮阳篷-三轮车 (awning-tricycle)、公交车 (bus) 和摩托车 (motor)。VisDrone2021[37] 数据集继承并扩展了 VisDrone2019 的数据和任务，在单张图片上提供了更多的场景和目标物体的精细标注。

实验环境为 Ubuntu 18.04 操作系统，CPU 为 Intel Xeon Platinum 8255C，基准频率为 2.50 GHz，GPU 为 NVIDIA GeForce RTX 3090 64 GB，编程语言为 Python 3.8，深度学习框架为 pytorch 1.8.1+CUDA 11.1。为了公平对比算法的性能，所有方法采用相同的超参数设置进行训练、验证，批次大小 (batch_size) 设为 16，初始学习率为 0.01，训练轮数 (epochs) 为 200。

3.2 评价指标

本实验采用查准率 Precision、召回率 Recall、每秒检测帧率 FPS、平均精度均值 AP、加权平均精度均值 mAP@0.5 (IoU 阈值取 0.5 时的 mAP 平均值)、mAP@0.5∶0.95 (IoU 阈值分别取 0.5、0.55、0.6、0.65、0.7、0.75、0.8、0.85、0.9、0.95 时的 mAP 平均值)、模型计算量 GFLOPs 对模型及各类别目标进行单独评价与综合测评。式 (18)~式 (22) 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。对应计算公式为

$$Precision=\frac{TP}{TP+FP}, \tag{18}$$

$$Recall=\frac{TP}{TP+FN}, \tag{19}$$

$$AP=\int_0^1 Precision(Recall)\mathrm{d}(Recall), \tag{20}$$

$$mAP=\frac{1}{n}\sum_{i=1}^{n}\int_0^1 Precision(Recall)\mathrm{d}(Recall), \tag{21}$$

$$FPS=\frac{Framenum}{ElapsedTime}. \tag{22}$$

3.3 消融实验

为验证改进模型对 YOLOv8s 的每个改进策略的有效性，使用 VisDrone2019 数据集以 YOLOv8s 为基线进行消融实验，实验所有数据参数和环境配置严格一致。实验结果如表 1 所示。其中“√”表示使用该改进策略，“×”表示未使用。

表 1 实验结果表明，单独引入 PFRC 全景特征分类层能够有效减少冗余信息，提高特征表示质量，Precision 提高了 3%，Recall 提高了 4.2%，mAP@0.5 提高了 2.4%。将 ATFS 双维特征采样模块替换标准卷积后，Precision 提高了 3.7%，Recall 提高了 5.3%，mAP@0.5 提高了 3.3%，GFLOPs 降低了 1.7%，添加 ATFS 能够增强特征提取能力，降低漏检率。MPFT 多路径全文信息整合模块通过引入上下文信息，帮助模型更准确地定位和识别目标物体，加快推理速度，mAP@0.5 提高了 4.2%，GFLOPs 降低了 3.3%。AFFDH 自适应前景聚焦检测头动态选择重要特征，减少背景噪声，mAP@0.5 提高了 1.3%，GFLOPs 降低了 2.0%。

将 PFRC 和 ATFS 模块同时引入基线模型，mAP@0.5 提高了 4.7%，这说明 ATFS 能够全面高效地采集 PFRC 中的关键前景特征信息。将 PFRC、ATFS 和 MPFT 模块添加到基线模型后，mAP@0.5 相比仅添加 PFRC 和 ATFS 模块提高了 1.1%，GFLOPs 降低了 2.8%，这表明 MPFT 对整体模型起到了轻量化作用，提高了训练速度。将 PFRC、ATFS、MPFT 和 AFFDH 同时添加到基线模型中，即本文的

全部改进策略，相比原模型，Precision 提高了 8.4%，Recall 提高了 7.1%，mAP@0.5 提高了 6.6%。根据消融实验结果可知，改进模型能够提升前景多尺度目标及小目标的检测精度，改善因信息丢失和缺少上下文信息导致的漏检和误检等问题，证明了改进模型的有效性。

