计算机工程与应用
Computer Engineering and Applications
ISSN 1002-8331,CN 11-2127/TP

《计算机工程与应用》网络首发论文

题目： 改进 RT-DETR 的无人机图像目标检测算法
作者： 姜贸翔，司占军，王晓喆
网络首发日期： 2024-09-13
引用格式： 姜贸翔,司占军,王晓喆.改进 RT-DETR 的无人机图像目标检测算法[J/OL].计算机工程与应用.https://link.cnki.net/urlid/11.2127.tp.20240912.1724.025

网络首发时间：2024-09-13 13:32:03
网络首发地址：https://link.cnki.net/urlid/11.2127.tp.20240912.1724.025

改进 RT-DETR 的无人机图像目标检测算法

姜贸翔[1]，司占军[1]，王晓喆[2]

1.天津科技大学 人工智能学院，天津 300457
2.北京航空航天大学 无人系统研究院，北京 100191

摘 要：针对轻小型无人机图像目标检测中由于目标灵活多样、环境复杂多变导致的检测精度低等问题，本文提出基于改进 RT-DETR 无人机目标检测算法。首先，综合考虑轻量级 SimAM 注意力和倒置残差模块改进 ResNet-r18 主干网络，提高目标检测模型的特征提取能力。其次，采用级联分组注意力机制优化倒置残差模块和特征交互模块，提升特征选择能力，实现目标检测信息的精细化获取。同时，颈部网络中引入 160×160 检测层，提升特征融合阶段小目标的感知能力。最后，基于 VisDrone2019 数据集的实验结果表明，改进后的模型具有更低的参数量和更高的检测精度。在 Alver_Lab_Ulastirma 和 HIT-UAV 数据集上进一步验证了改进方法的有效性和鲁棒性。

关键词：小目标检测；DETR；注意力机制；Transformer；残差链接
文献标志码:A 中图分类号:V279;TP391.41 doi：10.3778/j.issn.1002-8331.2405-0331

Improved Target Detection Algorithm for UAV Images with RT-DETR

JIANG Maoxiang[1], SI Zhanjun[1], WANG Xiaozhe[2]

1. College of Artificial Intelligence, Tianjin University of Science and Technology, Tianjin 300457, China
2. Beihang University, Institute of Institute of Unmanned System, Beijing 100191, China

Abstract: This paper proposes an improved RT-DETR algorithm for unmanned aerial vehicle (UAV) target detection in light and small-sized UAV image targets. Addressing issues such as low detection accuracy due to the flexible and diverse nature of targets and complex and variable environments, the proposed method enhances the feature extraction capability of the detection model by integrating lightweight SimAM attention and inverted residual modules into the ResNet-r18 backbone network. Furthermore, a cascaded group attention mechanism is employed to optimize the inverted residual modules and feature interaction modules, improving feature selection capability and achieving refined acquisition of target detection information. Additionally, a 160×160 detection layer is introduced in the neck network to enhance the perception capability of small targets during the feature fusion stage. Finally, the experimental results based on the VisDrone2019 dataset show that the improved model has lower number of parameters and higher detection accuracy. Further experiments on the Alver_Lab_Ulastirma and HIT-UAV datasets validate the effectiveness and robustness of the proposed improvements.

Key words: Small target detection; DETR; Cascaded Attention; Transformer; Residual link

基金项目：航空科学基金项目（2022Z012051001）。
作者简介：姜贸翔，男，天津科技大学人工智能学院研究生在读，研究方向为计算机视觉；司占军，通信作者，男，硕士研究生，教授/硕士生导师，研究方向：数字图像信息处理技术与虚拟增强现实技术，E-mail： szj@tust.edu.cn；王晓喆，工学博士，讲师，硕士研究生导师，主要研究方向：飞行器设计，气动弹性优化，无人机总体设计。

得益于人工智能在图像处理上的快速发展，近年来，航空图像分析不断提高其处理速度和准确度，成为现代科技应用中的重要组成部分。现今，现代无人机广泛应用于航空摄影、监控、搜索和救援等应用场景[1]。经典的目标检测算法分为两阶段目标检测算法和单阶段目标检测算法。单阶段目标检测算法有更快的推理速度，满足图像分析实时性的要求，能有效应用于无人机目标检测任务[2]。

在无人机航拍图像目标检测任务中，存在大量的小目标检测。小目标通常指的是在图像中占据相对较小比例的目标物体，其尺寸往往不足以提供充足的特征信息以便于准确检测[3]。这种情况在无人机监视任务中尤为常见，例如，需要识别和跟踪人员、车辆或其他小型物体。小目标检测面临的挑战不仅仅是目标尺寸小，还包括目标与背景的对比度低、目标形状不规则、以及目标可能被其他物体遮挡等问题。此外，无人机的运动也会引入图像模糊和运动模糊，进一步增加目标检测的难度。传统的目标检测方法往往无法有效应对这些挑战，因此需要针对小目标检测场景设计新的算法和技术。肖粲俊等人[4]引入自适应特征模块，并在模型内添加注意力机制实现对目标特 征的定位提取能力。李晓欢等人[5]提出一种视觉增强和特征加权的雷视融合车辆目标检测方法，基于特征加权的雷视融合车辆目标检测方法，对毫米波雷达特征图与视觉特征图进行加权融合，提高了低光照环境下的检测精度。郎磊等[6]提出基于YOLOX-Tiny 的轻量级遥感目标检测网络提高优化网络和多尺度预测方法来增强遥感图像检测。Wang等人[7]针对无人机图像的小目标特点，在YOLOv8的Neck中嵌入小目标检测结构STC（small target detection structure），充分捕获全局信息和上下文信息，并引入全局注意力减少采样过程中的特征信息丢失，得到较高的性能提升。

在传统的目标检测方法中，通常使用锚框或候选框来进行目标检测，DETR在于它能够将检测问题转化为无序序列的输出问题，从而将传统的“密集检测”转化为“稀疏检测”[8]。自从DETR(DEtection TRansformer)模型提出以来，Transformer框架在计算机视觉领域得到广泛应用[9]。DETR的核心思想是利用Transformer编码器对整个图像进行编码，然后通过解码器生成一组无序的边界框和相应的类预测。这种顺序输出方法消除了传统密集检测后处理步骤中对阈值滤波和非最大抑制的需求。Transformer模型结合自注意力机制，能够充分利用全局上下文信息，并在处理连续帧序列图像方面表现出显著的性能。文献[10]提出了可变形DETR，其注意力模块重点关注参考区域周围的一小部分关键采样点。可变形DETR能够实现比DETR更好的性能，尤其是在小目标检测方面，并且训练时间较DETR能够减少90%。Group DETR[11]引入多个object queries，既能保留DETR端到端的推理优势，同时还能利用训练中单到多优势来提升性能，加快模型的收敛速度。DAB-DETR[12]算法使用动态锚框作为Transformer解码器中的查询，并对其逐层进行动态更新，消除DETR训练收敛慢的问题。DINO(Detr with improved denoising anchor boxes)[13]模型是一种具有对比去噪训练、混合查询选择和两次前瞻的强端到端Transformer检测器DINO，可以显著提高训练效率和最终检测性能。DINO的“去噪思想”可以提升双边匹配的样本质量，加快训练的收敛速度。

