无人机视角下的小目标检测算法研究

摘要

无人机图像中包含大量小目标，由于目标密集分布、尺寸微小以及背景复杂等因素，当前的检测精度和特征提取效率仍有较大的提升空间。为提高小目标的检测效果，本文提出了一种基于 RT-DETR 的无人机小目标检测改进算法 RT-DETR-GH。为增强特征提取能力，在骨干网络中引入了卷积门控线性单元（CGLU），通过动态门控机制自适应调节特征通道的重要性，进而提高模型对有效特征的捕获能力，并抑制冗余特征对检测结果的干扰。此外，本文还在 AIFI 模块中引入了 HiLo 注意力机制，结合尺度内特征交互模块，进一步增强了模型对密集目标的关注。HiLo 注意力机制通过分离高频和低频注意力（Hi-Fi 和 Lo-Fi），分别处理细粒度特征和大尺度上下文信息，有效提升了模型在不同尺度特征上的捕捉能力。实验结果表明，改进后的模型，性能并没有显著提升，甚至出现了对冗余特征的过度关注，从而导致检测效果不如 RT-DETR。

关键词：小目标检测；注意力机制；RT-DETR；特征提取

1 引言

在计算机视觉领域，小目标检测一直是一个难以攻克的挑战。近年来，无人机技术的快速发展使其航拍图像采集功能得到了广泛应用。无人机能够轻松到达人类难以涉足的区域，例如偏远山区、茂密森林和危险建筑上空，从而获取这些地方的图像数据。然而，由于高空俯拍的视角，场景中包含大量小目标，这显著增加了目标检测的难度。

目前，针对航拍图像的目标检测方法主要依赖深度学习技术。基于深度学习的目标检测算法可以大致分为两阶段的 R-CNN [1] 系列、单阶段的 YOLO [2] 系列以及基于 DETR [3] 的改进系列。其中，单阶段检测技术在实时性方面具有显著优势。近年来，单阶段目标检测算法，尤其是 YOLOv5 之后的改进算法，通过对网络结构、激活函数、损失函数和训练策略等多个方面的优化，在自然图像检测中实现了更高的检测精度和更快的检测速度。然而，这些算法在无人机航拍图像的小目标检测中表现不佳。一方面，基于 CNN 的算法通常依赖非极大值抑制（NMS）进行后处理。当图像中存在大量目标时，NMS 会显著增加计算复杂度，从而潜在地影响实时检测的速度和稳定性。另一方面，航拍图像通常包含高分辨率的大型图像，其中的目标往往具有小物体、重叠的多尺度特性以及复杂的背景，使得卷积核难以有效捕获物体细节和空间关系，从而限制了检测精度的提升。

Transformer 通过多头自注意力机制能够捕获全局依赖关系，使其在建模图像中不同物体的复杂关系时表现出色。DETR 模型结合了传统卷积网络和 Transformer 的优点，采用混合架构充分发挥了 CNN 在图像特征提取方面的优势以及 Transformer 在捕获全局依赖关系方

面的能力。DETR 通过端到端学习直接预测目标的位置和类别，消除了对锚框设计和 NMS 后处理的依赖，使其在目标检测任务中实现了较高的检测精度和稳定的检测速度。然而，DETR 的高计算量限制了其在实际场景中的应用。

为了解决这一问题，百度提出了 RT-DETR [4]，一种基于 DETR 的实时目标检测方法。在检测精度和速度方面，RT-DETR 均优于现有的 YOLO 系列模型。RT-DETR 模型采用 ResNet [5] 作为骨干网络进行特征提取，并在特征提取网络的最后一层引入 Transformer 编码器，以实现全局特征相关性建模。在颈部网络中，设计了类似 PAFPN 结构的 CCFM 模块，通过自顶向下和自底向上的特征图融合来增强特征表达。在查询选择阶段，RT-DETR 使用 Top-k 选择策略，将融合模块的密集预测结果传递给解码器，以完成最终的目标检测任务。尽管 RT-DETR 在实时性和精度上实现了较好的平衡，但在无人机航拍小目标检测场景中仍然存在不足。骨干网络中使用的 ResNet 在下采样操作时丢失了大量关于小目标的信息，这显著限制了网络捕捉小目标细微特征的能力。为了解决这一问题，本文在 RT-DETR 的骨干网络中引入了卷积门控线性单元（CGLU） [6]。CGLU 通过动态门控机制，自适应调整特征通道的重要性，从而增强了模型对关键特征的捕获能力，同时抑制了冗余特征对检测结果的干扰。这一设计有效提升了模型在小目标场景中的特征提取效率。此外，CGLU 还能够动态平衡特征通道的权重，在保证计算效率的同时增强对小目标的表征能力。