为进一步验证改进模型的性能，本文在 VisDrone2019 数据集上对比消融实验的各类别平均精准度 mAP@0.5，结果如表 2 所示。表 2 显示，各类别物体的精度均有所提升，说明改进方法对大、中、小目标的检测效果显著，有效提升了模型的性能。

3.4 对比试验

本文将改进算法与目标检测领域经典算法和近年流行算法在 VisDrone2019 数据集上进行对比，包括 RetinaNet[38]、YOLOv5s、TPH-YOLOv5[39]、YOLOv7-tiny[40]、YOLOv8s、Deformmable-DETR[41]、YOLOv10s[42]、Improved YOLOv5[43]、Faster-RCNN[5]。对比实验结果如表 3 所示。

对比结果显示，改进模型相比 YOLOv5s，查准率 (Precision) 高出 9.4%，召回率 (Recall) 高出 9.0%，mAP@0.5 高出 10.7%，mAP@0.5∶0.95 高出 9.8%。与 TPH-YOLOv5 相比，改进模型在查准率上高出

表 1 所提算法在 VisDrone2019 数据集的消融实验

Table 1 Ablation experiments of the proposed algorithm in the VisDrone2019 dataset

Number	YOLOv8s	PFRC	ATFS	MPFT	AFFDH	Precision/%	Recall/%	mAP@0.5/%	GFLOPs
1	√	×	×	×	×	49.7	37.5	38.5	28.8
2	√	√	×	×	×	52.7	41.7	40.9	34.3
3	√	×	√	×	×	53.4	42.8	41.8	27.1
4	√	×	×	√	×	54.5	41.5	42.7	**25.5**
5	√	×	×	×	√	52.4	41.4	39.8	26.8
6	√	√	√	×	×	54.0	43.4	43.2	32.0
7	√	√	√	√	×	55.8	44.2	44.3	29.2
8	√	√	√	√	√	**58.1**	**44.6**	**45.1**	28.9

表 2 消融实验各类别精度对比结果/%

Table 2 Comparison results of the accuracy of ablation experiments by category/%

Number	Pedestrian	People	Bicycle	Car	Van	Truck	Tricycle	Awning-tricycle	Bus	Motor	mAP@0.5
1	37.2	27.6	14.7	77.4	42.9	39.0	23.8	21.5	56.1	39.4	38.5
2	38.2	22.3	18.9	81.0	44.1	41.2	18.7	25.5	58.1	35.6	40.9
3	39.0	31.0	18.4	80.9	44.9	42.9	24.6	19.9	60.3	39.7	41.8
4	41.8	33.2	17.6	83.8	45.4	40.6	25.8	26.1	61.5	45.6	42.7
5	33.3	22.1	16.3	73.4	39.4	41.0	21.1	22.9	57.4	40.9	39.8
6	44.8	34.7	18.9	84.7	46.6	**50.7**	26.8	25.4	58.3	49.1	43.2
7	52.8	**42.0**	20.1	83.3	46.7	43.8	28.0	26.7	60.8	52.7	44.3
8	**53.9**	41.3	**24.1**	**87.8**	**50.5**	45.3	**31.6**	**28.9**	**62.6**	**55.2**	**45.1**

表 3 VisDrone2019 数据集对比实验结果

Table 3 Results of comparison experiments on VisDrone2019 dataset

Model	Precision/%	Recall/%	mAP@0.5/%	mAP@0.5:0.95/%	GFLOPs	FPS
YOLOv5s	48.7	35.0	34.4	18.5	15.7	120
RetinaNet	35.5	21.9	20.3	12.5	92.9	26
TPH-YOLOv5	49.4	37.0	36.9	19.1	14.6	125
YOLOv7-tiny	50.1	41.2	37.6	22.6	**12.9**	113
Deformmable-DETR	52.4	**45.0**	44.2	25.7	179.3	69
YOLOv8s	49.7	37.5	38.5	22.1	28.8	129
YOLOv10s	55.4	40.7	41.1	23.8	21.6	133
Improved YOLOv5	57.7	43.0	43.9	24.9	34.3	99
Faster-RCNN	48.0	35.1	35.0	21.8	42.5	23
Ours	**58.1**	44.6	**45.1**	**28.3**	28.9	**145**

