Computer Science and Application 计算机科学与应用, 2026, 16(8), 220-230
Published Online August 2026 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2026.168276

Hans汉斯

基于改进Yolov12的航拍小目标检测算法

秦喜喜，董骏辉，丁富强，薛小维*

重庆文理学院数学与人工智能学院，重庆

收稿日期：2026年7月16日；录用日期：2026年8月18日；发布日期：2026年8月25日

摘　要

无人机航拍目标检测常受限于微小目标尺度与复杂背景干扰。机载边缘设备显存与算力严苛，难以兼顾精度与实时性；且半精度部署时极易因数值溢出导致网络崩溃。为此，本文提出一种基于YOLOv12的改进DGA-YOLO模型。首先，设计注意力特征融合模块，引入安全极小值与截断机制，解决FP16推理的数值稳定性问题，保障微小目标高频特征的平滑提取。其次，提出动态引导拼接模块，利用极速空间注意力与无重排零初始化残差，以极低内存访问开销实现多尺度特征高效融合。最后，引入无梯度抑制的MPDIoU损失函数，通过角点距离惩罚克服梯度消失，零推理开销下显著提升小目标边界框回归精度。实验表明，DGA-YOLO在保持极低参数量与计算量的同时，难点小目标bicycle、motorcycle精度分别提升2.6%、2.0%。在仅4 GB显存的入门级移动端设备上，模型半精度推理达46 FPS且精度无明显损耗，充分验证了其作为实时检测方案的可行性。

关键词

无人机航拍，小目标检测，YOLOv12，多尺度特征融合，边缘部署

Aerial Small Object Detection via Improved YOLOv12

Xixi Qin, Junhui Dong, Fuqiang Ding, Xiaowei Xue*

School of Mathematics and Artificial Intelligence, Chongqing University of Arts and Sciences, Chongqing

Received: July 16, 2026; accepted: August 18, 2026; published: August 25, 2026

Abstract

Aerial drone-based object detection is often constrained by the small scale of targets and complex

*通讯作者。

文章引用: 秦喜喜, 董骏辉, 丁富强, 薛小维. 基于改进 Yolov12 的航拍小目标检测算法[J]. 计算机科学与应用, 2026, 16(8): 220-230. DOI: 10.12677/csa.2026.168276

background interference. Onboard edge devices have strict limitations in memory and computational power, making it difficult to balance accuracy and real-time performance; furthermore, deploying models in half-precision (FP16) frequently leads to network failure due to numerical overflow. To address these challenges, this paper proposes an improved DGA-YOLO model based on YOLOv12. First, we design an attention-based feature fusion module incorporating a safe minimum value and truncation mechanism to resolve numerical instability issues during FP16 inference, ensuring smooth extraction of high-frequency features for tiny objects. Second, we introduce a dynamic guided concatenation module that leverages fast spatial attention and zero-initialized residual blocks without reordering, enabling efficient multi-scale feature fusion with minimal memory access overhead. Finally, we propose a gradient-free suppression MPDIoU loss function, which mitigates gradient vanishing through corner distance penalties, significantly improving bounding box regression accuracy for small objects at zero inference cost. Experiments show that DGA-YOLO achieves improvements of 2.6% and 2.0% in precision for challenging small targets—bicycles and motorcycles—while maintaining extremely low parameter and computational complexity. On entry-level mobile devices with only 4 GB GPU memory, the model reaches 46 FPS in FP16 inference with no significant loss in accuracy, fully demonstrating its feasibility as a real-time detection solution.

Keywords

Drone Aerial Photography, Small Object Detection, YOLOv12, Multi-Scale Feature Fusion, Edge Deployment

1. 引言

随着无人机技术的普及，航拍目标检测在智能交通与城市安防等领域应用广泛。然而，航拍图像面临目标尺度微小、分布密集且背景复杂等挑战，导致现有模型极易出现严重漏检[1]。此外，传统深层网络在特征提取过程中容易流失微弱的高频信息[2]，且无人机等边缘设备算力受限，难以支撑庞大模型的实时推理。因此，如何在低算力开销下保留多尺度特征并降低漏检率，已成为当前有待解决的工程痛点。

如今，卷积神经网络与 Transformer 等深度学习技术在目标检测领域取得了重大突破。尤其是 YOLO 系列与 RT-DETR 等端到端检测算法，因其具备检测速度快、全局感知强的优势，已被广泛应用于各类无人机航拍与侦察任务中[3] [4]。现有研究已对主流基线模型进行了诸多有效优化：例如，吕等人[5]和朱等人[6]分别通过改进 YOLOv8 架构与引入 Transformer 预测头，显著提升了无人机视角下微小目标的识别率；针对复杂背景干扰，王等人[7]和刘等人[8]探索了高效的多尺度特征聚合机制与边缘实时检测方案，有效增强了模型在密集遮挡环境下的鲁棒性。此外，在网络架构与轻量化方面，王等人[2]引入可编程梯度信息大幅提升了模型参数的利用效率；近期，赵等人[9]提出了具有强大全局建模能力的实时检测器 RT-DETR。然而，针对航拍图像中极端尺度的微小目标，现有轻量化网络在多尺度特征融合阶段仍容易流失高频细节，导致在复杂场景下依然存在较高的漏检风险。