此外，为了解决航拍图像中多尺度目标和复杂背景带来的问题，本文在 RT-DETR 的 AIFI（Anchor-Free Instance Interaction）模块中引入了 HiLo 注意力机制 [7]。HiLo 注意力机制通过分别处理高频和低频特征，将细粒度的局部特征（Hi-Fi）与大尺度的全局上下文信息（Lo-Fi）相结合。具体而言，Hi-Fi 部分聚焦于捕获小目标的边缘信息和细节特征，而 Lo-Fi 部分则用于建模目标间的全局关系和上下文语义信息。这种高低频特征分离的方法在不显著增加计算成本的前提下，有效提升了模型对多尺度目标的关注能力。结合尺度内特征交互模块，HiLo 注意力机制进一步增强了模型对密集小目标的检测效果，使其在复杂背景场景中表现出更强的鲁棒性。

综上所述，本文的主要改进和创新点如下：

（1）在骨干网络中引入了卷积门控线性单元，通过动态门控机制自适应地调节特征通道的重要性，提高了模型对小目标关键特征的捕获能力，同时抑制冗余特征的干扰。这有效解决了传统 ResNet 在下采样过程中导致的小目标特征信息丢失问题。

（2）在 AIFI 模块中引入了 HiLo 注意力机制，通过分离高频和低频特征处理，分别捕获局部细节信息和全局上下文关系，提升了模型对多尺度目标的适应能力，同时结合尺度内特征交互模块，进一步优化了模型对复杂场景中密集小目标的检测能力。

2 相关工作

近年来，UAV 目标检测技术的发展取得了显著进步。随着计算机视觉和深度学习技术的快速发展，目标检测算法在无人机应用场景中得到了广泛应用。传统目标检测方法通常基于图像处理和手工设计特征提取，但这些方法在面对复杂背景、多尺度目标以及小目标检测等问题时，表现出明显的局限性。相比之下，基于深度学习的目标检测算法能够更准确、更实时地检测目标，因而在无人机航拍图像的目标检测任务中取得了重要突破。

针对无人机航拍图像目标检测的挑战，近年来涌现了多种改进算法。例如，Zhai 等人 [8]

提出了一种基于优化 YOLOv8 的无人机目标检测方法，以解决无人机图像中小目标检测的挑战。该方法通过以下几个改进增强了模型性能：首先，在检测头中加入高分辨率检测头，同时裁剪了大目标检测头和冗余网络层，从而提高了小目标检测能力，并有效减少了网络参数；其次，在特征提取阶段引入 SPD-Conv 替代 Conv，以减少细粒度信息的丢失，增强多尺度特征的提取能力；最后，在颈部网络中融合 GAM 注意力机制，通过增强目标特征的融合，进一步提升了检测精度。Zhang 等人 [9] 提出了一种新的远程感知目标检测模型 RS-DETR，以应对遥感图像中目标尺度变化大、小目标占主导、背景复杂等挑战。该方法通过以下几方面改进提升了检测性能：首先，在注意力驱动的特征交互模块中引入级联分组注意力（CGA），通过捕获不同层次的特征并进行级联交互，不仅增强了特征的交互性，还提高了计算效率；其次，提出了增强型双向特征金字塔网络（EBiFPN），通过多尺度特征的融合提升了目标检测的准确性和鲁棒性；最后，设计了一种新的边界框回归损失函数 Focaler-GIoU，使模型更加关注难检测样本，从而改善了对小目标和重叠目标的检测性能。