8.7%，召回率高出 7.6%，mAP@0.5 高出 8.2%，mAP@0.5∶0.95 高出 9.2%。与经典算法 RetinaNet 和 Faster-RCNN 相比，改进模型每秒检测帧率 (FPS) 提高了 6 倍，GFLOPs 分别降低了 65 和 14.6，其他指标也有显著提升。与轻量化模型 YOLOv7-tiny 相比，改进模型的查准度高出 8.0%，召回率高出 3.4%，mAP@0.5 高出 7.0%，mAP@0.5∶0.95 高出 5.7%，FPS 高出 32，处理速度更快。与 YOLOv10s 相比，改进模型的查准率高出 2.7%，召回率高出 3.9%，mAP@0.5 高出 4.0%，mAP@0.5∶0.95 高出 4.5%，FPS 高出 12。与 Improved YOLOv5 相比，改进模型的召回率高出 1.6%，mAP@0.5 高出 1.2%，mAP@0.5∶0.95 高出 3.4%，GFLOPs 降低了 5.4，FPS 高出 46。Deformable-DETR 的召回率更高，mAP@0.5 与改进模型相近，但改进模型的 GFLOPs 仅为 Deformable-DETR 的 1/6，FPS 是 Deformable-DETR 的 2 倍。

图 8 显示了 YOLOv8s 算法与改进模型在训练过程中的各评价指标变化。

在训练初期，改进模型的 Precision、Recall、mAP@0.5 和 mAP@0.5∶0.95 等评价指标均优于基线模型。当训练轮数达到 120 轮时，模型开始收敛。这表明改进模型相比 YOLOv8s 具有更高的检测性能，能在多尺度目标检测中表现出更大优势。

综上所述，改进模型相比同领域的算法具有更高的检测精度、更快的检测速度和最佳的综合性能。在无人机航拍图像目标检测任务中，更具实时性优势。

本文从 VisDrone2019 数据集中选择了“暗光遮挡”、“复杂背景”和“广角多尺度”三组代表性图像，并对各算法的测试结果进行了可视化，结果如图 9 所示。从图中可以看出，Deformable-DETR 利用自适应采样点和局部注意力机制，提高了对小目标的检测能力，但在检测其他尺度目标时，容易出现漏检问题。与 RetinaNet、YOLOv5s、TPH-YOLOv5、YOLOv7-tiny、BDH-YOLO、YOLOv8s、YOLOv10s、Improved YOLOv5、Faster-RCNN 算法相比，改进模型在三组实验中均表现出最佳综合检测性能。在保证检测精度的同时，有效解决了误检和漏检问题。

为进一步验证改进模型的泛化能力和评估算法的鲁棒性，本文在 VisDrone2021 数据集上对各算法进行了对比实验。实验结果如表 4 所示。结果表明，改进模型在 Precision、Recall、mAP@0.5 和 mAP@0.5∶0.95 指标上均优于其他算法。