为了有效解决上述挑战，本文提出一种基于 YOLOv12s 的改进 DGA-YOLO 模型。本文创新性地设计了注意力特征融合模块[10]，旨在增强网络对异构分辨率特征的提取与对齐能力，确保微小目标的高频

细节在深层传递中不被高级语义信息淹没。同时，为了进一步优化前向推理效率，本文引入了动态引导拼接模块[11]，在不增加额外计算开销的前提下实现了特征流的高效重构。此外，在模型优化阶段，本文引入了无梯度抑制的 MPDIoU [12]作为边界框回归损失函数，有效突破了传统损失在微小及密集重叠目标上的定位瓶颈，显著提升了模型的回归精度与收敛速度。

2. Yolov12 模型简介

YOLOv12 是一种基于注意力机制的实时目标检测算法，以区域注意力替代传统卷积，在维持高推理速度的同时获得全局感受野。它兼顾了自注意力建模能力与低计算开销，延续 Anchor-free 轻量化范式，是当前效率与精度均衡的主流检测框架之一，其基础结构如图 1 所示。

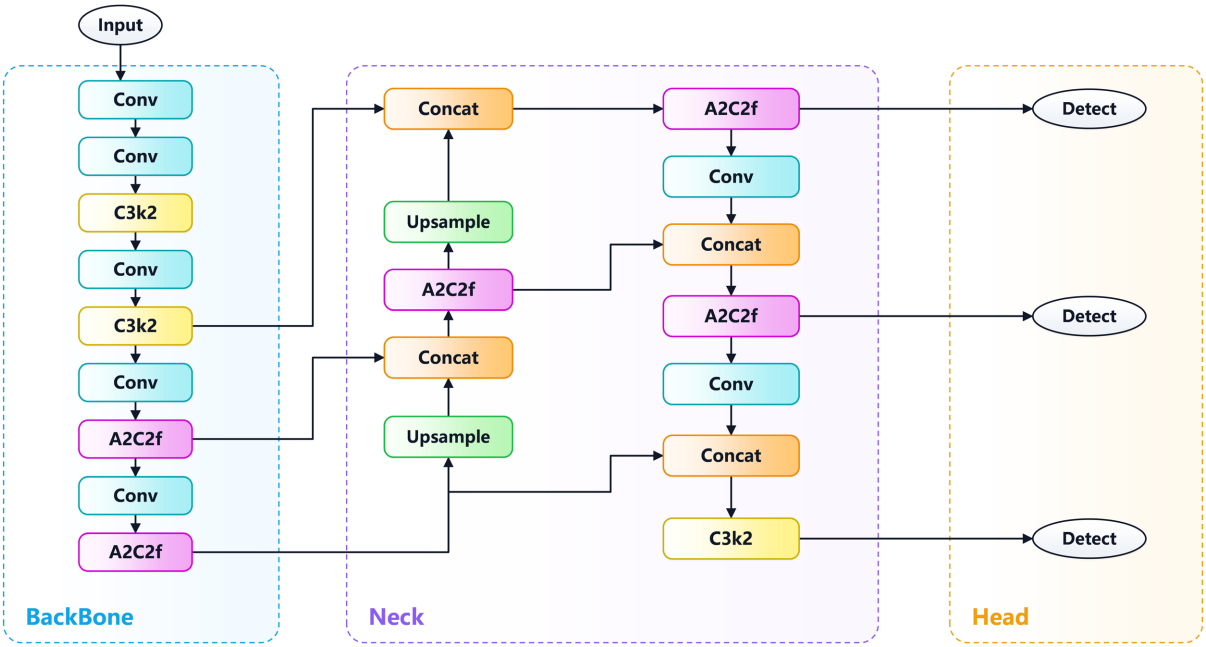


Figure 1. Structure diagram of the YOLOv12 model
图 1. YOLOv12 模型结构图

YOLOv12 由骨干网络、特征融合颈部与解耦检测头三部分组成。骨干网络结合区域注意力与 A2C2f 模块提取多层次特征，缓解了传统卷积在长距离依赖上的不足；颈部通过路径聚合网络对深层语义与浅层高分辨率特征进行双向跨尺度融合，增强多尺度感知能力；解耦检测头则将分类与回归分离，输出检测框与置信度分数。

3. DGA-YOLO 模型

本文以 YOLOv12s 为基线提出 DGA-YOLO，针对航拍小目标特征易丢失、复杂背景干扰强及边缘端半精度溢出等痛点，从特征鲁棒性、多尺度融合效率与定位精度三方面进行改进：首先，引入硬件感知注意力特征融合模块，通过安全极小值与截断机制保障半精度推理的数值稳定性，同时增强小目标高频细节信号；其次，设计动态引导拼接模块重构特征融合网络，利用极速空间注意力与无重排零初始化残差，在极低显存开销下实现跨尺度特征的动态筛选与高效融合，增强复杂背景辨识能力；最后，采用无梯度抑制的 MPDIoU 损失函数，去除梯度抑制机制，通过双角点欧氏距离惩罚实现零开销下更精准的微小目标定位。整体网络结构如图 2 所示。

