OEE | 光电工程 Opto-Electronic Engineering

Article

2026 年 1 月, 第 53 卷, 第 1 期

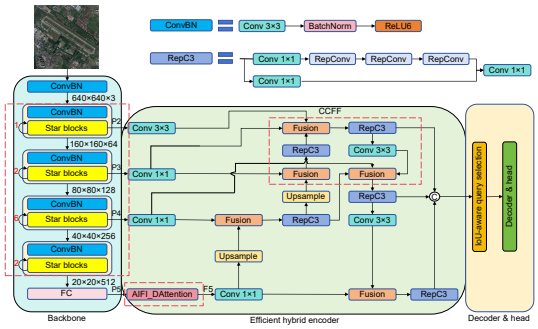
改进 RTDETR 的复杂背景下遥感图像目标检测算法

沈汉滨[1,2], 梅　婷[1,2], 林珊玲[2,3*], 林志贤[1,2,3], 林坚普[2,3], 吕珊红[2,3], 郭太良[1,2,3]

[1]福州大学物理与信息工程学院, 福建 福州 350116;

[2]中国福建光电信息科学与技术创新实验室, 福建 福州 350116;

[3]福州大学先进制造学院, 福建 泉州 362200



摘要：针对遥感图像中存在的背景复杂、目标尺度不一和小目标众多等因素所造成的目标检测精度低的问题，提出了一种改进 RTDETR 的遥感图像目标检测算法。首先引入轻量化主干网络 StarNet，在大幅度降低参数量和计算量的同时提高了网络的特征提取能力。其次在基于注意力的尺度内特征交互模块 (attention-based intra-scale feature interaction, AIFI) 中引入可变形注意力机制 (deformable attention, DAttention)，提高了网络对不同尺度目标的理解能力。最后，在特征融合部分采用加权双向特征金字塔网络 (Bi-directional feature pyramid network, BiFPN) 提高网络的特征融合能力。实验结果表明，改进算法在 DIOR 数据集上 mAP50 和 mAP50-95 指标分别提升了 1.8% 和 1.7%，同时参数量降低了 37.7%。此外，在 RSOD 数据集上进行泛化分析实验，改进算法 mAP50 和 mAP50-95 指标分别提升 1.6% 和 0.5% 且参数量降低 37.9%。改进后的算法在保证较低计算量和参数量的基础上，提升了检测精度，能够更好地应对复杂背景下的遥感图像目标检测任务。

关键词：遥感图像目标检测；RTDETR；轻量化；可变形注意力机制

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

DOI: 10.12086/oee.2026.250064 | CSTR: 32245.14.oee.2026.250064 | 中图分类号：TP391.41 | 文献标识码：A

沈汉滨，梅婷，林珊玲，等. 改进 RTDETR 的复杂背景下遥感图像目标检测算法 [J]. 光电工程，2026，**53**(1): 250064

Shen H B, Mei T, Lin S L et al. Improved RTDETR based remote sensing image object detection algorithm for complex backgrounds[J]. *Opto-Electron Eng*, 2026, **53**(1): 250064

1　引　言

遥感技术将地面物体的电磁波信息转化为图像形式，得到常用的遥感图像。对遥感图像进行分析和解译，可以得到图像中存在的地面物体的目标信息。遥感图像分析在环境检测、智能交通和国土资源管理等多个领域中起着关键作用。遥感图像具有背景复杂、目标类别多样、小目标聚集和目标尺度大小不一等特点，现有的技术不能有效地检测出目标。

传统的目标检测方法难以有效提取目标特征，在应对遥感图像时鲁棒性不足，导致检测效率低下。自 2014 年之后，深度学习有了突破性的发展。深度学习能够有效提取图像中的深层语义信息，具有非常强的数据挖掘和分析能力，因此被广泛应用于目标检测领域。目前，基于深度学习的目标检测算法主要有单阶段算法和两阶段算法两类。单阶段算法以 YOLO (you only look

收稿日期：2025-03-03；修回日期：2025-04-29；录用日期：2025-04-30；网络出版日期：2026-01-25

基金项目：国家重点研发计划 (2021YFB3600603)

*通信作者：林珊玲，sllin@fzu.edu.cn。

版权所有©2026 中国科学院光电技术研究所

once)[1-4] 系列和 SSD (single shot multibox detector)[5] 系列为代表。单阶段算法可以直接获取目标的坐标和类别信息，算法结构简洁，效率较高。两阶段算法以 RCNN (region-based convolutional neural networks)[6] 系列为代表，该算法需要单独训练分类器且输入图像需要固定大小，在检测时每个输入图像分配的候选框过多，导致整体效率太低。