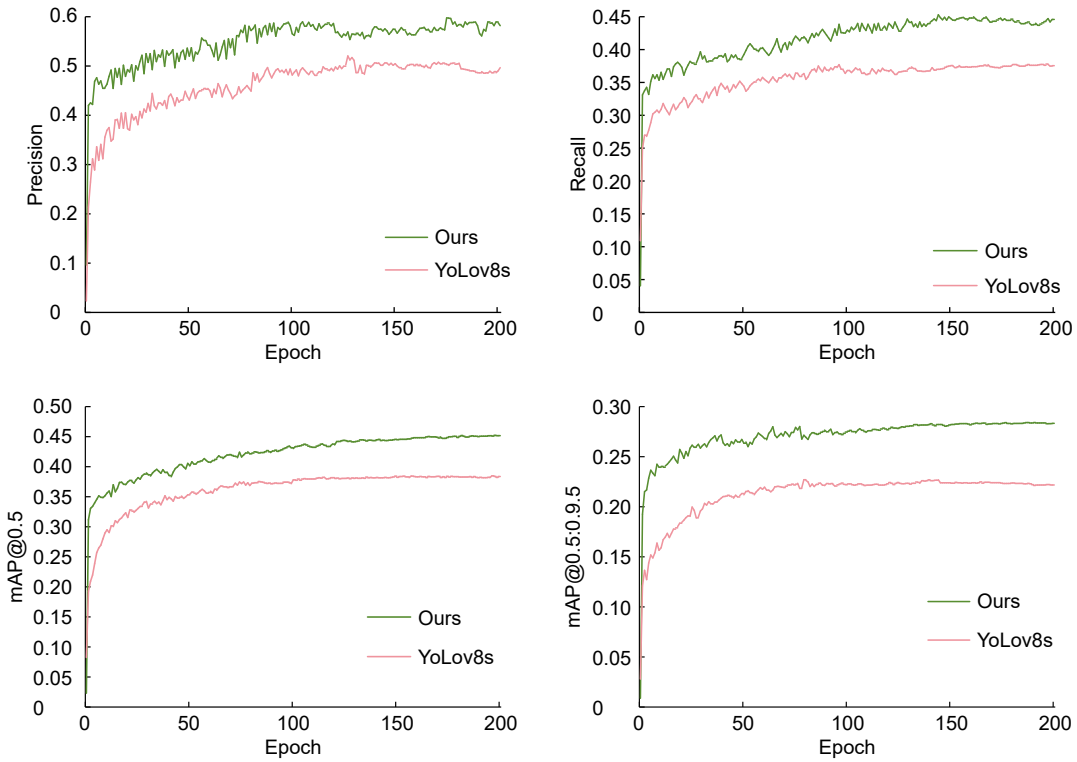


图 8 YOLOv8s 与改进模型评价指标对比

Fig. 8 Comparison of evaluation indicators between the YOLOv8s and improved mode

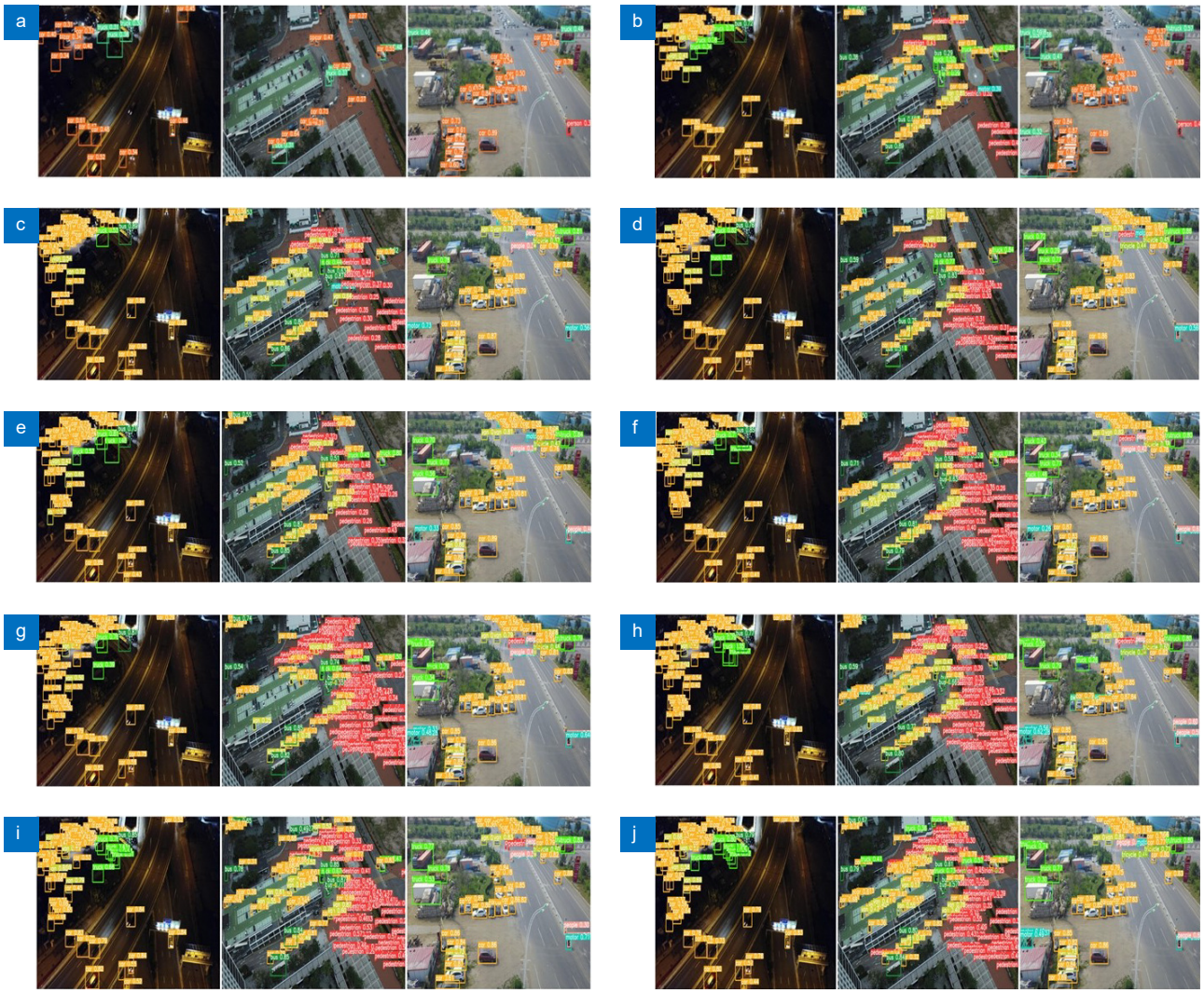


图 9 对比试验可视化结果。(a) RetinaNet; (b) YOLOv5s; (c) Faster-RCNN; (d) TPH-YOLOv5; (e) YOLOv7-tiny; (f) YOLOv8s; (g) YOLOv10s; (h) Improved YOLOv5; (i) Deformmable-DETR; (j) Ours

Fig. 9 Comparison test visualisation results. (a) RetinaNet; (b) YOLOv5s; (c) Faster-RCNN; (d) TPH-YOLOv5; (e) YOLOv7-tiny; (f) YOLOv8s; (g) YOLOv10s; (h) Improved YOLOv5; (i) Deformmable-DETR; (j) Ours

表 4 VisDrone2021 数据集对比实验结果

Table 4 Results of comparison experiments on VisDrone2021 dataset

Model	Precision/%	Recall/%	mAP@0.5/%	mAP@0.5:0.95/%	GFLOPs	FPS
YOLOv5s	46.6	33.1	31.9	16.1	15.7	115
RetinaNet	31.5	18.9	15.3	10.5	92.9	21