最新的RT-DETR[14]是一个基于Transformer的实时端到端目标检测器，RT-DETR具有相对更少参数量和更低的计算成本。RT-DETR在速度和精度方面都优于同等尺寸的YOLO模型。然而，其在小目标检测方面的表现仍然有待提升。结合上述内容，本文对RT-DETR算法进行了改进，主要贡献包含以下几点：

首先，使用改进后的轻量级的ResNet-r18(Residual Network)[15]主干网络作为本文的主干网络，其中残差模块中集成了轻量级的SimAM(A Simple, Parameter-Free Attention Module，无参数卷积神经网络注意模块)[16] 和改进的iRMB(Inverted Residual Mobile Block)[17]模块；其次，为使模型能够在保留全局信息的同时更多地关注局部细节，通过将级联分组注意力(Cascaded Group Attention，CGA)[18]模块引入到AIFI模块中构成SFCA(Scale wise feature interaction based on cascaded group attention)模块，增强模型的特征选择能力；最后，为增强模型对小目标和密集目标的感知，在RT-DETR的颈部网络中增加了针对小目标的SODL(Small Object Detection Layer)结构。

1 RT-DETR 算法

RT-DETR在同等测试条件下展现出了出色的性能和平衡。RT-DETR算法是一款实用性较高的基于DETR的实时检测器，且解决了关于“两套阈值”的问题，克服了 NMS 对实时检测器推理速度的延迟和对精度的影响。RT-DETR算法和传统的目标检测网络具有相似的结构，结构如图1所示。RT-DETR从主干中提取不同层级的输出，并将它们融合。该高效的混合编码器通过多尺度内特征交互(Attention-based Intrascale Feature Interaction，AIFI)和跨尺度特征融合模块(cross-scale feature-fusion module，CCFM)将多尺度特征转换为一系列图像特征。CCFM由N个RepBlock块组成，双路径输出通过元素级加法进行融合。过程表述如以下公式所示。

$$\begin{aligned} Q &= K = V = Flatten(S_5) \\ F_5 &= RS(Attn(Q,K,V)) \\ Output &= CCFM(\{S_3,S_4,F_5\}) \end{aligned} \tag{1}$$

式中:S3、S4、S5 表示主干网络生成最后三个阶段的特性作为编码器的输入；表示生成查询(Query，Q)、键(Key，K)和值(Value，V)；Attn 表示多头自注意；RS

表示 Reshape 操作，表示将特征的形状恢复为与作为 Flatten 的逆运算的 S5 相同。

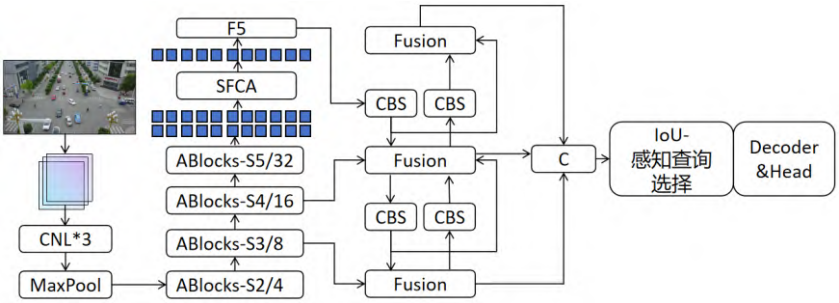


图 1　RT-DETR 模型框架

Fig.1　RT-DETR model framework

RT-DETR采用IoU(Intersection over Union)感知的查询选择来选择固定数量的图像特征作为解码器的初始对象查询。最后，解码器使用辅助预测头，迭代优化对象查询，生成边界框和置信度分数。RT-DETR在分配阶段和计算损失的阶段，分类的标签均使用了“IoU-aware”的设计。模型的分类损失会考虑到预测边界框与真实边界框之间的IoU，"IoU-aware"的训练策略中预测边界框与真实边界框的IoU超过某个阈值时，模型才会将目标的类别标签视为正确的分类。如下公式所示：

$$L(\hat{y},y)=L_{box}(\hat{b},b)+L_{cls}(\hat{c},c,IoU) \qquad (2)$$

式中：$\hat{y}$ 和 y 表示预测和地面真实值，$y=\{c,b\}$ 且 $\hat{y}=\{\hat{c},\hat{b}\}$，c 和 b 分别表示类别和边界框，将 IoU 评分引入到分类分支的目标函数中，以实现对正样本的分类和定位的一致性约束。

2　算法原理

2.1　算法介绍

结合无人机航拍图像目标检测的需求，提出适用于无人机场下的 EMRT-DETR 模型(Multi-scale RT-DETR Target Detection Model in UAV scene，无人机场景下多尺度的 RT-DETR 目标检测模型)。针对 RT-DETR 在小目标的处理方面的不足，引入适合图像特征的注意力机制，并且通过增加 Transformer 层数，以增强模型的表达能力，以更好地利用 Transformer 架构执行目标检测任务。EMRT-DETR 模型结构如图 2 所示。

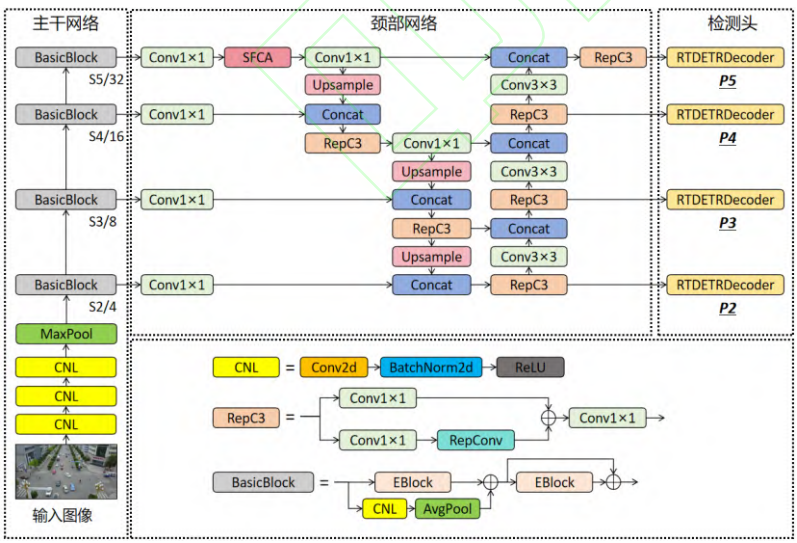


图 2　EMRT-DETR 算法流程实现

Fig.2　Implementation of EMRT-DETR algorithm process

2.2　EBlock 模块

无人机探测中存在大量小目标和大量密集目标，为此需要进一步增强目标检测模型的特征提取能力。相比于传统的主干网络中使用的残差模块，iRMB是传统残差模块的改进版本，可以更好地捕获特征之间的关系。在此基础上进一步对iRMB进行了改进，改进后的iRMBC(Inverted Residual Mobile Block and Cascaded Group Attention)模块可以更好地捕捉图像中的详细信息。通过集成 SimAM 和 iRMBC 模块构成