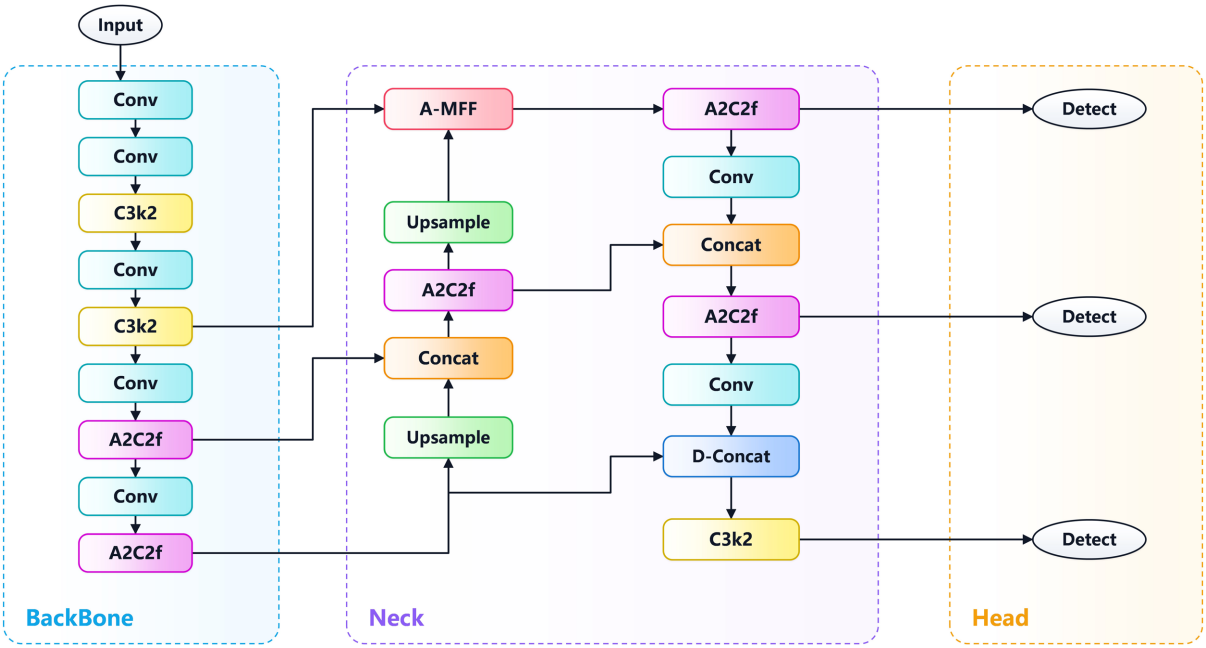


Figure 2. DGA-YOLO model structure diagram
图 2. DGA-YOLO 模型结构图

3.1. 注意力特征融合模块

针对跨尺度特征融合中深层语义易淹没小目标信号，以及边缘设备 FP16 推理时，常规局部能量操作易引发数值溢出导致部署崩溃的双重痛点，本文设计了一种具备数值安全保障的注意力多尺度特征融合模块，其内部结构如图 3 所示，X_{shallow} 和 X_{deep} 分别代表来自网络横向连接的浅层特征张量与自下而上传递的深层特征张量。

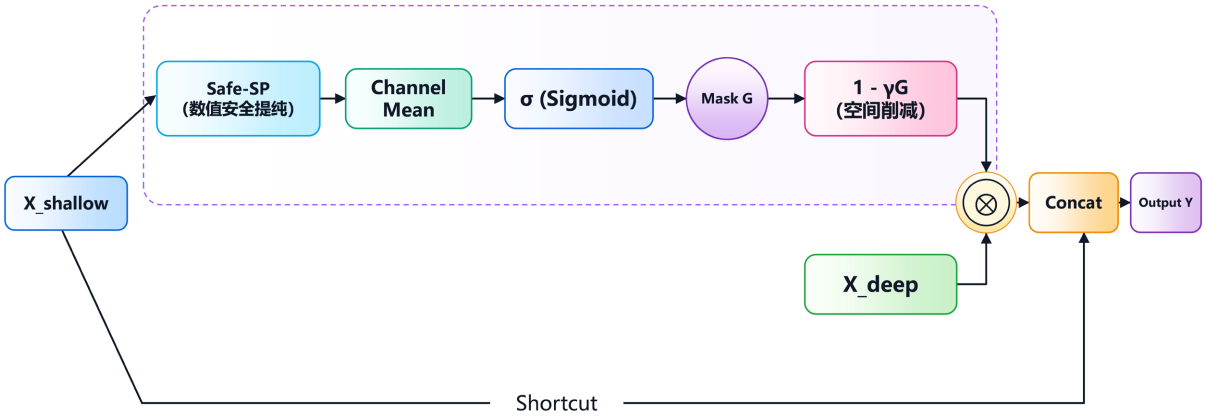


Figure 3. Attention feature fusion module structure diagram
图 3. 注意力特征融合模块结构图

在图 3 中，A-MFF 的核心思想是在安全数值空间内提取浅层高频细节生成空间雷达图，进而对深层特征进行局部语义退让调制。具体而言，为精准提取浅层特征 X_{shallow} 的高频轮廓并彻底规避 FP16 溢出风险，本文在能量计算前设计了 Safe-SP 机制。通过对输入特征幅值施加 $[-100,100]$ 的 Clamp 截断，并辅以极小值保护 ϵ，高频能量图 F_{energe} 的计算如式(1)和式(2)所示：

$$F_{sq} = \left(\text{Clamp}\left(X_{\text{shallow}}, -100, 100\right)\right)^2 + \epsilon \tag{1}$$

$$F_{\text{energy}} = \sqrt{F_{sq}} \tag{2}$$

获取安全的能量图后，模块沿通道维度聚合空间响应，并使用 Sigmoid 函数将其映射为高频空间掩码 G：

$$G = \sigma\left(\text{Channel}_{\text{Mean}}\left(F_{\text{energy}}\right)\right) \tag{3}$$