TPH-YOLOv5	45.4	34.0	33.6	16.3	14.6	119
YOLOv7-tiny	48.8	39.2	34.6	19.9	**12.9**	105
Deformmable-DETR	51.5	42.1	42.0	22.4	179.3	62
YOLOv8s	47.7	37.5	37.4	19.8	28.8	120
YOLOv10s	50.7	38.1	40.5	21.2	21.6	127
Improved YOLOv5s	52.2	40.7	42.1	22.6	34.3	88
Faster-RCNN	46.1	33.6	32.3	18.8	42.5	20
Ours	**53.1**	**42.6**	**43.1**	**24.3**	28.9	**138**

4 结 论

为解决无人机航拍图像目标检测中的漏检和误检问题，本文提出了一种自适应前景聚焦无人机航拍图像目标检测算法。通过设计全景特征细化分类层，显著增强了网络对前景特征的聚焦能力，提高了算法的检测精度。自适应双维特征采样单元，采用 SPC 卷积与 SK 注意力机制，进一步提高算法对多尺度特征信息提取与细节信息保留能力，改善了误检和漏检问题，降低了模型计算量。多路径信息整合模块通过引入全局上下文信息，实现多尺度特征信息的充分融合，显著提升了算法检测性能，加快了模型推理速度。自适应前景聚焦检测头利用动态聚焦机制，加强了网络对焦点的特征敏感度，使模型在检测精度和资源占用之间达到平衡。在 VisDrone2019 数据集与 VisDrone2021 数据集进行实验。结果表明，该算法较基线模型具有更高的检测精度与检测速度。相比同领域其他算法，改进模型的综合检测性能最佳，满足无人机航拍图像目标检测所需的性能与要求。在未来的研究中，将考虑进一步优化算法的检测性能，在模型检测精度与轻量化之间达到平衡，实现边缘设备上的高精度检测，并进一步提高模型的实际应用价值。