EBlock(Enhanced Block)模块。EBlock模块具有更小的参数量，同时可以提高骨干网络的特征提取能力，结构如图3所示。

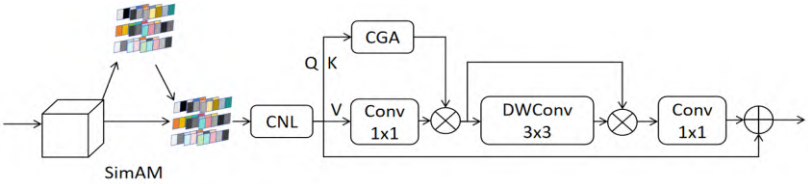


图 3 EBlock 模块

Fig.3 EBlock module

注意力模块在深度学习中被广泛应用，以提高特征提取能力。SimAM模块为每个神经元定义了一个能量函数，有效地帮助模型学习小目标的特征，抑制非目标背景干扰，增强特征映射的表示能力，从而提高模型检测精度，弥补了RT-DETR算法在小目标处理上的不足。SimAM是一个轻量级的注意模块，可以自适应地调整特征图中不同位置的重要性，并通过计算相似度得分和加权特征来提高目标检测的准确性。为了更好地实现注意力，找到重要神经元最简单的方法是测量神经元之间的线性可分性。在SimAM模块中定义的能量函数，如式(3)所示。SimAM注意机制模块不需要在原始网络中添加参数，而是在一层中推断出特征图的三维注意权值。

$$e_t(w_t,b_t,y,x_i)=\left(y_t-\hat{t}\right)^2+\frac{1}{M-1}\sum_{i=1}^{M-1}(y_o-\hat{x_i})^2 \quad (3)$$

式中：$\hat{t}=w_t t+b_t$；$\hat{x_i}=w_t x_i+b_t$ 是 t 和 x_i 的线性变换，其中 t 和 x_i 是输入特征 $X\in R^{C\times H\times W}$ 的单一通道中的目标神经元和其他神经元。i 是空间维度上的指数，$M=H\times W$ 是该通道上的神经元数量。w_t 和 b_t 是变换的权重和偏差。最小化上述方程相当于训练同一通道内神经元 t 与同一通道内其他神经元之间的线性可分性。事实上，由于同一通道内的所有神经元都遵循相同的分布，因此可以预先计算 H 和 W 维数中输入特征的均值和方差，以避免冗余计算。最终得到以下公式：

$$e_t(w_t,b_t,y,x_i)=\frac{1}{M-1}\sum_{i=1}^{M-1}(-1-(w_t x_i+b_t))^2+\left(1-(w_t t+b_t)\right)^2+\lambda w_t^2 \quad (4)$$

每个通道都有M个能量函数。求解方程计算是繁重的。可以通过以下方法很容易地得到线性变换的权重和偏差，线性变换的权重 w_t 的解：

$$w_t=-\frac{2(t-\mu_t)}{(t-\mu)^2+2\dfrac{1}{M-1\sum_i^{M-1}(x_i-\mu_i)^2}+2\lambda} \quad (5)$$

偏差 b_t 的解：

$$b_t=-\frac{1}{2}(t+\mu_t)w_t \quad (6)$$

式(5)和(6)中：μ_t 为除t外的所有神经元的均值和方差：

$$\mu_t=\frac{1}{M-1}\sum_{i=1}^{M-1}x_i \quad (7)$$

通过计算 w_t 和 b_t 的解析解，以及通道中所有神经元的均值和方差，得到公式为：

$$e_t^*=\frac{4(\hat{\sigma}^2+\lambda)}{(t-\hat{\mu})^2+2\hat{\sigma}^2+2\lambda} \quad (8)$$

式(8)中：e_t^* 的值越小，目标神经元 t 附近与周围神经元的区别就越大。因此，它被赋予更高的权重，表明其在视觉处理中的重要性更大。在这种情况下，$1/e_t^*$ 的值越大，意味着每个神经元的重要性越高。因此，每个神经元的重要性都可以通过评估 $1/e_t^*$ 得出。

$\hat{\mu}$ 代表均值：

$$\hat{\mu}=\frac{1}{M}\sum_{i=1}^{M}x_i \quad (9)$$

$\hat{\sigma}^2$ 代表方差：

$$\hat{\sigma}^2=\frac{1}{M}\sum_{i=1}^{M}(x_i-\hat{\mu})^2 \quad (10)$$

综上所述SimAM模块可以描述如下：

$$X^*=sigmoid(\frac{1}{E})\otimes X \quad (11)$$

其中 E 在通道和空间维度上将 e_t^* 分组，sigmoid 函数用于避免权值过大。

无人机获得的图像中的目标往往相对较小，如小型车辆、行人等。由于目标的尺寸较小，其在图像中所占据的像素数量有限，这增加了目标检测的难度。iRMB模块结合了多头自注意力和深度可分离卷积(Depth-wise Conv，DWConv)，该模块类似于轻量级CNN 中的反向残差结构。反向残差结构使用扩展层将特征维度映射到更高维度的空间，并使用深度可分卷积在更高维度的空间中获取更多信息。这种设计综合考虑CNN 的局部特征建模效率和类似Transformer的动态建模能力来学习远距离交互，并取得取得平衡，且利用动态全局建模和静态局部信息融合的优势，同时有效地提高模型的接受域，提高其处理下游任务的能力。为了更有效地捕获远程交互和本地上下文信息，增强模型捕获全局和本地信息的能力，将iRMB模块与

级联分组注意力相结合构成iRMBC模块。

级联分组注意力模块用于关注不同目标和它们之间的信息交换，该模块如图4所示。级联分组注意模块在几个方面与之前的多头自我注意(multi-head self-attention，MHSA)有所不同[14]。首先，在生成查询(Query，Q)、键(Key，K)和值(Value，V)之前，它将注意力头分成几组。这种分组允许每个组内的注意力头注意特征图中特定的通道子集。此外，为了学习更多样化的特征映射，增加模型容量，级联分组注意力模块将每个注意头的输出添加到同一组内的下一个注意头的输入中。这种迭代的细化过程有助于获取更详细的信息，并增强网络的鉴别能力。最后，将多个注意力头的输出连接在一起，并通过一个线性层获得最终的输出。如下所示：

$$\begin{aligned}\tilde{X}_{ij} &= Attn(X_{ij}W_{ij}^{Q}, X_{ij}W_{ij}^{K}, X_{ij}W_{ij}^{V}) \\ \tilde{X}_{i+1} &= Concat[\tilde{X}_{ij}]_{j=1:h}W_{i}^{P} \\ X_{ij}^{'} &= X_{ij} + \tilde{X}_{i(j-1)}, 1< j \le h\end{aligned} \tag{12}$$