Vaswani 等[7] 于 2017 年提出了 Transformer 算法，自提出之后就被大量运用于自然语言处理领域，该算法在自然语言处理领域取得了很好的效果。Carion 等[8] 提出了 DETR (detection Transformer) 算法，第一次将 Transformer 算法应用于目标检测任务中，该算法通过全局上下文来预测图像中的目标，不需要使用候选框，实现了端到端的目标检测，但该算法计算量消耗过大且训练收敛速度慢。Zhu 等[9] 提出了 Deformable DETR 算法，通过可变形自注意机制解决标准 Transformer 中注意力的问题，加快算法的收敛速度并且降低算法的复杂度；同时又引入了多尺度特征解决小目标检测性能不足的问题，但该方法在复杂背景下鲁棒性不足。Meng 等[10] 提出了 Conditional DETR 算法，通过引入条件交叉注意力机制解决了训练收敛慢的问题，该算法仍然存在对多尺度信息捕获能力不足的问题。Zhang 等[11] 提出了 DINO (DETR with improved denoising anchor boxes) 算法，引入改进的去噪锚框到 DETR 中，在平均精度上取得了显著改进，但该算法在特定任务上的泛化能力太差。Zhao 等[12] 首次提出了一种基于 Transformer 的实时目标检测算法 RTDETR (real-time detection Transformer)，它通过解耦内部尺度交互和跨尺度特征融合来高效地处理多尺度特征，并引入了基于交并比 (intersection over union, IoU) 的查询选择机制，优化了目标查询的初始化过程，大大提高了检测速度，是第一个实时端到端的 DETR 检测算法。金睿蛟等[13] 在 DETR 中引入一种多尺度旋转可变形注意力模块，提升了算法对小目标的检测能力，但该算法对背景复杂度的适应能力较差。肖振久等[14] 通过将感知注意力与移动反向卷积结合形成一个检测头增强对小目标感知能力，但也增加了算法计算量和参数量。白金燕等[15] 通过在 RTDETR 主干网络中引入注意力机制和添加小目标检测检测头，提高了遥感图像小目标检测精度的同时也增加了算法的计算量。周韩莲[16] 等提出了一种复杂背景下的轻量级遥感飞机目标检测算法，通过融合参数化和细节增强提高算法推理速度，但其算法特征提取能力较弱。

综上所述，有许多学者将 Transformer 应用于遥感图像目标检测领域。但是，当前的遥感图像目标检测算法存在不能够有效地提取目标的特征信息、特征融合过程中存在信息丢失和捕捉不同尺寸目标的能力不强等问题，在算法大小和检测精度方面仍然有很大的提升空间。针对遥感图像目标检测尚存在的问题，本文通过对 RTDETR 算法进行改进，提出了一种改进的遥感图像目标检测算法。主要工作如下：

1) 在算法主干网络中引入轻量化主干网络 StarNet[17]，通过元素乘法将输入数据映射到一个极高维的非线性特征空间中，大大降低主干网络参数量且有效提高特征提取能力。

2) 使用可变形注意力机制 (DAttention)[18] 改进高效混合编码器中的基于注意力的尺度内特征交互模块 (attention-based intra-scale feature interaction, AIFI)，在不同尺度上自适应地选择关键信息，增强算法对局部细节和多尺度特征的理解。

3) 在混合编码器的基于 CNN 的跨尺度特征融合模块 (CNN-based cross-scale feature fusion, CCFF) 中采用加权双向特征金字塔网络 (BiFPN)[19] 充分利用低维的细节信息和高维的语义信息，提升了多尺度特征的表达能力。

2 RTDETR 算法介绍

RTDETR 是一种实用性较高的 DETR 系列的实时目标检测器，该算法共有三个模块：主干网络 (backbone)、颈部网络 (neck) 和头部网络 (head)，本文选择以 ResNet-18[20] 为主干网络的 RTDETR 算法作为基准算法。该算法通过主干网络对输入图像分别进行 8 倍、16 倍和 32 倍下采样，三个尺度的输出 P3、P4 和 P5 分别作为下一阶段的输入。RTDETR 算法的颈部网络是由基于注意力的尺度内特征交互和基于 CNN 的跨尺度特征融合组成的高效混合编码器 (efficient hybrid encoder)[21]。AIFI 是单尺度的 Transformer 编码器，对主干网络输出的 P5 层进行学习提取，输出特征被标记为 F5。CCFF 将主干网络输出的 P3、P4 和 P5 三个尺度的特征图进行深度特征融合，以捕捉图像中更为丰富的信息。头部网络使用 IoU 查询选择机制，从混合编码器的输出序列中筛选出一定数量的图像特征作为解码器的初始目标查询。随后，解码器在辅助预测头的帮助下，通过多轮迭代和对象查