3 本文方法

3.1 本文方法概述

RT-DETR 是一种创新的实时端到端目标检测框架，它依托 Transformer 架构，巧妙地应对了多尺度特征处理的挑战，本文选用 RT-DETR 作为基线模型，其详细的网络设计如图 1所示。

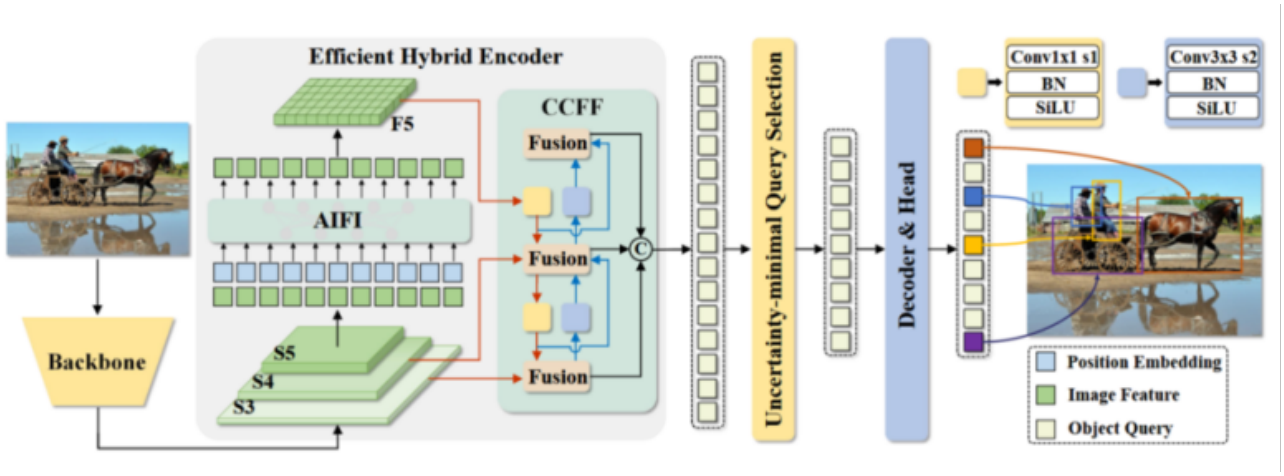

图 1. RT-DETR 网络结构图

RT-DETR 模型主要由骨干网络、AIFI 和 CCFM 组成的混合编码器，以及带有辅助预测头的解码器共同构成。

3.2 特征提取模块

骨干网络通过卷积层和 BasicBlock 模块实现初步的特征提取，将 S3、S4、S5 这三层的输出送入混合编码器来提供丰富的多层次信息。混合编码器内 AIFI 模块专注于处理高级图像特征，通过自注意力机制在 S5 特征图上进行内部尺度交互，提高模型在对象检测和识别方面的性能，CCFM 模块则利用自底向上和自顶向下的双路径融合策略，通过上采样和下采样有

效整合 S3、S4、F5 三个特征图的多尺度特征。解码器部分包括 loU-aware query 和 Decoder head 模块，其中 loU 感知查询模块根据分类分数选择的排名靠前的 K 个预测框，将排名靠前的预测框输出，loU 感知查询选择可以为对象查询提供更多具有准确分类和精确位置的编码器特征，从而提高检测器的准确度。随后通过迭代优化过程，解码器逐步生成精确的边界框预测和对应的置信度评分，以完成高效且准确的检测任务。