式中:第 j 个头计算在 X_{ij} 上的自我注意力，这是输入特征 X_i 的第 j 次分割，h 是头的总数，W_{ij}^{Q} W_{ij}^{K} W_{ij}^{V} 是映射输入特征分成不同子空间的投影层，W_{i}^{P} 是一个线性层，它将连接的输出特征投影到与输入一致的维度。$X_{ij}^{'}$ 表示将每个头部的输出添加到随后的头部中，以逐步细化特征。

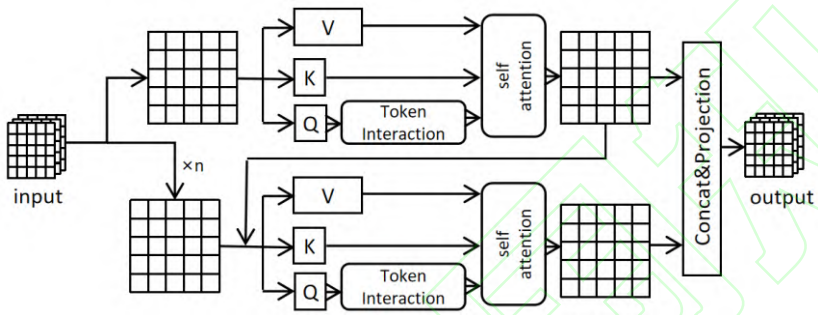


图 4　级联分组注意力模块

Fig.4　Cascaded Group Attention

2.3　改进颈部网络

在无人机图像目标检测任务中存在大量的小目标检测。小目标是在图像中占据相对较小比例的目标物体，其尺寸往往不足以提供充足的特征信息以便于准确检测。在特征经过主干网络后，由于经过的卷积少，低层特征分辨率更高，往往包含更多位置信息和细节信息，但是其语义更低，噪声更多。高层级特征具有更强的语义信息，但知道的细节少。颈部的设计是为了更好地利用由骨干提取的特征，对不同阶段提取的特征进行重新处理和合理利用。为了增强模型对小目标和密集目标的感知，对 RT-DETR 中的颈部网络进行改进，具体做法是在网络中增加了针对小目标的检测层，称为 SODL(Small Object Detection Layer)结构。为缓解 SODL 结构会增大模型参数和计算量的问题，将 RT-DETR 的 Deconder 的隐藏层维度从默认的 256 层降低为 128 层。从表 1 中可以看出减少 Deconder 中的隐藏层维度可以有效降低模型参数和计算量。

表 1　测试集中的 TIDE 指标

Table 1　TIDE indicators in the test set

模型	隐藏层维度	参数量(M)	计算量(G)
EMRT-DETR	256	16.7	88.2
EMRT-DETR	128	13.8	68.1

2.4　SFCA 模块

为使网络能够更好的在不同的空间尺度上进行特征选择，本文使用级联分组注意力改进 AIFI 模块得到 SFCA 模块，如图 5 所示。作为 Transformer 模型的 Encoder 部分，SFCA 模块使网络能够在不同的空间尺度上进行精细的特征选择。通过引入级联分组注意力，使该模块将注意力权重细化到更小的空间尺度，从而捕获更详细的信息。通过这种方法，SFCA 模块可以提高模型对致密和小物体的感知。它可以自适应地调整注意力的权重，使网络能够更好地关注重要的特征。模型可以更准确地定位和识别目标，从而提高目标检测精度。

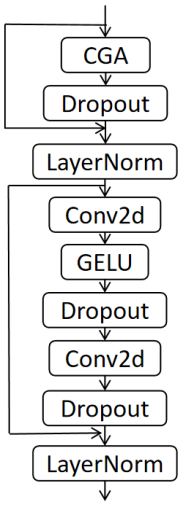


图 5 SFCA 模块

Fig.5 SFCA module

3 实验结果与分析

3.1 数据集

VisDrone2019[19]数据集提供丰富的真实世界图像和视频数据。这些数据包含复杂的场景和各种物体，为研究人员提供了丰富的实验对象和挑战性任务。数据集由10209幅静态图像(6471张用于训练，548张用于验证，3190张用于测试)组成，由各种无人机摄像头捕获，覆盖范围广泛，包括位置、环境、物体和密度，如图6所示。该数据集含有大量的小目标，且复杂的环境因素和目标遮挡严重的情况。此外，数据集还涵盖了丰富的场景和背景，如城市道路、建筑物等，提供了更真实、复杂的测试环境。数据集中的大量小型目标和模糊物体，对于无人机场景下的目标检测具有重要意义，有助于评估算法在实际应用中的性能和鲁棒性。

红外热图像 HIT-UAV[20]数据集上的实验进一步证明了改进方法的泛用性，如图 7 所示。包含从 43470 帧中提取的 2898 张红外热图像，由无人机从不同的场景(学校、停车场、道路、操场等)捕获，涵盖了广泛的方面，如物体(人、自行车、汽车、其他车辆)、飞行高度数据(从 60 到 130 米)、相机透视数据(从 30 到 90 度)和阳光强度(白天和晚上)。数据集包含包含大量小型目标和模糊物体的特点，对于无人机场景下的目标检测和跟踪具有重要意义。

最后在 ALU(Alver_Lab_Ulastirma)[21]数据集中进行实验，其中包含 1593 张无人机航拍图像，其中 1187 张用于训练、117 张用于验证和 50 张用于测试。共 6 个类别，分别为公共汽车、长途卡车、小型货车、摩托车、卡车和车辆，如图 8 所示。

图6 VisDrone2019实例

Fig.6 VisDrone2019 example

图7 HIT-UAV的实例

Fig.7 HIT-UAV example

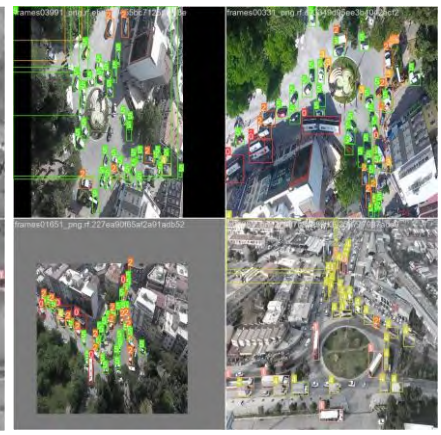


图8 ALU实例

Fig.8 ALU example

3.2 超参数

如表 2 所示，网络实验环境为 Windows 11，GPU 使用 RTX 4090(24GB)，CPU 使用的 i9 13900k 24(8+16)核 32 线程处理器，CUDA 版本为 11.8，PyTorch 版本为 2.1.0，Python 版本为 3.9.18。

表 2 实验设备

Table 2 Experimental equipment

Parame-ters	Configuration
CPU	i9-13900K 3.00 GHz
GPU	NVIDIA GeForce RTX 4090（24g）
Python	3.9.18
torch	2.1.0
CUDA	11.8