该掩码 G 准确标定了微小目标或复杂纹理所在的空间位置。最终，利用 G 形成反向门控，对深层特征 X_{deep} 进行空间衰减调制，并与原始浅层特征完成高质量拼接。整体计算如式(4)所示：

$$Y = \text{Concat}\left(X_{\text{deep}} \otimes \left(1 - \gamma \cdot G\right), X_{\text{shallow}}\right) \tag{4}$$

3.2. 动态引导拼接模块

在目标检测网络的特征融合阶段，不同尺度与层级的特征交互至关重要。结合 DGA-YOLO 网络结构图可知，位于 Neck 层的 D-Concat 模块接收两个方向的输入箭头。这两个箭头代表着来自不同网络层级的两股特征流。在面临背景复杂的场景时，如果仅对这两股异构特征进行传统通道维度的粗暴拼接，极易导致关键目标的有效表征被冗余背景淹没。为此，本文提出了一种具有零初始化保底策略的动态引导拼接模块。其内部结构如图 4 所示。

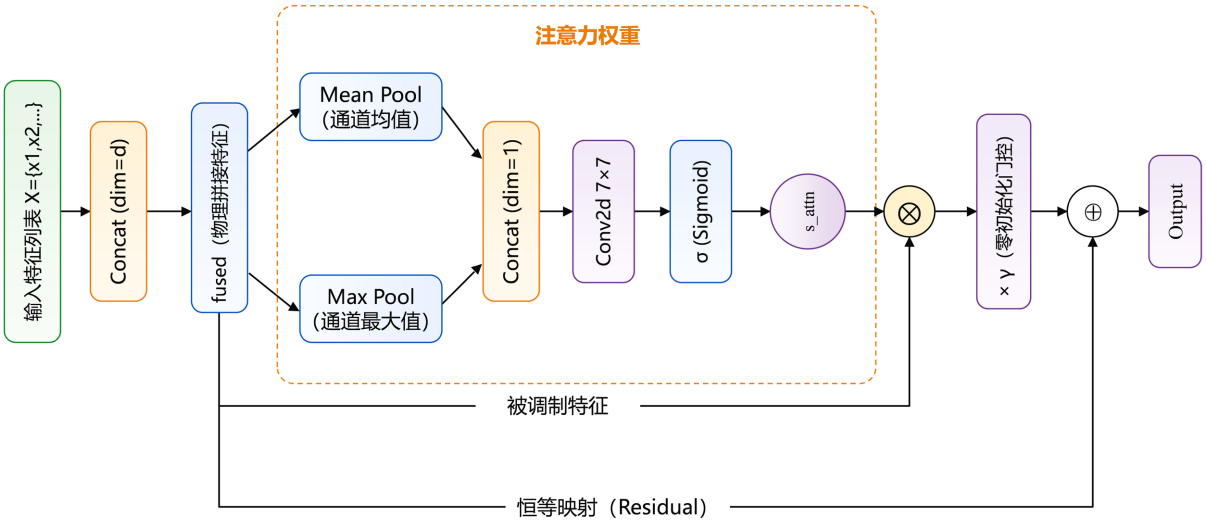


Figure 4. Dynamic guidance concatenation module structure diagram
图 4. 动态引导拼接模块结构图

如图 4 所示，在此模块内对应为输入特征列表 $x = [x_1, x_2]$。为了完整保留这两条路径上最原始的特征表示，网络首先在顶端直接将它们沿通道维度进行标准拼接，生成基础的全局融合特征 F_{fused}：

$$F_{\text{fused}} = \text{Concat}\left(\left[x_1, x_2\right], \text{dim} = d\right) \tag{5}$$

随后，数据流向下进入极速空间注意力提取分支。为了引导网络自动过滤融合后的干扰噪声并聚焦于关键目标区域，F_{fused} 被同时送入通道均值池化和最大池化。这两个操作直接沿通道维度执行，将全局背景分布与局部显著边缘信息压缩并拼接，彻底避免了传统注意力机制中耗时的张量维度重排操作，从

而保证了模块的极低延迟。拼接后的特征紧接着经过一个具有较大感受野的 7×7 卷积层，以捕获更广阔的空间上下文关系，并通过 Sigmoid 函数映射，生成二维的空间注意力权重掩码 M_s：

$$M_s=\sigma\left(\mathrm{Conv}_{7\times7}\left(\left[\mathrm{MeanPool}\left(F_{\mathrm{fused}}\right);\mathrm{MaxPool}\left(F_{\mathrm{fused}}\right)\right]\right)\right) \tag{6}$$