3.3 损失函数定义

RT-DETR 的损失函数设计旨在优化目标检测的精度和效率，主要包括分类损失、边界框回归损失和辅助损失。

3.3.1 分类损失（Classification Loss）

使用 Focal Loss 作为分类损失函数，用于解决目标检测中正负样本不平衡的问题。Focal Loss 通过引入调制因子，减少易分类样本的权重，使模型更加关注难分类样本。公式如下：

$$L_{\text{cls}} = -\alpha_t (1 - p_t)^{\gamma} \log(p_t)$$

其中，p_t 是模型预测的目标类别概率，α_t 是类别权重，γ 是调制因子。

3.3.2 边界框回归损失（Bounding Box Regression Loss）

使用 GIoU Loss 作为边界框回归损失函数，优化预测框与真实框之间的重叠程度。公式如下：

$$L_{\text{box}} = 1 - \text{GIoU}(B_{\text{pred}}, B_{\text{gt}})$$

其中，B_{pred} 是预测的边界框，B_{gt} 是真实的边界框。

3.3.3 辅助损失（Auxiliary Loss）

引入辅助损失函数，包括 Query 分类损失和 Query 回归损失，通过多任务学习帮助模型更好地收敛。

3.3.4 总损失函数

总损失函数是分类损失、边界框回归损失和辅助损失的加权和：

$$L_{\text{total}} = \lambda_1 L_{\text{cls}} + \lambda_2 L_{\text{box}} + \lambda_3 L_{\text{aux}}$$

其中，λ_1、λ_2、λ_3 是各损失项的权重系数。

4 复现细节

4.1 与已有开源代码对比

本文提出的 RT-DETR-GH 是在骨干网络 ResNet18 中引入了 CGLU 模块，CGLU 结合了卷积操作与门控机制，允许网络在特征提取阶段动态选择和强化重要的通道，而抑制不相

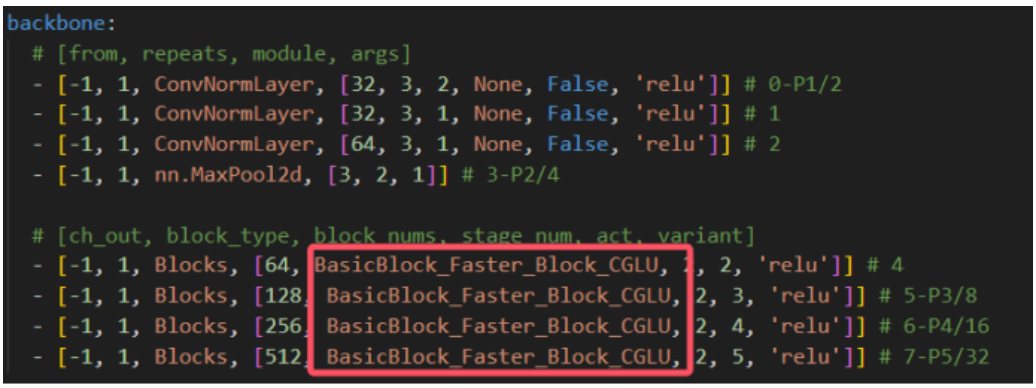

图 2. 骨干网络中引入 CGLU 模块

关或冗余的通道，从而使得特征表示更加精炼。这一设计不仅提升了模型对小目标的识别能力，还避免了传统卷积神经网络中常见的计算冗余问题。如图 2所示。

针对无人机图像中目标的多尺度特性，本文在 RT-DETR 的 AIFI 模块中引入了 HiLo 注意力机制。HiLo 注意力机制将高频特征和低频特征进行分离处理，以分别捕捉小目标的细节信息和全局上下文信息，如图 3所示。