如表 3 所示，表中展示了研究中使用的主要超参数。实验在均不使用预训练模型的情况下进行。实验

中的超参数中训练批次（batch）设置为 4，小批次训练可以促使模型学习到更一般化的特征，因为每个批次的数据样本都是从整个数据集中随机抽取的，有助于模型更好地泛化到新数据。合适的训练轮数(Epoch)可以帮助提升模型的泛化能力，即在未见过的数据上的表现。通过验证集的表现来确定最佳的训练轮数，通常可以找到折中点，避免过拟合同时又能提升性能。过大的训练轮数也意味着增加了训练的时间和计算成本。在实际应用中，需要权衡训练时间和模型性能之间的关系。基于此在 VisDrone2019 数据集的消融实验中使用 150 轮的训练轮数进行实验，以此验证不同组件和改进方法对检测性能的影响。优化器选择 SGD，SGD 在 Transformer 模型中能够提供快速收敛、高效的参数更新和优化，以及广泛的调优可能性。640×640 像素的图像需要的内存和计算资源更少，对于大多数目标检测模型来说是比较平衡的选择，可以提供足够的信息量，使得模型能够准确地检测和定位物体。

表 3　实验参数

Table 3　Experimental parameters

超参数	设置
训练批次	4
训练轮数	150/250
优化器	SGD
并行工作线程数	8
图像尺寸	640×640

3.3　评价指标

为了全方位评估模型的性能，实验设置的主要评价指标为 Parameters(参数量)、GFLOPs(计算量)、mAP@0.5、mAP@0.5:0.95。

精确率是指所有预测为正样本的结果中被预测正确的比例；召回率是指所有的正样本当中，被正确的预测为正样本的比例，精确率计算公式如下：

$$p = \frac{TP}{TP + FP} \tag{13}$$

式中: p (Precisions)表示精确率；真阳性 TP、真阴性 TN、假阳性 FP 和假阴性 FN 为混淆矩阵的指标。

$$R = \frac{TP}{TP + FN} \tag{14}$$

式中: R (Recall)表示召回率。

AP (Average Precision)是通过计算 P-R 曲线下的面积来评估模型在目标检测任务中的性能，反映了模型在不同召回率下的准确率平均水平，计算公式如下：

$$AP = \int_0^1 \mathrm{p}(r)dr \tag{15}$$

式中: $\mathrm{p}(r)$ 表示在每个召回率点处的最大准确率，r 表示召回率。

mAP 是用 p 和 R 作为两轴作图后围成的面积，m 表示平均，@后面的数表示判定 IoU 为正负样本的阈值，@0.5:0.95 表示阈值取 0.5:0.05:0.95 后取均值。计算公式如下：

$$\mathrm{m}AP = \frac{\sum_{\mathrm{i}=1}^{k} AP_i}{\mathrm{K}} \tag{16}$$

TIDE（Toolkit for Identifying Detection and segmentation Errors）[22]是用于目标检测和实例分割任务的失效分析工具。它旨在帮助研究人员识别目标检测和分割任务中的错误，并深入了解这些错误的原因。在实际应用中，很多看似是背景错误或定位错误的情况实际上是由于 GT(ground truth)数据没有正确标注或标注错误所导致的。使用 IoU^{max}来表示假阳性与给定类别的 GT 的最大 IoU 重合度。前景 IoU 临界值用 t^f 表示，背景临界值用 t^b表示，这两个临界值分别设置为 0.5 和 0.1。具体指标如下：

(1) CLs（Classification Error，分类错误）：

$\mathrm{IoU}^{max} \geq t^f$ 对于错误类别的 box，定位正确但分类错误。

(2) Loc（Localization Error，定位误差）：

$t^b \leq \mathrm{IoU}^{max} \leq t^f$ 对于正确类别的 box，分类正确但定位不正确。

(3) Both（Both Cls and Loc Error，Cls 和 Loc 错误）：$t^b \leq \mathrm{IoU}^{max} \leq t^f$ 对于错误类别的 box，分类错误和定位错误。

(4) Dupe（Duplicate Detection Error，重复检测错误）：$\mathrm{IoU}^{max} \geq t^f$ 分类正确，但是有另外一个置信度更高、分类正确的 Bounding Box。

(5) Bkg（Background Error，背景误差）：

$\mathrm{IoU}^{max} \leq t^b$ for all box（检测到的背景为前景）。

(6) Miss（Missed GT Error，假阴性错误）：除了分类错误和定位误差以外，所有没有检测到的 GT。

3.4　性能评估

3.4.1　消融实验

研究使用 VisDrone2019 数据集进行了详细的消融实验。表 4 所示为 RT-DETR 模型为基准模型的消融实验结果。图 9 所示为表 4 中实验编号 A、C、E 和 F 的 P、R、mAP@0.5 和 mAP@0.5-0.95 的变化曲线。实验编号 A、B、C、D、E、F 的训练轮数设置为 150 轮。实验 A 为 RT-DETR 的实验结果。从实验 B 和 C 可以看出在主干网络中增加了 SimAM 注意力模块和 SFCA 模块后，增强了主干网络的特征选择能力和特征选择能力。从实验 D 和 E 可以看出对 iRMB 模块的改进是有效的，不仅进一步降低了模型的参数量，检测精度和收敛能力也得到进一步提升。从实验 E 可以

看出在引入 iRMBC 模块、SimAM 注意力模块和 SFCA 模块后进一步细化特征提取过程，并增强网络的特征选择能力，使网络的特征提取能力得到增强，检测精度 mAP@0.5 提升至 46.3。从实验 F 可以看出，通过加入 SODL 结构改进颈部网络，提升了模型捕获小目标的能力，改进模型的检测精度 mAP@0.5 提升至 48.8。与 RT-DETR 相比改进后模型的参数量和模型大小下降了 31%，增强了模型在实际应用的可部署性。消融实验证明了改进方法增强了模型的学习能力，同时具有更强的可部署性。在使用 RTX 4090 进行模型的 FPS 测试时，实验结果显示改进后的模型的检测速度有所降低。

表 4 VisDrone2019 消融实验结果

Table 4 VisDrone2019 ablation experiment results

编号	模型	训练轮数	参数量(M)	计算量(G)	P	R	mAP@0.5	mAP@0.5-0.95	FPS(bs=1)	模型大小(M)
A	RT-DETR	150	20.0	57.3	59.5	42.9	44.8	27.1	102	40
B	A+SimAM	150	20.0	57.3	59.3	43.4	45.0	27.1	101	40
C	B+SFCA	150	19.8	57.4	60.0	43.5	45.5	27.5	102	39
D	C+iRMB	150	16.3	49.5	60.4	44.2	45.8	27.9	78	33
E	C+iRMBC	150	15.2	47.2	61.0	44.7	46.3	28.2	54	31
F	E+SODL	150	13.8	68.1	62.0	46.6	48.8	30.5	52	28