询优化，最终生成目标的预测框和相应的置信度得分。RTDETR 算法结构如图 1 所示。

3 本文算法

RTDETR 算法是首种能够在较高精度下达到实时性的 DETR 算法，但是其算法结构仍然存在诸多缺陷，因此本文对 RTDETR 算法进行改进。首先，使用轻量化主干网络 StarNet 改进 ResNet18 主干网络，这一改进能够有效降低算法复杂度和提升主干网络的特征提取能力。其次，在高效混合编码器的 AIFI 中引入 DAttention 改进其标准注意力机制，其灵活的注意力分配机制提高了网络对不同尺度目标的理解能力。最后，在高效混合编码器的特征融合中使用双向特征金字塔网络模块 BiFPN，该改进使算法能够捕捉到图像中的不同尺度特征信息且能有效增强算法的特征融合能力。改进的 RTDETR 算法结构如图 2 所示。上述三处改进在图 2 中分别用红色虚框进行标记。首先在主干网络中使用 ConvBN 和 Star blocks 作为基础模块；其次在 AIFI 中引入了可变形注意力机制；最后在特征融合部分增加了 P2 层且在相同层级之间增加一个额外的输入。

3.1 轻量化特征提取主干网络

在遥感图像目标检测任务中，原 RTDETR 算法使用 ResNet18 作为主干网络，在复杂背景下对目标的关注度和特征的提取能力较弱。此外，ResNet18 残差网络堆叠过深，不仅导致参数量过大，还存在大量冗余计算，大大增加了算法的推理时间，使得算法难以部署在无人机等资源受限设备上。为了解决上述问题，本文使用轻量化网络 StarNet 来替换原来的主干网络 ResNet18，在保持较低计算量和算法复杂度的前提下，有效提升网络对不同目标的特征提取能力。改进后的主干网络结构如图 3 所示。

StarNet 算法是 2024 年微软公司提出的一种基于元素乘法 (element-wise multiplication) 的轻量级网络算法。其通过元素乘法将输入数据映射到一个极高维的非线性特征空间，这种映射方式能够生成丰富的特征表示，使得算法在处理复杂数据时更加有效。其次，StarNet 利用元素乘法在不同通道上进行成对特征乘法，实现特征的非线性组合。这种方法能够在紧凑的特征空间内工作，