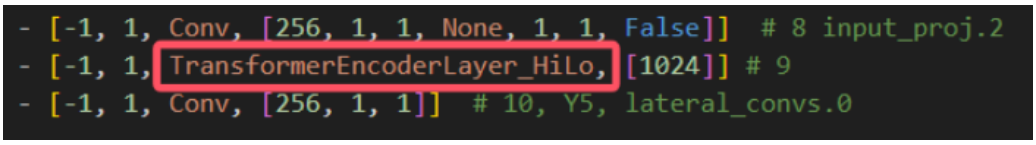

图 3. 在 AIFI 模块中引入 HiLo 注意力机制

4.2 实验环境搭建

本文实验使用的数据集为 VisDrone2019 公共数据集。其中，6471 个训练集，548 个验证集，1610 个测试集。但是由于算力原因，并没有使用全部数据集，最后选用了 1500 张图片，按照 7:2:1 的比例划分数据集。无人机捕获的图像具有鲜明的特征，包括显著的尺寸变化，具有各种干扰的复杂环境，以及灵活可变的多种物体形状。该数据集一共有十个类别。每个类别的实体数量差异很大，行人和汽车占据了总数的大部分。如图 4所示。在训练过程中，输入图像大小设置为 640 ×640，批处理大小设置为 8,epoch 数设置为 100。本文采用基于 RT-DETR 的超参数，选取前 300 个编码特征初始化解码器的对象查询模块，选用 Adamw 优化器。实验环境如 5所示。实验参数设置如图 6所示。

4.3 创新点

RT-DETR 在骨干网络和 AIFI 模块中引入了多项创新设计，显著提升了模型的特征提取能力和多尺度目标检测性能。

4.3.1 骨干网络中的卷积门控线性单元（Conv-GLU）

在骨干网络中引入了卷积门控线性单元（Conv-GLU），通过动态门控机制自适应地调节特征通道的重要性。这种设计能够有效增强模型对小目标关键特征的捕获能力，同时抑制冗余

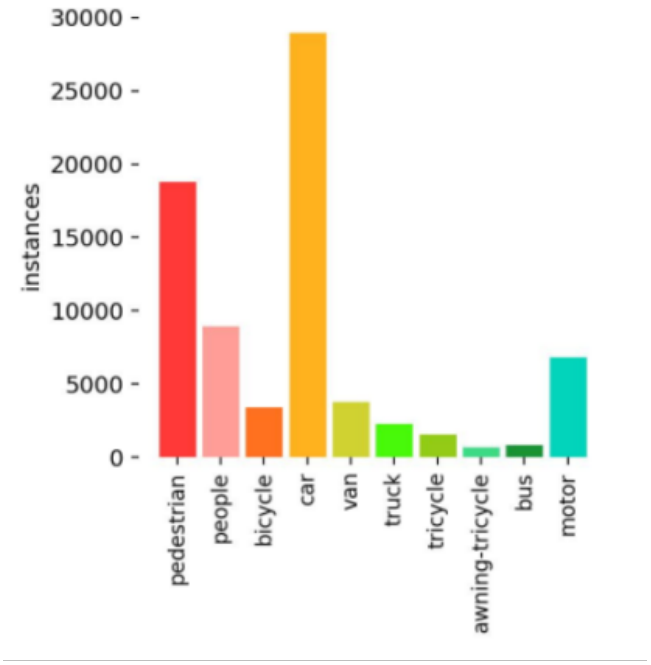

图 4. 数据集信息

参数	配置
CPU	16 vCPU Intel(R) Xeon(R) Platinum 8481C
GPU	RTX 4090D(24GB)
操作系统	Ubuntu22.04
Python	3.8

图 5. 实验环境

特征的干扰。传统 ResNet 在下采样过程中容易导致小目标特征信息的丢失，而 Conv-GLU 通过动态门控机制，能够自适应地选择重要特征通道，从而缓解这一问题。具体来说，Conv-GLU 通过以下方式实现：

- **动态门控机制**：根据输入特征动态生成门控权重，调节特征通道的激活强度。
- **特征选择**：增强对小目标关键特征的表达能力，同时抑制无关特征的干扰。

这一创新设计显著提升了模型对小目标的检测性能，尤其是在复杂场景中。

4.3.2 AIFI 模块中的 HiLo 注意力机制

在 AIFI 模块中引入了 HiLo 注意力机制，通过分离高频和低频特征处理，分别捕获局部细节信息和全局上下文关系。这种设计能够有效提升模型对多尺度目标的适应能力。HiLo 注意力机制的核心思想是将特征图分解为高频和低频两部分：