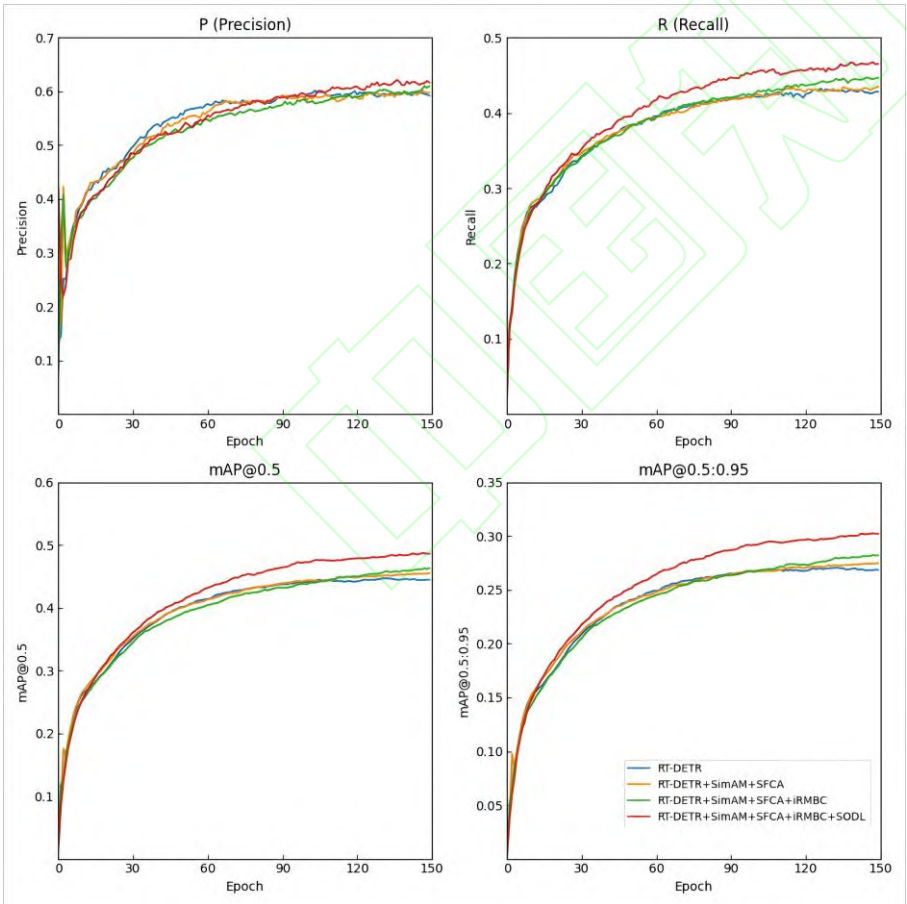


图9 消融实验变化曲线

Fig.9 Change curve of the ablation experiment

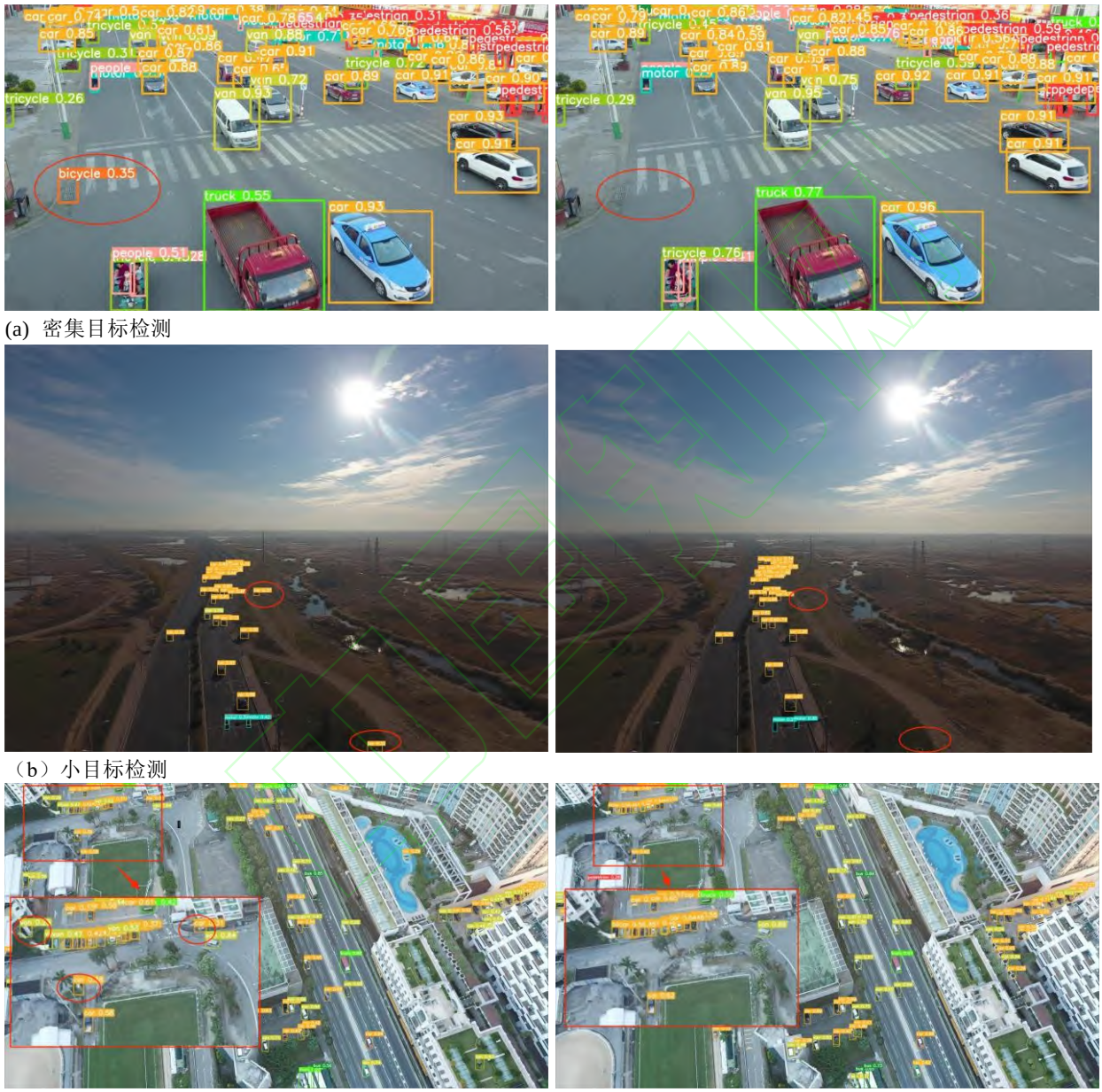
综上所述，消融实验可以看出，改进后的模型虽然检测速度下降，但却带来显著的模型精度提升和模型参数的减少。这些优化可以显著改善无人机图像目标检测任务的表现，使其更适应各种复杂的现实应用场景。

通过观察 RT-DETR 的结果，发现模型正确分类目标后，定位目标时却存在偏差或错误。此外，模型会在同一目标周围生成多个重叠的边界框，导致重复检测。如表 5 所示，可以看出 EMRT-DETR 相比 RT-DETR 优化了分类正确但定位不准确的情况，同时有效改善重复检测的错误情况。