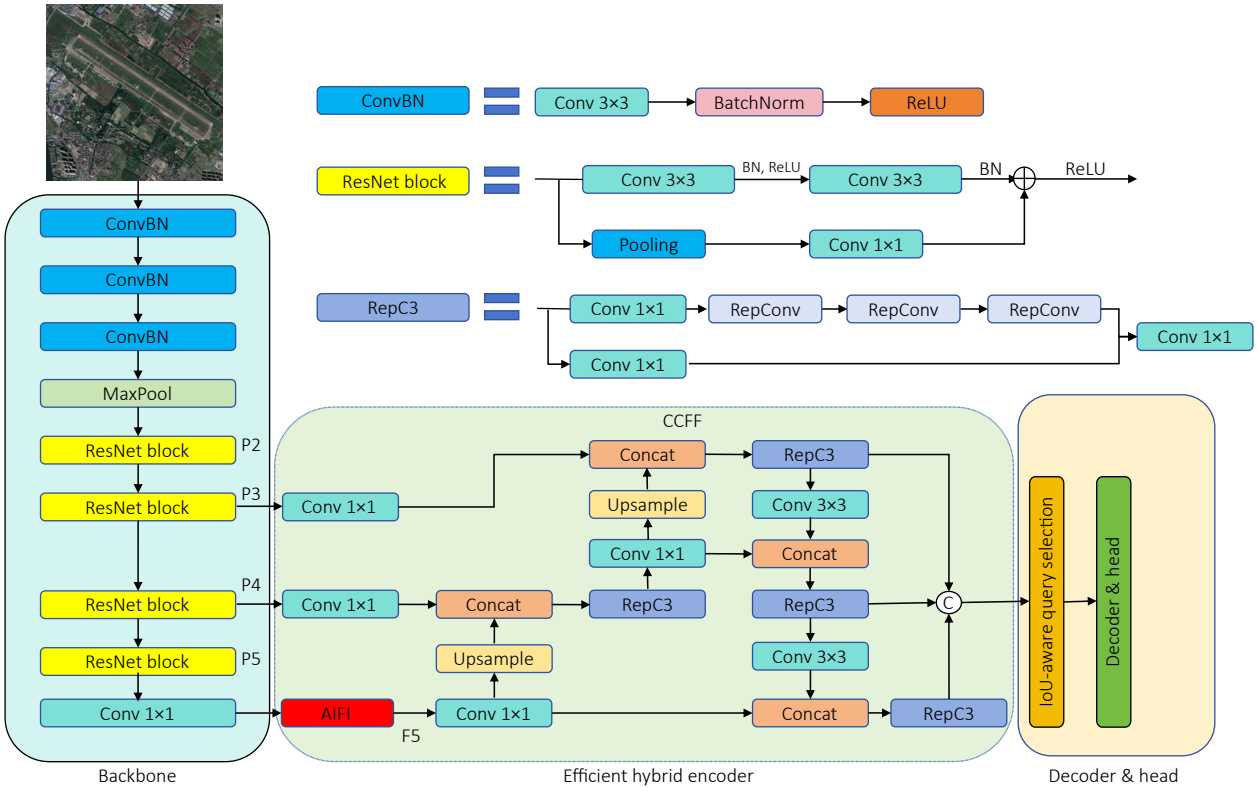


图 1 RTDETR 算法结构图

Fig. 1 Structure diagram of the RTDETR algorithm

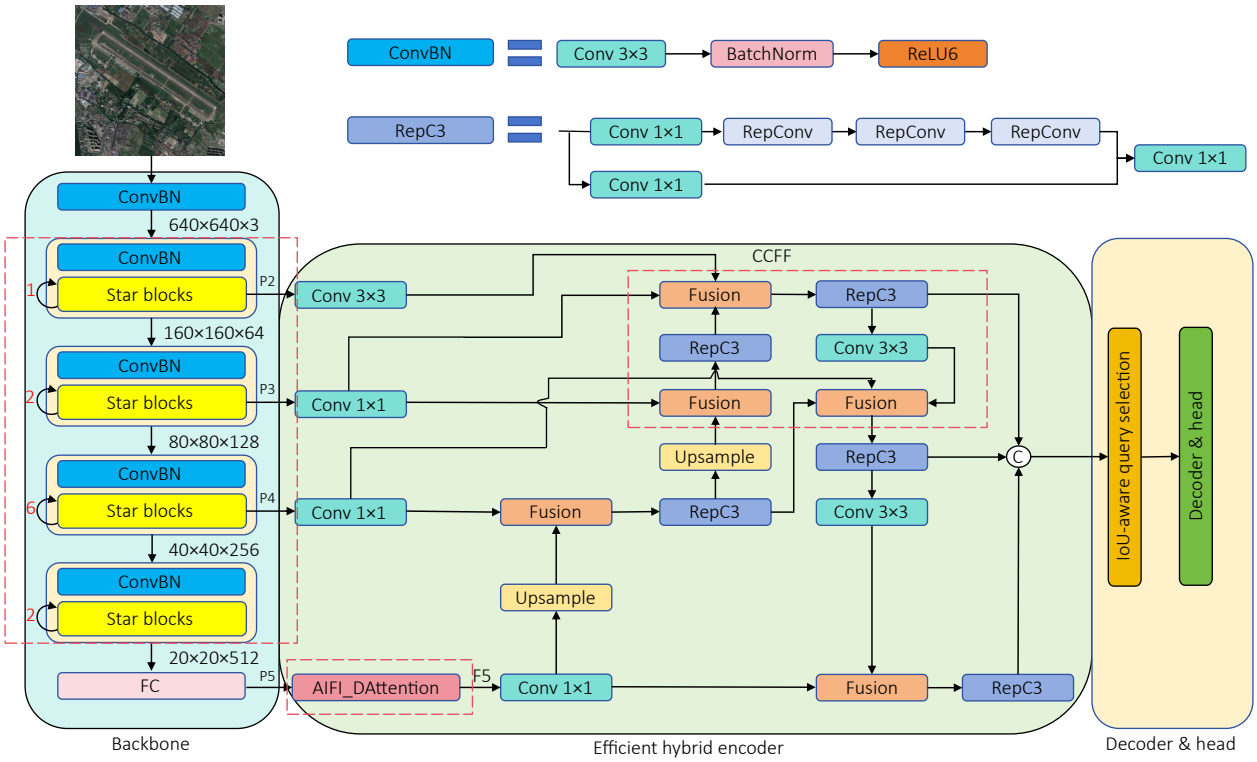


图 2 改进 RTDETR 算法结构图

Fig. 2 Structure diagram of the improved RTDETR algorithm

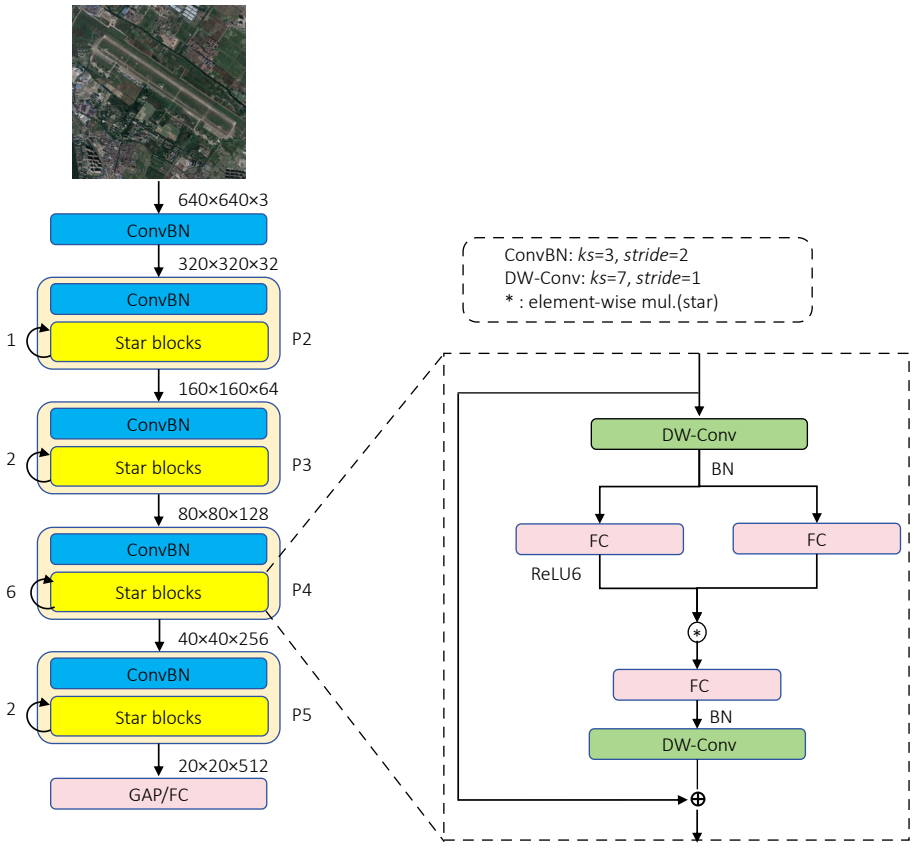


图 3 改进的主干网络结构

Fig. 3 Structure diagram of the improved backbone network

同时受益于隐含的高维特征表示，在不增加计算成本的情况下具备丰富的特征表示能力。这种高效的特征融合方式使得 StarNet 能够在紧凑的特征空间内实现近乎无限的维度，极大地提高了算法的表示能力和性能。在单层神经网络中元素乘法操作为$(\boldsymbol{w}_1^{\mathrm{T}}\boldsymbol{x}+\boldsymbol{b}_1)\times(\boldsymbol{w}_2^{\mathrm{T}}\boldsymbol{x}+\boldsymbol{b}_2)$，$\boldsymbol{w}_1$、$\boldsymbol{w}_2$、$\boldsymbol{x}\in\mathbf{R}^{(d+1)\times 1}$，其中 $\boldsymbol{w}$ 为权重矩阵，$\boldsymbol{x}$ 为输入矢量，$\boldsymbol{b}$ 为偏移量矩阵，d 为输通道数。元素乘法可重写为

$$\begin{aligned}\boldsymbol{w}_1^{\mathrm{T}}\boldsymbol{x}\times\boldsymbol{w}_2^{\mathrm{T}}\boldsymbol{x}&=\left(\sum_{i=1}^{d+1}\boldsymbol{w}_1^i\boldsymbol{x}^i\right)\times\left(\sum_{j=1}^{d+1}\boldsymbol{w}_2^j\boldsymbol{x}^j\right)=\sum_{i=1}^{d+1}\sum_{j=1}^{d+1}\boldsymbol{w}_1^i\boldsymbol{w}_2^j\boldsymbol{x}^i\boldsymbol{x}^j\\&=\alpha_{(1,1)}\boldsymbol{x}^1\boldsymbol{x}^1+\cdots+\alpha_{(4,5)}\boldsymbol{x}^4\boldsymbol{x}^5+\cdots\\&\quad+\alpha_{(d+1,d+1)}\boldsymbol{x}^{d+1}\boldsymbol{x}^{d+1},\end{aligned}\tag{1}$$