- **高频特征**：捕捉局部细节信息，增强模型对小目标和边缘特征的感知能力。
- **低频特征**：捕捉全局上下文关系，增强模型对大目标和复杂场景的理解能力。

```python
if __name__ == '__main__':
    model = RTDETR('ultralytics/cfg/models/rt-detr/rtdetr-r18.yaml')
    # model.load('') # loading pretrain weights
    model.train(data='dataset/data.yaml',
                cache=False,
                imgsz=640,
                optimizer='AdamW',
                epochs=100,
                batch=8,
                workers=16,
                project='runs/train',
                name='exp',
                )
```

图 6. 实验参数设置

此外，HiLo 注意力机制与尺度内特征交互模块相结合，进一步优化了模型对复杂场景中密集小目标的检测能力。具体来说：

- **尺度内特征交互**：在特征图内部进行多尺度特征交互，增强模型对多尺度目标的表达能力。
- **密集目标检测**：通过高频特征的增强，模型能够更好地处理密集小目标的检测任务。

这一创新设计显著提升了模型在多尺度目标检测任务中的性能，尤其是在复杂场景中。

5 实验结果分析

通过观察图 7，RT-DETR 在精度，召回率，map 值上都是优于 YOLO 系列模型，说明 RT-DETR 可能对数据集中的各种场景和目标类型有更好的适应性。它在处理不同尺度、不同背景和不同姿态的目标时，能够有效地提取特征并进行准确的检测。改进后的模型 RT-DETR-GH 在精度上由于 RT-DETR，但是在召回率和 map 值上比基线模型差一些。RT-DETR-GH 可能存在过拟合，将一些非检测目标的物体检测进来。

根据下图 8，在模型参数数量方面，YOLOv5 和 YOLOv8 的参数数量较为接近，分别为 25.1M 和 25.8M。RT-detr 的参数数量最少，为 19.8M，而 RT-detr-GH (ours) 的参数稍微多一点，为 23.2M。这说明 RT-DETR 在模型参数量上是具有优势的，由于我替换了骨干网络中的卷积块和添加了注意力机制层，导致改进后的模型参数量具有一定的上升。在 GLOPs 和 ModelSize 上都可以反应出 RT-DETR 系列的模型比 YOLO 系列的轻量一些，可能更便于部署在边缘设备上。但是在 FPS 上，RT-DETR 实时处理的效率并不是特别高。

从下图 9可以看出来，RT-DETR-GH 应该是过拟合了，由于在改进的时候，对特征提取和特征融合进行了重点操作，导致模型过于关注微小的物体，反而导致模型的 map 下降。

6 总结与展望

本文针对无人机视角下的目标检测进行研究，旨在克服无人机航拍图像中密集分布，尺度变化大和背景复杂等问题，本文提出了基于 RT-DETR 的改进算法 RT-DETR-GH，通过在骨干网络中引入卷积门控线性单元 CGLU，解决了 ResNet 在下采样过程中小目标特征信