在取到空间注意力权重掩码 M_s 后，流程进入底部的特征调制与残差融合环节。该掩码 M_s 首先与原始融合特征 F_{fused} 进行元素级相乘，以动态抑制冗余背景并增强关键响应。为了确保该创新模块的引入不会破坏主干网络已建立的优秀特征分布，我们在该环节设计了核心的零初始化残差策略。注意力调制后的特征会乘以一个可学习的门控参数 γ，最后再与旁路传入的原始 F_{fused} 进行恒等相加，输出最终特征 F_{out}：

$$F_{\mathrm{out}}=F_{\mathrm{fused}}+\gamma\cdot\left(F_{\mathrm{fused}}\otimes M_s\right) \tag{7}$$

3.3. 无梯度抑制的 MPDIoU 损失函数

在目标检测的边界框回归阶段，主流模型常采用 CIoU 作为损失函数。然而，CIoU 强依赖于预测框与真实框之间的长宽比一致性。当预测框与真实框的长宽比相近但实际尺寸或位置存在偏差时，长宽比惩罚项会迅速衰减至 0，导致模型陷入“梯度饥饿”状态，严重拖慢了网络对微小偏移的收敛速度。此外，部分引入非线性衰减因子的 IoU 变体，会过度压制高质量预测框的微调梯度，限制了模型对复杂几何形态的完美贴合。为了释放模型最原始的学习潜力并打破梯度抑制的瓶颈，本文在损失函数层面摒弃了复杂的长宽比计算，直接引入了一种最小点距离交并比作为边界框回归的评估核心。

具体而言，设真实的边界框为 $B^{gt}=\left(x_1^{gt},y_1^{gt},x_2^{gt},y_2^{gt}\right)$，网络预测的边界框为 $B^{prt}=\left(x_1^{prt},y_1^{prt},x_2^{prt},y_2^{prt}\right)$。首先，计算两者的交并比：

$$\mathrm{IoU}=\frac{\left|B^{gt}\cap B^{prd}\right|}{\left|B^{gt}\cup B^{prd}\right|} \tag{8}$$

随后，放弃了对中心点距离和长宽比的冗余计算，转而直接计算预测框与真实框在左上角点和右下角点的欧氏距离的平方。具体推导如式(9)与式(10)所示：

$$d_1^2=\left(x_1^{gt}-x_1^{prd}\right)^2+\left(y_1^{gt}-y_1^{prd}\right)^2 \tag{9}$$

$$d_2^2=\left(x_2^{gt}-x_2^{prd}\right)^2+\left(y_2^{gt}-y_2^{prd}\right)^2 \tag{10}$$

为了使距离惩罚具备尺度不变性，需要引入包围 B^{gt} 与 B^{prt} 的最小外接凸框，并计算其对角线距离的平方 c^2：

$$c^2=\left(x_{\max}^c-x_{\min}^c\right)^2+\left(y_{\max}^c-y_{\min}^c\right)^2 \tag{11}$$

最终，通过将两个角点的归一化距离惩罚直接作用于 IoU 之上，得到纯净版 MPDIoU 的计算公式：

$$\mathrm{MPDIoU}=\mathrm{IoU}-\frac{d_1^2}{c^2}-\frac{d_2^2}{c^2} \tag{12}$$

4. 实验与结果分析

4.1. 数据集与实验环境

本研究的实验验证阶段采用了 Version 平台提供的开源航拍数据集：aerial (版本：v4，许可协议：CC

by 4.0)。该数据集共包含 2985 张图片，按约 7:2:1 的比例划分为训练集、验证集和测试集，图片数量分别为 2090 张、597 张和 298 张。图像场景涵盖了不同的飞行高度、光照条件及天气状况等。数据集中共定义了 6 个目标类别，具体包括：bicycle、bus、car、motorcycle、person 和 truck。

实验基于 Linux 深度学习服务器环境开展。硬件核心采用 NVIDIA GeForce RTX 4090 (24 GB)显卡，软件配置为 Python 3.8、PyTorch2.0.0。具体的实验环境参数设置如下表 1 所示。

Table 1. Parameter settings
表 1. 参数设置

Parameter	Value
epochs	200
batch	16
imgsz	640
workers	16
optimizer	AdamW
Lr0	0.001

4.2. 消融实验

实验采用常见的评估指标包括召回率 R、以及平均精度均值 mAP50 和 mAP50-95。为评估每项改进的有效性，实验以 YOLOv12s 为基准线，依次添加 A-MFF 模块、D-Concat 模块以及改进损失函数进行了一系列消融实验。

Table 2. Ablation experiments
表 2. 消融实验

Model	Params/M	GFLOPs	R/%	mAP50/%	mAP50-95/%	FPS/桢
基线模型	9.23	21.2	0.452	0.457	0.265	109.9
+A	9.23	21.2	0.454	0.461	0.267	151.5
+A+C	9.23	21.2	0.474	0.467	0.271	144.9
+B+C	9.23	21.2	0.459	0.464	0.270	142.9
+A+B	9.23	21.2	0.461	0.467	0.272	138.9
+A+B+C	9.23	21.2	0.459	0.471	0.271	140.8