表5 测试集中的TIDE指标

Table 5 TIDE indicators in the test set

模型	Cls	Loc	Both	Dupe	Bkg	Miss
RT-DETR	20.8	4.1	0.5	0.17	2.3	13.4
EMRT-DETR	20.9	3.7	0.5	0.14	2.4	13.9



(a) 密集目标检测

（b）小目标检测

（c）小目标检测

图10 检测效果比较

Fig.10 Comparison of detection effects

3.4.2 可视化实验与结果分析

如图 10 所示，呈现了两种模型在不同场景下的检测结果，左侧为表 4 中实验组 A 的可视化实验结果，右侧为表 4 中实验组 F 的可视化实验结果。在实验章节中，通过对 RT-DETR 基线模型与 EMRT-DETR 模型的全面可视化比较，展示了它们在 VisDrone2019 数据集上的性能差异。左侧展示了 RT-DETR 基线模型的检测结果，右侧展示了 EMRT-DETR 模型的检测结果。分别呈现复杂背景中的密集和高空条件下，模型改进前后的可视化效果。可以看出，红色标示在 RT-DETR 基线模型中存在的错检情况，但在 EMRT-DETR 模型中未出现。表明 EMRT-DETR 模型相比基线模型在误差检测方面的性能优势。同时，观察到 EMRT-DETR 模型在检测精度方面有所提升。

综上所述，在稀疏小目标场景中改进模型能有效

减少漏检和误检情况。有效地改善了 RT-DETR 对无人机视角下小目标的检测性能，进一步验证本文改进方法的有效性。

3.4.3 泛化性实验

表6和表7分别为红外热图像HIT-UAV数据集和ALU无人机航拍图像数据集上的泛化性实验结果。

表 6 HIT-UAV 数据集的实验结果

Table 6 Experimental results of the HIT-UAV dataset

模型	P	R	mAP@0.5	mAP@0.5:0.95
RT-DETR	84.1	72.5	75.6	47.5
EMRT-DETR	88.2	73.2	78.2	50.9

表 7 ALU 数据集的实验结果

Table 7 Experimental results of the ALU dataset

模型	P	R	mAP@0.5	mAP@0.5:0.95
RT-DETR	61.2	60.5	63.9	43.3
EMRT-DETR	61.4	60.2	64.8	45.2

如表6所示，实验证明改进方法对于红外热图像数据集能够实现有效的性能提升。RT-DETR的P、R、mAP@0.5、 mAP@0.5：0.95指标，改进后分别提升了5%、1%、3%、7%，证明了本文改进方法的泛用性。

如表7所示，实验证明在其他无人机视角下的数据集上改进算法也同样有效。可以明显观察到经过改进的EMRT-DETR模型的mAP@0.5：0.95指标提升了4%，证明了改进方法的有效性。

3.4.4 对比实验

在 VisDrone2019 数据集中将改进模型与YOLOv3-tiny、YOLOv3[23]、YOLOv5[24]、YOLOv6n[25]、文献[26]、Faster R-CNN[27]、SSD[28]、RDS-YOLOv5[29]、THP-YOLOv5[30]、Deformable-DETR[10]、YOLOv8 等目标检测算法进行比对。Faster R-CNN、SSD 和 Deformable-DETR 为经典目标检测算法。文献[26]和 RDS-YOLOv5 为 YOLO 系列算法在无人机图像目标检测的最新变体。THP-YOLO 为 VisDrone 挑战中 YOLO 系列算法的 SOTA 算法，在本实验中引用 THP-YOLO 作者最新发布的代码在 YOLOv8 环境中进行复现。实验结果如表 8 所示，改进后的 RT-DETR 模型取得了显著的检测精度提升，达到了最高水平。模型在 mAP 指标上显著超过 YOLO 系列模型。尽管改进模型在精度上表现出色，但其运行时的帧率（FPS）略低于 YOLO 系列模型。改进模型在帧率上超过 DETR 系列中的 Deformable-DETRMO 模型。虽然改进模型在速度上有牺牲，但改进模型仍能在实时或近实时的应用场景中提供符合要求的检测速度，显示出改进模型在精度和速度之间做出了有效的权衡。

表 8 对比实验结果

Table 8 Comparative experimental results

模型	Param(M)	GFLOPs	FPS(bs=1)	ModelSize(MB)	mAP@0.5	mAP@0.5-0.95
YOLOv3-tiny	12.1	18.9	800	23.8	24.2	13.5
YOLOv3[23]	103.7	282.3	100	791.9	45.4	27.9
YOLOv5n[24]	2.5	7.1	320	5.1	31.7	18.3
YOLOv5m-SPPF	25.1	64.0	220	48.1	39.5	23.6
YOLOv6n[25]	4.2	11.8	330	8.3	29.4	17.1
YOLOv6m[25]	52.0	161.2	160	99.5	40.9	24.9
文献[26]	-	38.5	-	24.2	45.7	27.7
Faster R-CNN[27]	-	-	-	106.1	33.0	17.0
SSD[28]	-	62.7	-	26.3	24.2	10.7
RDS-YOLOv5[29]	14.0	38.5	-	-	46.9	29.2
THP-YOLOv5n[30]	2.7	11.8	138	5.6	35.6	21.0
THP-YOLOv5m[30]	26.3	85.0	120	50.7	47.0	28.9
Deformable-DETR[10]	-	196	29	40	43.1	27.1
YOLOv8n	3.0	8.1	365	6.0	33.1	19.2
YOLOv8s	11.7	28.5	327	22.0	38.3	22.7
YOLOv8m	25.8	78.7	214	50.8	41.8	25.4
YOLOv8l	43.6	164.9	145	85.6	43.8	27.0
YOLOv8x	68.1	257.4	108	133.5	44.7	27.6
RT-DETR	20.0	57.3	102	39	44.7	27.1
EMRT-DETR(ours)	13.8	68.1	52	28	48.8	30.5

4 结束语

针对轻小型无人机图像目标检测中由于目标灵活多样、环境复杂多变导致的检测精度低等问题提出 EMRT-DETR 模型。首先，通过引入 SimAM 注意力机制有效提升模型的特征提取能力，通过将倒残差模块

与级联分组注意力相结合增强模型捕获全局和本地信息的能力，最后构成了 EBlock 模块。其次，通过使用级联分组注意力模块优化 AIFI 模块，构成的 SFCA 模块使网络能够更好的在不同空间尺度上进行特征选择。最后，通过在颈部网络中增加的 SOLD 结构有效提升模型对于小目标敏感度。经过实验和改进方法的探究发现相较于 RT-DETR 模型，虽然检测速度有所降低，但改进后模型可以更快速的收敛，同时具备更强的可部署性和更强的检测精度。

经过对 RT-DETR 的研究，RT-DETR 模型展现出的极强的泛用性和应用潜力，相比 YOLO 系列算法模型自身依然存在收敛困难和模型体积大的问题。在无人机图像目标检测任务中需要更加轻量和精度更高的目标检测模型。基于此目的改进后的模型在计算量上有所增加，导致检测速度一定程度的下降，但考虑到检测性能和模型体积上的优化，这种权衡是值得的。后续可以通过部署和优化进一步提升检测速度。未来研究将继续致力于进一步优化模型的结构和算法，以提高目标检测系统的实时性和自适应性。尽管 Transformer 模型在目标检测中取得了显著进展，但其训练和推理时间延长的问题仍然存在。相信在进一步的研究和探索中，Transformer 模型在计算机视觉领域将发挥更大的作用。通过优化模型结构、改进训练策略以及利用硬件加速等手段有望克服这些挑战，为目标检测任务带来更好的性能和效果。

参考文献:

[1] CHENG N, WU S, WANG X, et al. AI for UAV-assisted IoT applications: A comprehensive review[J]. IEEE Internet of Things Journal, 2023, 10(16): 14438-14461.