$$\begin{cases}i==j, & \alpha_{(i,j)}=\boldsymbol{w}_1^i\boldsymbol{w}_2^j\\ i!=j, & \alpha_{(i,j)}=\boldsymbol{w}_1^i\boldsymbol{w}_2^j+\boldsymbol{w}_1^i\boldsymbol{w}_2^j\end{cases},\tag{2}$$

式中：i 和 j 是通道的索引数。

在多层神经网络中元素乘法操作如式 (3)~(6) 所示。

$$O_1=\sum_{i=1}^{d+1}\sum_{j=1}^{d+1}\boldsymbol{w}_{(1,1)}^i\boldsymbol{w}_{(1,2)}^j\boldsymbol{x}^i\boldsymbol{x}^j,\tag{3}$$

$$O_2=\boldsymbol{w}_{(2,1)}^{\mathrm{T}}O_1\times\boldsymbol{w}_{(2,2)}^{\mathrm{T}}O_1,\tag{4}$$

$$O_3=\boldsymbol{w}_{(3,1)}^{\mathrm{T}}O_2\times\boldsymbol{w}_{(3,2)}^{\mathrm{T}}O_2,\tag{5}$$

$$O_l=\boldsymbol{w}_{(l,1)}^{\mathrm{T}}O_{l-1}\times\boldsymbol{w}_{(l,2)}^{\mathrm{T}}O_{l-1},\tag{6}$$

式中：O_l代表第 l 层元素乘法操作。

StarNet 主要由 ConvBN、Star blocks 和 FC 组成，它采用 4 级分层架构，利用卷积层进行下采样，并使用修改后的 Star blocks 进行特征提取，通道扩展因子始终设置为 4，网络宽度在每个阶段加倍。ConvBN 由卷积、激活函数、批归一化处理组成。Star blocks 由含有控制中间全连接层扩展比例的卷积层和深度可分离卷积组成，P2 层至 P5 层 Star blocks 的个数分别为 1、2、6、2。Star blocks 首先将输入图像先经过深度可分离卷积再进行批归一化处理，输出被划分成两路分支分别进行卷积操作且对其中一路输出分支进行非线性激活；然后进行元素级乘法，紧接着经过深度可分离卷积；最后把输出结果和原始输入进行残差连接。