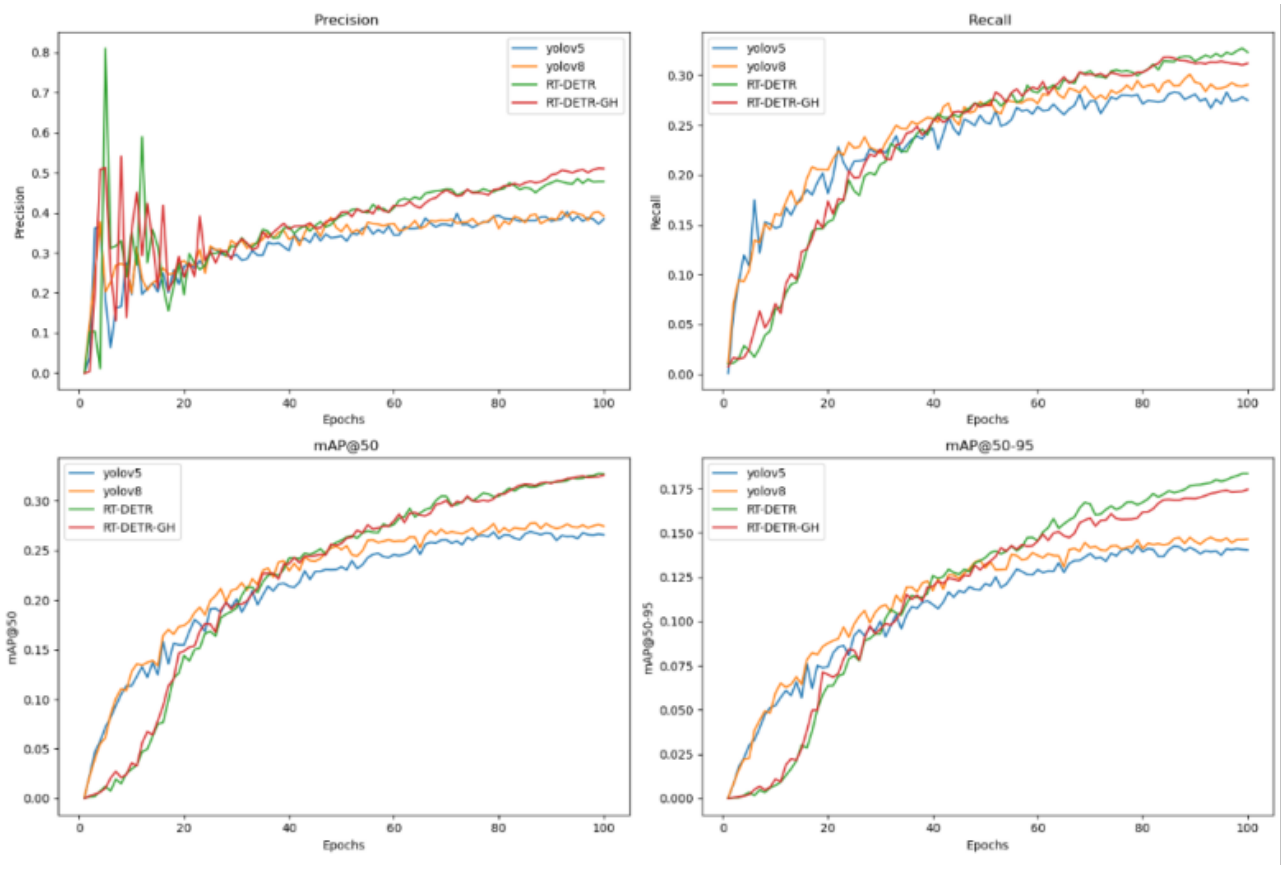

图 7. 模型对比图

模型	Param(M)	GLOPs	FPS	ModelSize(MB)	map@50	map@0.5-0.95
YOLOv5	25.1	64.0	226.5	48.1	26.9	14.3
YOLOv8	25.8	78.7	240.9	49.6	27.7	14.7
RT-detr	19.8	57.0	66.3	38.6	32.7	18.3
RT-detr-GH(ours)	23.2	62.5	41.7	57.2	32.6	18.1

图 8. 模型效果对比图

息丢失的问题，增强了模型对小目标关键特征的捕获能力，并抑制了冗余特征的干扰；同时在 AIFI 模块中引入 HiLo 注意力机制，结合尺度内特征交互模块，有效提升了模型对多尺度目标和密集小目标的检测能力，使其在复杂背景场景中表现出更强的鲁棒性。然而，通过在 VisDrone2019 公共数据集上的实验分析发现，改进后的模型 RT-DETR-GH 虽然在一定程度上提升了对小目标的识别能力，但由于对特征提取和融合的重点操作，导致模型出现过拟合现象，过于关注微小物体，使得召回率和 map 值反而不如基线模型 RT-DETR，且模型参数量有所增加。未来的研究可以继续探索更有效的特征提取和融合方法，以及如何更好地平衡模型的复杂度和泛化能力，以提高无人机小目标检测的精度和效率。