[2] AL-LQUBAYDHI N, ALENEZI A, ALANAZI T, et al. Deep learning for unmanned aerial vehicles detection: A review[J]. Computer Science Review, 2024, 51: 100614.

[3] WU H, ZHU Y, LI S. CDYL for infrared and visible light image dense small object detection[J]. Scientific Reports, 2024, 14(1): 3510.

[4] 肖粲俊,潘睿志,李超,等.基于改进 YOLOv5s 绝缘子缺陷检测技术研究[J/OL].电子测量技术:1-7[2023-01-11].
XIAO C J, PAN R Z,LI C, et al.Research on defect detection technology based on improved YOLOv5s insulator[J/OL].Electronic Measurement Technology:1-7 [2023-01-11].

[5] 李晓欢,霍科辛,颜晓凤,等.基于特征加权视觉增强的雷视融合车辆检测方法[J].公路交通科技，2023,40(2):182-189.
LI X H, HUO K X, YAN X F, et al. Thunder-vision fusion vehicle detection method based on feature-weighted visual enhancement [J]. Highway Traffic Technology, 2023,40 (2): 182-189.

[6] 郎磊，刘宽，王东. 基于 YOLOX-Tiny 的轻量级遥感图像目标检测模型[J].激光与光电子学进展,2023,60(2):0228004.
LANG L,LIU K,WANG D.Lightweight Remote Sensing Object Detector based on YOLOX-Tiny, 2023,60(2): 0228004.

[7] WANG F, WANG H, QIN Z, et al. UAV target detection algorithm based on improved YOLOv8[J]. IEEE Access, 2023, 11: 116534-116544.

[8] DAI X, CHEN Y, YANG J, et al. Dynamic detr: End-to-end object detection with dynamic attention[C]// Proceedings of the IEEE/CVF international conference on computer vision. 2021: 2988-2997.

[9] HAN K, WANG Y, TIAN Q, et al. Ghostnet: More features from cheap operations[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 1580-1589.

[10] ZHU X, SU W, LU L, et al. Deformable detr: Deformable transformers for end-to-end object detection[J]. arXiv preprint arXiv:2010.04159, 2020.

[11] CHEN Q, CHEN X, ZENG G, et al. Group detr: Fast training convergence with decoupled one-to-many label assignment[J]. arXiv preprint arXiv:2207.13085, 2022.

[12] LIU S, LI F, ZHANG H, et al. Dab-detr: Dynamic anchor boxes are better queries for detr[J]. arXiv preprint arXiv:2201.12329, 2022.

[13] ZHANG H, LI F, LIU S, et al. Dino: Detr with improved denoising anchor boxes for end-to-end object detection[J]. arXiv preprint arXiv:2203.03605, 2022.

[14] LV W, XU S, ZHAO Y, et al. Detrs beat yolos on real-time object detection[J]. arXiv preprint arXiv:2304. 08069, 2023.

[15] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[16] YANG L, ZHANG R Y, LI L, et al. Simam: A simple, parameter-free attention module for convolutional neural networks[C]//International conference on machine learning. PMLR, 2021: 11863-11874.

[17] ZHANG J, LI X, LI J, et al. Rethinking mobile block for efficient attention-based models[C]//2023 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE Computer Society, 2023: 1389-1400.

[18] LIU X, PENG H, ZHENG N, et al. Efficientvit: Memory efficient vision transformer with cascaded group attention[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 14420-14430.

[19] CAO Y, HE Z, WANG L, 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.

[20] SUO J, WANG T, ZHANG X, et al. HIT-UAV: A high-altitude infrared thermal dataset for Unmanned Aerial Vehicle-based object detection[J]. Scientific Data, 2023, 10(1): 227.

[21] new. (2024). Alver_Lab_Ulastirma Dataset. Roboflow Universe. [Online]. Retrieved from https://universe. roboflow.com/new-0ikav/alver_lab_ulastirma (accessed on 2024-04-29).

[22] BOLYA D, FOLEY S, HAYS J, et al. Tide: A general toolbox for identifying object detection errors[C]// Computer Vision–ECCV 2020: 16th European Con- ference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16. Springer International Publishing, 2020: 558-573.

[23] REDMON J, FARHADI A. Yolov3: An incremental improvement[J]. arXiv preprint arXiv:1804.02767, 2018.

[24] WU W, LIU H, LI L, et al. Application of local fully Convolutional Neural Network combined with YOLO v5 algorithm in small target detection of remote sensing image[J]. PloS one, 2021, 16(10): e0259283.

[25] LI C, LI L, JIANG H, et al. YOLOv6: A single-stage object detection framework for industrial applications[J]. arXiv preprint arXiv:2209.02976, 2022.

[26] 潘玮,韦超,钱春雨,等.面向无人机视角下小目标检测的YOLOv8s 改进模型[J].计算机工程与应用,2024, 60(9): 142-150.
PAN W, WEI C, QIAN C Y, et al. Improved YOLOv8s Model for Small Object Detection from Perspective of Drones [J]. Computer Engineering and Applications, 2024,60 (9): 142-150.

[27] SEO D M,WOOHJ,KIMMS, et al. Identification of asbes tos slates in buildings based on faster region-based convolu tional neural network (Faster R-CNN) and drone-based aerial imagery[J]. Drones, 2022, 6(8): 194.

[28] LIU W, ANGUELOV D, ERHAN D, et al. Ssd: Single shot multibox detector[C]//Computer Visioner region-based convolu tional neural network (Faster R-CNN) and drone-based aerial imagroceedings, Part I 14. Springer International Publishing, 2016: 21-37.

[29] 陈佳慧,王晓虹.改进 YOLOv5 的无人机航拍图像密集小目标检测算法[J].计算机工程与应用,2024,60(3): 100-108.
CHEN J H, WANG X H. Dense Small Object Detection Algorithm Based on Improved YOLOv5 in UAV Aerial Images[J]. Computer Engineering and Applications, 2024,60 (3): 100-108.

[30] ZHU X, LYU S, WANG X, et al. TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 2778-2788.