3.2 可变形注意力机制

RTDETR 算法的 AIFI 模块中的标准自注意力机制使用标准的固定位置编码来并注入位置信息到特征中。这种全局的注意力机制在处理高分辨率图像或长序列数据时，计算复杂度过高，难以高效处理且对不同尺度目标理解能力太弱。因此，本文使用可变形注意力机制来改进 AIFI 模块。

可变形注意力机制结合了稀疏注意力的思想和动态偏移的创新，通过生成动态偏移来调整采样位置，从而在稀疏注意力的基础上实现更加灵活的注意力分配。可变形注意力机制首先通过稀疏选择减少计算范围，然后在这些关键位置上应用动态偏移，以获取更加准确的注意力权重。每个查询点根据自身的内容动态生成偏移量，这个偏移量决定了算法应该注意的特征点的位置。偏移量的生成可以通过一个轻量级的卷积网络来实现，输入是当前查询点的特征向量。这种动态调整的能力使得算法能够更灵活地捕捉到输入中的重要信息。动态偏移使得可变形注意力能够处理更复杂的空间结构信息，改善了标准自注意力机制的局限性。此外，可变形注意力机制通过在多个尺度上操作，使得算法可以同时关注到图像的局部细节和全局结构。

对于给定的特征图$x\in\mathbf{R}^{H_{\mathrm{G}}\times W_{\mathrm{G}}\times 2}$，生成均匀网络点$p\in\mathbf{R}^{H_{\mathrm{G}}\times W_{\mathrm{G}}\times 2}$，网格的大小是在特征图按因子 r 下采样得到的，长宽大小如式 (7) 所示，参考点的值是线性间隔的二维坐标如式 (8) 所示。

$$\begin{cases}H_{\mathrm{G}}=\dfrac{H}{r}\\ W_{\mathrm{G}}=\dfrac{W}{r}\end{cases},\tag{7}$$

$$\{(p_x,p_y)|(0,0),\cdots,(W_{\mathrm{G}}-1,H_{\mathrm{G}})\}.\tag{8}$$

式中：H 和 W 代表输入特征图的高度和宽度；x 和 y 表示坐标轴索引。

偏移量的计算首先需通过线性投影W_q来处理特征图，从而获得 q。q、k 和 v 分别表示 Query、Key 和 Value，m 表示多头注意力机制中的头索引。然后，利用偏移生成网络$\theta_{\mathrm{offset}(q)}$来生成 q 的偏移量。将偏移量加到输入图像中参考点的位置上来确定变形点实际的采样位置。此外，使用采样函数$\varphi(z;(p_x,p_y))$确保采样过程可微，从而确保采样位置的区域特征。在每个变形点上通过双线性插值方法生成特征向量，接着分别使用投影W_k和W_v得到变形后的和。为实现可变形多头注意力机制，计算每个查询$q^{(m)}$变形后的键并加上先前的位置偏移，然后计算与键的内积得到注意力分数，进而建立查询特征与图像关键区域之间的关系。在可变形注意图的引导下，将变形后的特征值融合，并通过投影W_{O}得到最终输出变

形后的特征 z。计算公式为

$$g(a,b)=max(0,1-|a-b|)\,, \tag{9}$$

$$\varphi=\sum_{(r_x,r_y)}g(p_x,r_y)g(p_y,r_y)z[r_y,r_x,:]\,, \tag{10}$$

$$\begin{cases} q=xW_q \\ \tilde{k}=\tilde{x}W_k \\ \tilde{v}=\tilde{x}W_v \end{cases}, \tag{11}$$

$$z^{(m)}=\sigma\left(\frac{q^{(m)}\tilde{k}^{(m)\mathrm{T}}}{\sqrt{d}}\right)\tilde{v}^{(m)}\,, \tag{12}$$

$$z=Concat(z^{(1)},...,z^{(m)})W_{\mathrm{O}}\,, \tag{13}$$

式中：$\tilde{x}=\varphi(x;p+\Delta p)$，$\Delta p=\theta_{\mathrm{offset}}(q)$，$g(a,b)$表示双线性采样权重函数；$(r_x,r_y)$表示$z\in\mathbf{R}^{H_{\mathrm{G}}\times W_{\mathrm{G}}\times 2}$上的所有位置；$\tilde{k}$和$\tilde{v}$分别表示变形的键和值嵌入；$\varphi(\hat{B};R)\in\mathbf{R}^{HW\times H_{\mathrm{G}}W_{\mathrm{G}}}$表示与先前对应的位置嵌入。