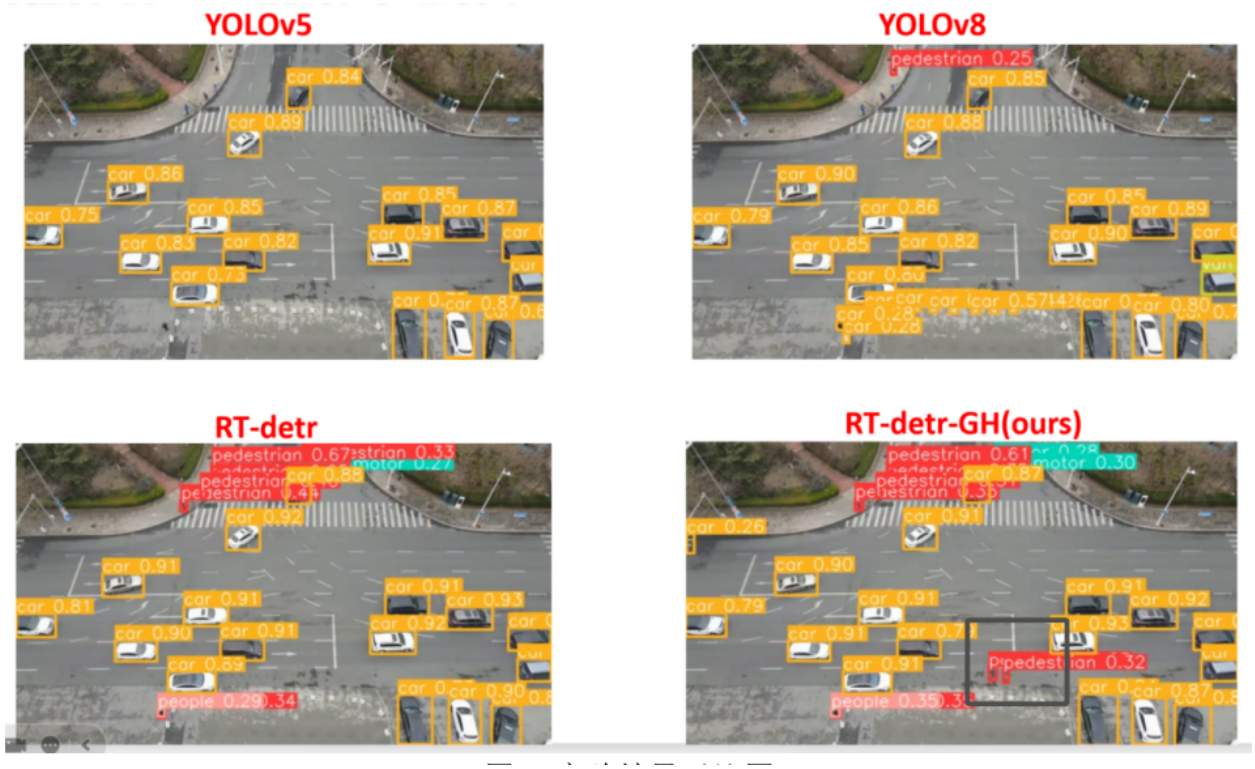

图 9. 实验结果对比图

参考文献

[1] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[2] J Redmon. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[4] Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. Detrs beat yolos on real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16965–16974, 2024.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[6] Dai Shi. Transnext: Robust foveal visual perception for vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17773–17783, 2024.

[7] Zizheng Pan, Jianfei Cai, and Bohan Zhuang. Fast vision transformers with hilo attention. *Advances in Neural Information Processing Systems*, 35:14541–14554, 2022.

[8] Xianxu Zhai, Zhihua Huang, Tao Li, Hanzheng Liu, and Siyuan Wang. Yolo-drone: an optimized yolov8 network for tiny uav object detection. *Electronics*, 12(17):3664, 2023.

[9] Hao Zhang, Zheng Ma, and Xiang Li. Rs-detr: An improved remote sensing object detection model based on rt-detr. *Applied Sciences*, 14(22):10331, 2024.