利益冲突: 所有作者声明无利益冲突

参考文献

[1] Chen X, Peng D L, Gu Y. Real-time object detection for UAV images based on improved YOLOv5s[J]. *Opto-Electron Eng*, 2022, **49**(3): 210372.
陈旭, 彭冬亮, 谷雨. 基于改进 YOLOv5s 的无人机图像实时目标检测[J]. 光电工程, 2022, **49**(3): 210372.

[2] Xiong X R, He M T, Li T Y, et al. Adaptive feature fusion and improved attention mechanism-based small object detection for UAV target tracking[J]. *IEEE Internet Things J*, 2024, **11**(12): 21239−21249.

[3] Ma L, Guo Y T, Lei T, et al. Small object detection based on multi-scale feature fusion using remote sensing images[J]. *Opto-Electron Eng*, 2022, **49**(4): 210363.
马梁, 苟于涛, 雷涛, 等. 基于多尺度特征融合的遥感图像小目标检测[J]. 光电工程, 2022, **49**(4): 210363.

[4] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]//*2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2005: 886–893. https://doi.org/10.1109/CVPR.2005.177.

[5] 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.

[6] Redmon J, Divvala S, Girshick R, et al. You only look once: unified, real-time object detection[C]//*Proceedings of 2016 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]//*2017 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] Bochkovskiy A, Wang C Y, Liao H Y M. YOLOv4: optimal speed and accuracy of object detection[Z]. arXiv: 2004.10934, 2020. https://doi.org/10.48550/arXiv.2004.10934.

[10] 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.

[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] Zhang Z, Yi H H, Zheng J. Focusing on small objects detector in aerial images[J]. *Acta Electron Sin*, 2023, **51**(4): 944−955.

[13] Li S C, Yang X D, Lin X X, et al. Real-time vehicle detection from UAV aerial images based on improved YOLOv5[J]. *Sensors*, 2023, **23**(12): 5634.

[14] Li K, Wang Y N, Hu Z M. Improved YOLOv7 for small object detection algorithm based on attention and dynamic convolution[J]. *Appl Sci*, 2023, **13**(16): 9316.

[15] 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.

[16] Zhu M L, Kong E. Multi-scale fusion uncrewed aerial vehicle detection based on RT-DETR[J]. *Electronics*, 2024, **13**(8): 1489.

[17] Shao Y F, Yang Z X, Li Z H, et al. Aero-YOLO: an efficient vehicle and pedestrian detection algorithm based on unmanned aerial imagery[J]. *Electronics*, 2024, **13**(7): 1190.

[18] Zhan W, Sun C F, Wang M C, et al. An improved Yolov5 real-time detection method for small objects captured by UAV[J]. *Soft Comput*, 2022, **26**(1): 361−373.

[19] Chen P L, Wang J T, Zhang Z W, et al. Small object detection in aerial images based on feature aggregation and multiple cooperative features interaction[J]. *J Electron Meas Instrum*, 2023, **37**(10): 183−192.
陈朋磊, 王江涛, 张志伟, 等. 基于特征聚合与多元协同特征交互的航拍图像小目标检测[J]. 电子测量与仪器学报, 2023, **37**(10): 183−192.

[20] Sui J C, Chen D K, Zheng X, et al. A new algorithm for small target detection from the perspective of unmanned aerial vehicles[J]. *IEEE Access*, 2024, **12**: 29690−29697.

[21] Li X, Wang W H, Hu X L, et al. Selective kernel networks[C]//*Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019: 510–519. https://doi.org/10.1109/CVPR.2019.00060.

[22] Zhao X B, Liu K Q, Gao K, et al. Hyperspectral time-series target detection based on spectral perception and spatial-temporal tensor decomposition[J]. *IEEE Trans Geosci Remote Sens*, 2023, **61**: 5520812.