偏移量由偏移生成网络生成，该网络具备局部特征感知能力，能够学习出合理的偏移。该网络首先将输入特征图用卷积核大小为 5×5 的深度可分离卷积来提取特征，接着利用 GELU 激活函数、批归一化处理和 1×1 卷积进一步处理，并且忽略 1×1 卷积中的偏差项。相对位置偏差为每对查询和键之间的相对位置进行编码，通过空间信息增强了普通注意力，考虑形状为 $H\times W$ 的特征图，其相对坐标位移在 $[-H;H]$ 和 $[-W;W]$ 两个维度上，偏移量生成网络如图 4 所示。

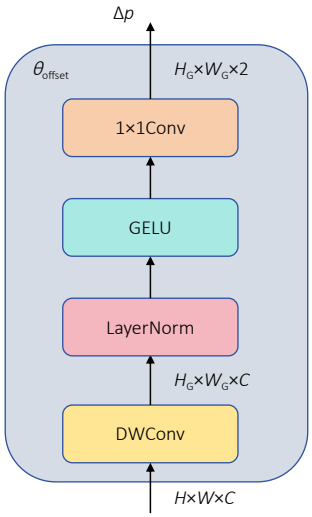


图 4 偏移生成网络

Fig. 4 Offset generation network

遥感图像不仅存在背景复杂、小目标聚集和多尺度目标众多等问题，还存在遮挡和几何畸变等场景挑战。可变形注意力机制在应对这些挑战时能够很好地发挥其自身优势。在处理遮挡问题时，遥感图像中的云层阴影和建筑重叠等会导致目标特征信息的丢失，可变形注意力机制通过其灵活的注意力采样机制，能够有效避开遮挡区域，重点关注可见的目标区域，同时结合上下文信息进行推理，能够有效提高对遮挡物体的检测与识别能力。在应对多尺度目标与密集小目标时，可变形注意力机制使得编码器模块能够更好地结合深层特征图的高维信息和浅层特征图的低维信息，且能更能够灵活地捕捉到输入中的重要信息，提高对不同尺度目标的理解能力，在目标密集区域也能够更好地区分相邻目标。在应对复杂背景和几何畸变挑战时，可变形注意力通过灵活的空间采样和自适应特征聚合，能够聚焦于更重要的目标区域并且更加准确地提取不规则目标，有效抑制复杂背景和几何畸变带来的干扰，提升目标检测任务的鲁棒性。

3.3 加权双向特征金字塔网络

在原 RTDETR 算法中 CCFF 模块采用简单的拼接或者加和来进行特征融合，这种方式并未能有效地利用金字塔特征图之间的相关性，特别是在处理遥感图像小目标时，不能够充分利用多尺度信息且会造成小目标信息丢失。为了解决上述问题，本文引入加权双向特征金字塔网络 BiFPN 对 CCFF 进行改进。BiFPN 结构可以分为自顶向下的特征传播和自低向上的特征传播两个阶段，丰富了特征的上下文表示，能在一定程度上提高遥感图像目标检测的效果。在加权双向特征金字塔网络结构中，首先删除没有与其他部分进行特征融合的单一节点来减少计算量；其次在相同层级的输入输出节点上增加一条边来更好地进行特征融合更深层次的特征融合。此外，还引入了可学习的权重，最后把每条双向路径看成一个网络特征层来实现通过自动调整不同输入特征的权重更好的实现特征融合。

在 RTDETR 算法的 CCFF 模块中引入 BiFPN，BiFPN 结构如图 5 所示。通过创新性地增加 P2 层，减少小目标在特征提取过程中的信息丢失，并且增强算法对不同尺度目标的特征融合能力。在图 5 中，P2、P3、P4、P5 层分别是对输入图像进行 4、8、16、32 倍下采样获得的。在 BiFPN 结构中，P2 层具有较高的分辨率，该层中存在小目标丰富的细粒度特征信息。此外，通过将有大量低维特征信息的 P2 层与有更高维特征的 P3 层进行融合，能够增强算法对小目标的识别能力和特征信息的表达力。由于 P3 层和 P4 层的输入与输出节点处于相同层级，因此需要在这两个节点之间增加额外的边。通

过 BiFPN 模块进行改进，增强了算法对小目标的识别能力，从而提高算法的检测精度。

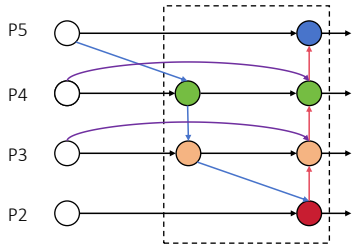


图 5 BiFPN 模块

Fig. 5 BiFPN module

4 实验与结果分析

4.1 实验设备与参数

本文实验使用 Ubuntu 22.04 LTS 操作系统，16 核的 Intel Xeon Platinum 8481C 处理器，显存大小为 24 GB 的 NVIDIA GTX 4090D 显卡，CUDA 版本为 12.1。深度学习框架使用的版本为 Pytorch 1.13.1、Torchvision 0.14.1，Python 版本为 3.8.16。本文的所有实验在训练阶段均未载入预训练权重。具体的实验环境参数设置如表 1 所示。