表 2 表明，各改进策略均在特定维度上带来了明显的性能增益，其中 A 代表 A-MFF 模块，B 代表 D-Concat 模块，C 代表改进损失函数。在基线模型上单独引入 A-MFF 模块后，模型的 mAP50 由 0.457 小幅提升至 0.461。尤其值得关注的是，在参数量和计算量均保持不变的前提下，推理帧率从 109.9 FPS 显著跃升至 151.5 FPS。这一结果充分表明，A-MFF 模块在增强特征提取能力的同时，有效优化了前向推理的计算效率。进一步组合引入 D-Concat 模块与改进损失函数后，虽然推理帧率略微回落至 142.9 FPS，但检测精度稳步提升，召回率 R 达到 0.459，mAP50 达到 0.464。最终，将 A-MFF、D-Concat 与改进损失函数三项策略进行综合集成，所构建的模型能够充分融合改进的优势。相较于基线模型，在保持参数量和计算量“零增长”的前提下，最终模型的 mAP50 与 mAP50-95 分别提升至 0.471 与 0.271，且推理速

度仍显著优于基线，实现了检测精度与推理效率的良好平衡。

Table 3. The mAP50 of each target category
表 3. 各目标类别的 mAP50

Model	bicycle/%	bus/%	car/%	motorcycle/%	person/%	truck/%
YOLOv12s	0.119	0.599	0.721	0.406	0.412	0.486
DGA-YOLO	0.145	0.620	0.729	0.426	0.418	0.490

此外，表 3 可以看出改进后的综合模型在绝大多数类别上均取得了最优或次优的检测精度。尤其针对航拍场景中检测难度较大的小尺寸目标及易混淆目标，改进策略展现出了更为精准的定位与识别能力：bicycle 的检测精度从 0.119 大幅提升至 0.145，motorcycle 从 0.406 提高至 0.426，bus 与 car 的精度亦均有稳定增长。这表明，本文提出的综合改进方案能够有效抑制航拍图像中的复杂背景干扰，显著改善了对困难目标的漏检问题。

4.3. 对比实验

从表 4 可以看出，与 YOLOv8s、YOLOv11s 以及基线模型 YOLOv12s 的对比中，本文提出的 DGA-YOLO 模型展现出显著的性能优势。在同样保持 9.23 参数量和 21.2 计算量的前提下，DGA-YOLO 的 mAP50 提升至 0.471，涨幅达 1.4%，mAP50-95 达到 0.271，召回率 R 更以 0.459 的成绩取得全场最优。更为重要的是，得益于网络结构的针对性优化，DGA-YOLO 模型有效突破了基线模型的计算瓶颈，将推理帧率大幅回升至 140.8。综上所述，DGA-YOLO 模型未引入任何额外的存储与计算开销，在检测精度与推理效率上均实现了显著提升，在“轻量化、高精度、高实时性”三者之间达成了更优的平衡。

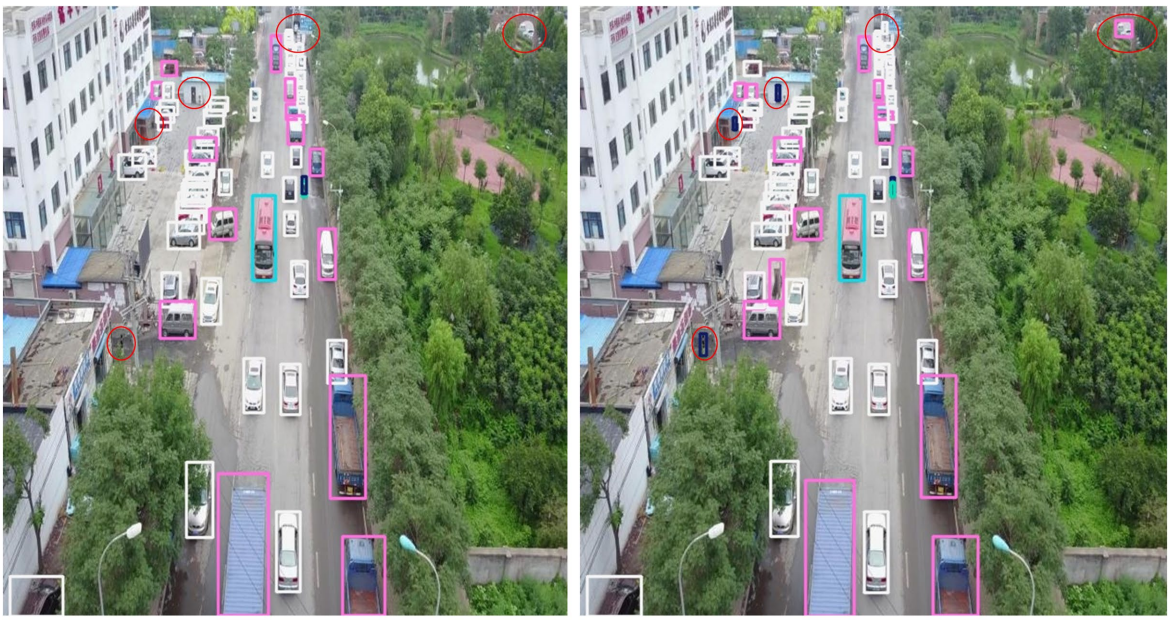
Table 4. Controlled experiment
表 4. 对比实验