[23] Wu Y X, He K M. Group normalization[C]//*Proceedings of the 15th European Conference on Computer Vision*, 2018: 3–19. https://doi.org/10.1007/978-3-030-01261-8_1.

[24] Yin X Y, Goudriaan J A N, Lantinga E A, et al. A flexible sigmoid function of determinate growth[J]. *Ann Bot*, 2003, **91**(3): 361−371.

[25] Tanaka M. Weighted sigmoid gate unit for an activation function of deep neural network[J]. *Pattern Recognit Lett*, 2020, **135**: 354−359.

[26] 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.

[27] Howard A G, Zhu M L, Chen B, et al. MobileNets: efficient convolutional neural networks for mobile vision applications[Z]. arXiv: 1704.04861, 2017. https://doi.org/10.48550/arXiv.1704.04861.

[28] Zhang P F, Lo E, Lu B T. High performance depthwise and pointwise convolutions on mobile devices[C]//*Proceedings of the 34th AAAI Conference on Artificial Intelligence*, 2020: 6795–6802. https://doi.org/10.1609/aaai.v34i04.6159.

[29] Lin M, Chen Q, Yan S C. Network in network[C]//*2nd International Conference on Learning Representations*, 2013.

[30] Yan S, Shao H D, Wang J, et al. LiConvFormer: a lightweight fault diagnosis framework using separable multiscale convolution and broadcast self-attention[J]. *Expert Syst Appl*, 2024, **237**: 121338.

[31] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//*Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017: 6000–6010.

[32] 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.

[33] Wu H, Wen C L, Shi S S, et al. Virtual sparse convolution for multimodal 3D object detection[C]//*Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023: 21653–21662. https://doi.org/10.1109/CVPR52729.2023.02074.

[34] Feng M K, Yu H C, Dang X Y, et al. Category-aware dynamic label assignment with high-quality oriented proposal[Z]. arXiv: 2407.03205, 2024. https://doi.org/10.48550/arXiv.2407.03205.

[35] Verelst T, Tuytelaars T. Dynamic convolutions: exploiting spatial sparsity for faster inference[C]//*Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 2317–2326. https://doi.org/10.1109/CVPR42600.2020.00239.

[36] 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 2019 IEEE/CVF International Conference on Computer Vision Workshops*, 2019: 213–226. https://doi.org/10.1109/ICCVW.2019.00030.

[37] Cao Y R, He Z J, Wang L J, et al. VisDrone-DET2021: the vision meets drone object detection challenge results[C]//*Proceedings of 2021 IEEE/CVF International Conference on Computer Vision*, 2021: 2847–2854. https://doi.org/10.1109/ICCVW54120.2021.00319.

[38] Wang Y Y, Wang C, Zhang H, et al. Automatic ship detection based on RetinaNet using multi-resolution Gaofen-3 imagery[J]. *Remote Sens*, 2019, **11**(5): 531.

[39] Zhu X K, Lyu S C, Wang X, et al. TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios[C]//*Proceedings of 2021 IEEE/CVF International Conference on Computer Vision*, 2021: 2778–2788. https://doi.org/10.1109/ICCVW54120.2021.00312.

[40] Liu C, Hong Z Y, Yu W H, et al. An efficient helmet wearing detection method based on YOLOv7-tiny[C]//*Proceedings of the 6th International Conference on Machine Learning and Machine Intelligence*, 2023: 92–99. https://doi.org/10.1145/3635638.3635652.

[41] Zhu X Z, Su W J, Lu L W, et al. Deformable DETR: deformable transformers for end-to-end object detection[C]//*9th International Conference on Learning Representations*, 2021.

[42] Wang A, Chen H, Liu L H, et al. YOLOv10: real-time end-to-end object detection[Z]. arXiv: 2405.14458, 2024. https://doi.org/10.48550/arXiv.2405.14458.

[43] Li S X, Liu C, Tang K W, et al. Improved YOLOv5s algorithm for small target detection in UAV aerial photography[J]. *IEEE Access*, 2024, **12**: 9784−9791.