Model	Params/M	GFLOPs	R/%	mAP50/%	mAP50-95/%	FPS/桢
YOLOv8s	11.13	28.4	0.445	0.449	0.260	196.1
YOLOv11s	9.42	21.3	0.449	0.453	0.258	149.3
YOLOv12s	9.23	21.2	0.452	0.457	0.265	109.9
DGA-YOLO	9.23	21.2	0.459	0.471	0.271	140.8

4.4. 实验结果可视化

图 5 中，目标极易与建筑阴影或复杂地面背景发生融合。观察左侧基准模型的检测结果可知，处于画面左侧建筑物边缘及深色阴影下的行人由于特征微弱，发生了彻底的漏检；同时，画面右上角极远距离处的微小目标也未能被识别。相比之下，DGA-YOLO 模型凭借强大的多尺度特征提取能力，成功穿透了阴影干扰，不仅精准召回了建筑物旁的隐蔽目标，也敏锐地捕捉到了远处的极限小目标。

图 6 中，目标尺度极小且分布密集，基准模型在此处出现了大面积的“群发性漏检”；此外，分布在画面左侧边缘及正下方被树木冠盖部分遮挡的零星目标，基准模型也毫无响应。而 DGA-YOLO 模型得益于内部高效的特征融合与对齐机制，彻底改善了这一缺陷。它不仅在十字路口中心精准揪出了密集的微小目标群，还成功检出了被树冠遮挡边缘的极小目标。



YOLOv12s　　DGA-YOLO

Figure 5. Street scene viewed from a low-altitude perspective
图 5. 低空视角的街道场景

YOLOv12s　　DGA-YOLO

Figure 6. The crossroads viewed from an elevated perspective
图 6. 高空视角的十字街口

为了直观展示本文改进模型在复杂航拍场景下的目标捕捉能力，尤其对微小目标的召回效果，接下来引入归一化混淆矩阵。上面两张图分别是 YOLOv12s 与 DGA-YOLO 的归一化混淆矩阵。在按真实类别归一化的矩阵中，对角线数值直接表征各类别的召回率，底行的数值则反映真实目标被漏检的比例。图 7 可知，改进模型在几乎所有类别上的召回率均显著提升。尤其针对航拍视角下特征极不明显的小目

标，bicycle 的召回率从 0.11 提升至 0.21，近乎翻倍；person 和 motorcycle 也分别从 0.43 和 0.40 提升至 0.56 和 0.45。与此同时，改进策略有效缓解了基线模型的严重漏检问题——原始 YOLOv12s 对 bicycle 和 person 的漏检率曾高达 75%和 55%，改进后则分别降至 0.55 和 0.38，motorcycle 的漏检率亦降至 0.30，bus、truck 等较大目标的漏检率均降至 0.15。上述可视化结果充分表明，本文构建的改进模型能够有效穿透航拍图像中的复杂背景干扰，显著提升了对困难样本的召回与定位能力。