作者简介

肖振久 (1968−)，男，副教授，硕士研究生导师，主要从事机器学习和图像与视觉信息计算方面的研究。

E-mail：xiaozhenjiu@lntu.edu.cn

张杰浩 (2000−)，女，硕士研究生，主要从事遥感图像目标检测方面的研究。

E-mail：zjhao0409@163.com

【通信作者】吴正伟 (1998−)，男，硕士研究生，主要从事机器学习和图像与视觉信息计算方面的研究。

E-mail：1525545769@qq.com

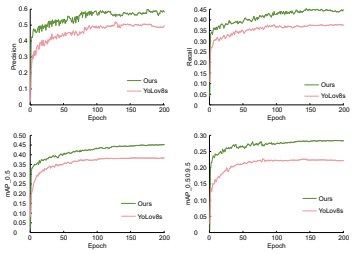
曲海成 (1981−)，副教授，硕士研究生导师，副院长，CCF 会员，主要从事遥感影像高性能计算、视觉信息计算、目标检测与识别。

E-mail：quhaicheng@lntu.edu.cn

扫描二维码，获取PDF全文

Adaptive foreground focusing for target detection in UAV aerial images

Xiao Zhenjiu, Wu Zhengwei*, Zhang Jiehao, Qu Haicheng



Comparison of evaluation indicators between YOLOv8s and improved mode

Overview: To address the issues of missed and false detections due to significant scale variations of foreground targets, uneven sample distribution, and high background redundancy in UAV aerial images, we propose an adaptive foreground-focused object detection algorithm based on the YOLOv8s model. This algorithm incorporates several novel components designed to enhance detection accuracy and efficiency. First, a panoramic feature refinement classification (PFRC) layer is introduced. This layer enhances the algorithm's focus capability and improves the representation quality of foreground samples through re-parameterized spatial pixel variance and shuffle operations. The PFRC layer effectively refines the spatial pixel distribution, highlighting important features while reducing noise. This ensures that the foreground representation is prominent and clear, thereby improving the algorithm's ability to detect objects accurately. Second, we incorporate an adaptive two-dimensional feature sampling (ATFS) unit. This unit employs a separate-learn-merge strategy, which strengthens the extraction of foreground features and retains essential background details. By dynamically adjusting the sampling grid to various scales and orientations, the ATFS unit enhances fine-grained detail extraction. This not only reduces false detections but also accelerates inference, making the algorithm more efficient. Third, a multi-path full-text information integration (MPFT) module is introduced. This module utilizes a multi-branch structure and a broadcast self-attention (BSA) mechanism to address the ambiguity mapping issues caused by downsampling. The MPFT module optimizes feature interaction and integration, enhancing the algorithm's ability to recognize and locate targets accurately. By processing different feature types simultaneously, the multi-branch structure and BSA mechanism reduce the computational load while maintaining high detection accuracy. Finally, we propose an adaptive foreground focus detection head (AFF_Detect). This detection head employs a dynamic focusing mechanism that adjusts based on input characteristics. The AFF_Detect head improves the detection accuracy of foreground targets and suppresses background interference. This dynamic adjustment ensures that the algorithm performs well across various scenarios, enhancing its robustness and generalization capabilities. Experimental results on the VisDrone2019 and VisDrone2021 datasets demonstrate the effectiveness of our proposed algorithm. The mAP@0.5 values achieved are 45.1% and 43.1%, respectively, representing improvements of 6.6% and 5.7% over the baseline model. These results indicate that our algorithm outperforms other state-of-the-art methods, showcasing significant enhancements in detection accuracy, robustness, generalization, and real-time performance. In conclusion, our adaptive foreground-focused object detection algorithm introduces innovative components that address the challenges of UAV aerial image analysis. The integration of the PFRC layer, ATFS unit, MPFT module, and AFF_Detect head results in a comprehensive solution that enhances the representation of foreground features, reduces false detections, and optimizes computational efficiency. These advancements make our algorithm a valuable contribution to UAV-based object detection, offering a significant improvement in performance and reliability.

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; DOI: 10.12086/oee.2024.240149

Foundation item: Project supported by Basic Scientific Research Project of Liaoning Provincial Universities (LJKMZ20220699), and Subject Innovation Team Project of Liaoning Technical University (LNTU20TD-23)

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

* E-mail: 1525545769@qq.com