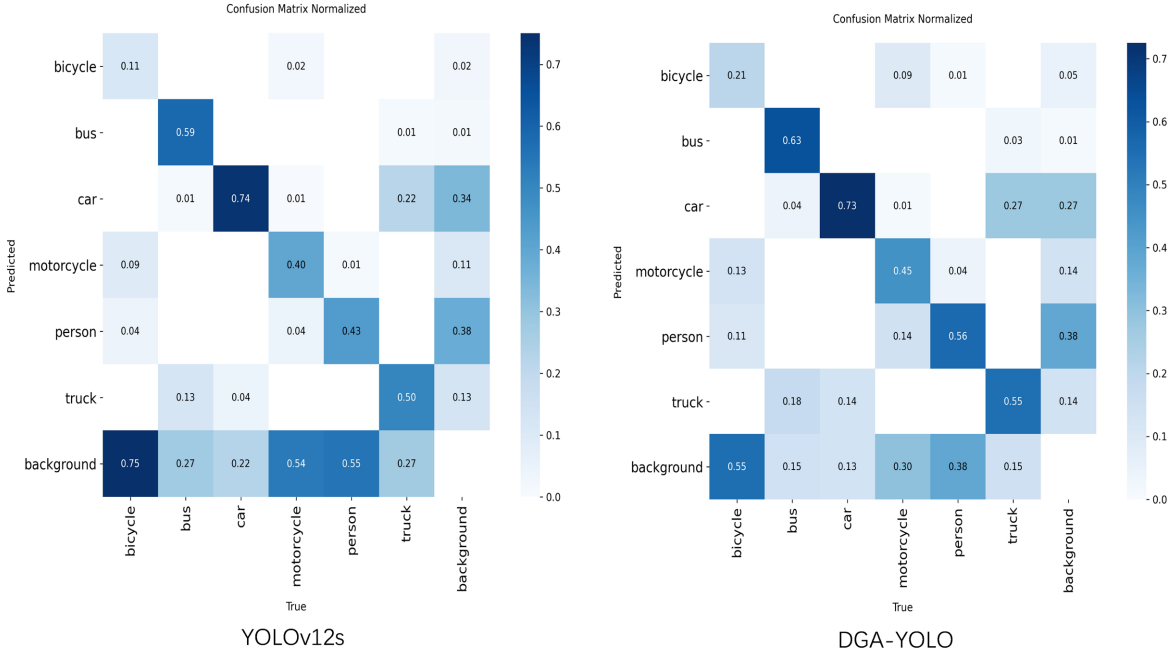


Figure 7. Normalized confusion matrix
图 7. 归一化混淆矩阵

5. 结论

针对无人机航拍场景中目标尺度微小、背景干扰严重及边缘设备算力受限等痛点，本文提出一种基于 YOLOv12 的改进 DGA-YOLO 模型，该模型创新性地引入了注意力特征融合模块与动态引导拼接模块，有效增强了网络对高频微弱特征的响应与多层级语义的对齐能力。

综合实验表明，DGA-YOLO 成功突破了基线模型的计算瓶颈，在实现“零额外参数与计算开销”的前提下，不仅显著提升了模型的前向推理效率，更大幅降低了微小目标的漏检率。可视化分析进一步证实，该模型在阴影遮挡、目标密集等复杂侦察场景下展现出卓越的特征感知能力与鲁棒性。综上所述，DGA-YOLO 在“轻量化、高精度、高实时性”之间取得了理想的平衡，为无人机空中巡查、城市交通监控等资源受限的边缘视觉任务提供了一种极具潜力的工程解决方案。

基金项目

本研究得到了重庆市自然科学基金创新发展联合基金(项目编号：CSTB2025NSCQ-LZX0065)、重庆市教育委员会科学技术研究项目(项目编号：KJQN202201343、KJZD-K202401304)的资助。

参考文献

[1] Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., *et al*. (2024) YOLOv10: Real-Time End-to-End Object Detection.

Advances in Neural Information Processing Systems 37, Vancouver, 10-15 December 2024, 107984-108011. https://doi.org/10.52202/079017-3429

[2] Wang, C., Yeh, I. and Mark Liao, H. (2024) YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information. In: Leonardis, A., *et al*., Eds., *Computer Vision—ECCV* 2024, Springer Nature Switzerland, 1-21. https://doi.org/10.1007/978-3-031-72751-1_1

[3] Wu, X., Li, W., Hong, D., Tao, R. and Du, Q. (2022) Deep Learning for Unmanned Aerial Vehicle-Based Object Detection and Tracking: A Survey. *IEEE Geoscience and Remote Sensing Magazine*, **10**, 91-124. https://doi.org/10.1109/mgrs.2021.3115137

[4] Akyon, F.C., Onur Altinuc, S. and Temizel, A. (2022) Slicing Aided Hyper Inference and Fine-Tuning for Small Object Detection. 2022 *IEEE International Conference on Image Processing* (*ICIP*), Bordeaux, 16-19 October 2022, 966-970. https://doi.org/10.1109/icip46576.2022.9897990

[5] Lyu, Y., Zhang, T., Li, X., Liu, A. and Shi, G. (2024) LightUAV-YOLO: A Lightweight Object Detection Model for Unmanned Aerial Vehicle Image. *The Journal of Supercomputing*, **81**, Article No. 105. https://doi.org/10.1007/s11227-024-06611-x

[6] Zhu, X., Lyu, S., Wang, X. and Zhao, Q. (2021) TPH-YOLOv5: Improved YOLOv5 Based on Transformer Prediction Head for Object Detection on Drone-Captured Scenarios. 2021 *IEEE*/*CVF International Conference on Computer Vision Workshops* (*ICCVW*), Montreal, 11-17 October 2021, 2778-2788. https://doi.org/10.1109/iccvw54120.2021.00312

[7] Wang, C., He, W., Nie, Y., Guo, J., Liu, C., Wang, Y., *et al*. (2023) Gold-YOLO: Efficient Object Detector via Gather-And-Distribute Mechanism. *Advances in Neural Information Processing Systems* 36, New Orleans, 10-16 December 2023, 51094-51112. https://doi.org/10.52202/075280-2224

[8] Liu, S., Zha, J., Sun, J., *et al*. (2023) EdgeYOLO: An Edge-Real-Time Object Detector. arXiv:2302.07483. https://doi.org/10.48550/arXiv.2302.07483

[9] Zhao, Y., Lv, W., Xu, S., *et al*. (2024) DETRs Beat YOLOs on Real-Time Object Detection. arXiv:2304.08069. https://doi.org/10.48550/arXiv.2304.08069

[10] Liu, S., Huang, D. and Wang, Y. (2019) Learning Spatial Fusion for Single-Shot Object Detection. arXiv:1911.09516. https://doi.org/10.48550/arXiv.1911.09516

[11] Chen, Y., Dai, X., Liu, M., Chen, D., Yuan, L. and Liu, Z. (2020) Dynamic Convolution: Attention over Convolution Kernels. 2020 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 11027-11036. https://doi.org/10.1109/cvpr42600.2020.01104

[12] Ma, S. and Xu, Y. (2023) MPDIoU: A Loss for Efficient and Accurate Bounding Box Regression. arXiv:2307.07662. https://doi.org/10.48550/arXiv.2307.07662