表 1 实验参数

Table 1 Experimental parameter

Type	Setting
Epochs	200
Batch size	4
Image size	640×640
Workers	4
Learning rate	0.0001
Optimize	AdamW

4.2 数据集

为了评估和验证算法性能，本文使用由西北工业大学提出的 DIOR 数据集进行消融实验和对比实验，该数据集中一共有 23463 张遥感图像和 190288 个目标实例，适合用于训练和验证复杂的检测算法。数据集图像是在不同天气、季节和成像条件下拍摄，具有丰富的变化。图像大小为 800 pixel×800 pixel，空间分辨率为 0.5 m 到 30.0 m，均为高分辨率光学遥感图像，适合用于精确的目标检测。每个目标实例都用轴向对齐的边界框手动标记，标注信息详细且准确。数据集中包含高速公路服务区 (expressway-service-area)、机场 (airport)、桥 (bridg)、棒球场 (baseballfield)、港口 (harbor)、烟囱 (chimney)、水坝 (dam)、飞机 (airplane)、高速公路收费站 (expressway-toll-station)、篮球场 (basketballcourt)、高尔夫球场 (golffield)、地面轨道场 (groundtrackfield)、风车 (windmill)、船 (ship)、网球场 (tenniscourt)、储罐 (storagetank)、体育场 (stadium)、火车站 (trainstation)、车辆 (vehicle) 和天桥 (overpass) 等 20 个目标类别。本文将 DIOR 数据集按照 1:1:2 的比例随机划分为训练集、验证集和测试集，其中训练集有 5863 张图像、验证集有 5862 张图像、测试集有 11738 张图像。

考虑到遥感图像目标检测的实际应用，为了防止算法出现过拟合，提高算法的泛化性能，本文在 RSOD 数据集上进行泛化分析实验 RSOD 数据集包含飞机、运动场、立交桥和油桶四个对象类别，包含 976 张图像和 6950 个目标。为了在 RSOD 数据集上学习到比较多的特征，将 RSOD 数据集按照 7:1:2 进行分配，即训练集 657 张图像、验证集 93 张图像和测试集 186 张图像。

4.3 评价指标

本文实验均在同一环境下进行训练和测试，为了评估本文改进算法应用于遥感图像目标检测任务中的效果，使用精确率 (precision, P)、召回率 (recall, R)、平均精度均值 (mean average precision, mAP)、参数量 (parameters)、每秒千兆浮点运算次数 (gigabyte floating-point operations per second, GFLOPs) 和每秒检测帧数 (frames per second, FPS) 等指标进行评估。其中 mAP、AP、R 和 P 的公式为

$$P=\frac{TP}{FP+TP}, \tag{14}$$

$$R=\frac{TP}{FN+TP}, \tag{15}$$

$$AP=\int_0^1 P(R)\mathrm{d}R, \tag{16}$$

$$mAP=\frac{1}{n}\sum_{i=0}^{n} AP_i, \tag{17}$$

式中：TP 表示预测为正样本且与实际标签相符的样本数量；FP 表示预测为正样本但与实际标签不符的样本数量；FN 表示预测为负样本且与真实标签不符的样本数量；AP 表示在 Precision-Recall 曲线下所计算出的面积值；n 表示数据集中目标类别的总数。

4.4 消融实验

本文在 RTDETR 算法的基础上进行了一系列改进，旨在评估改进算法应用于遥感图像目标检测任务中的效果。通过设计一组消融实验，分别验证了三个改进模块所带来的性能影响，消融实验结果如表 2 所示，使用“√”表示该组实验在 RTDETR 算法中引入的部分，数据加粗代表本文模型或本文使用的改进点，在实验中逐步将各个改进模块嵌入到算法中。RTDETR 算法和改进算法的精确率、召回率、mAP50 和 mAP50-95 训练曲线如图 6 所示。由表 2 数据可知，单独使用轻量化主干网络

表 2 消融实验结果

Table 2 Ablation experiment results

RTDETR	StarNet	DAttention	BiFPN	*P*/%	*R*/%	mAP50/%	mAP50-95/%	Parameter/M	GFLOPs	FPS
√				86.3	75.3	80.3	57.4	19.9	57.0	184.2
√	√			87.1	76.7	80.7	57.8	12.0	31.9	253.3
√		√		86.6	75.9	80.7	57.6	19.9	57.2	188.6
√			√	86.8	75.8	80.7	57.8	20.3	64.4	164.1
√	√	√		87.7	77.8	81.9	58.9	12.0	32.1	318.1
√	√		√	87.6	77.8	81.8	58.7	12.4	38.3	233.0
√		√	√	86.7	76.3	81.0	57.8	20.3	64.6	171.7
√	**√**	**√**	**√**	**88.1**	**78.3**	**82.1**	**59.1**	**12.4**	**38.5**	**257.8**

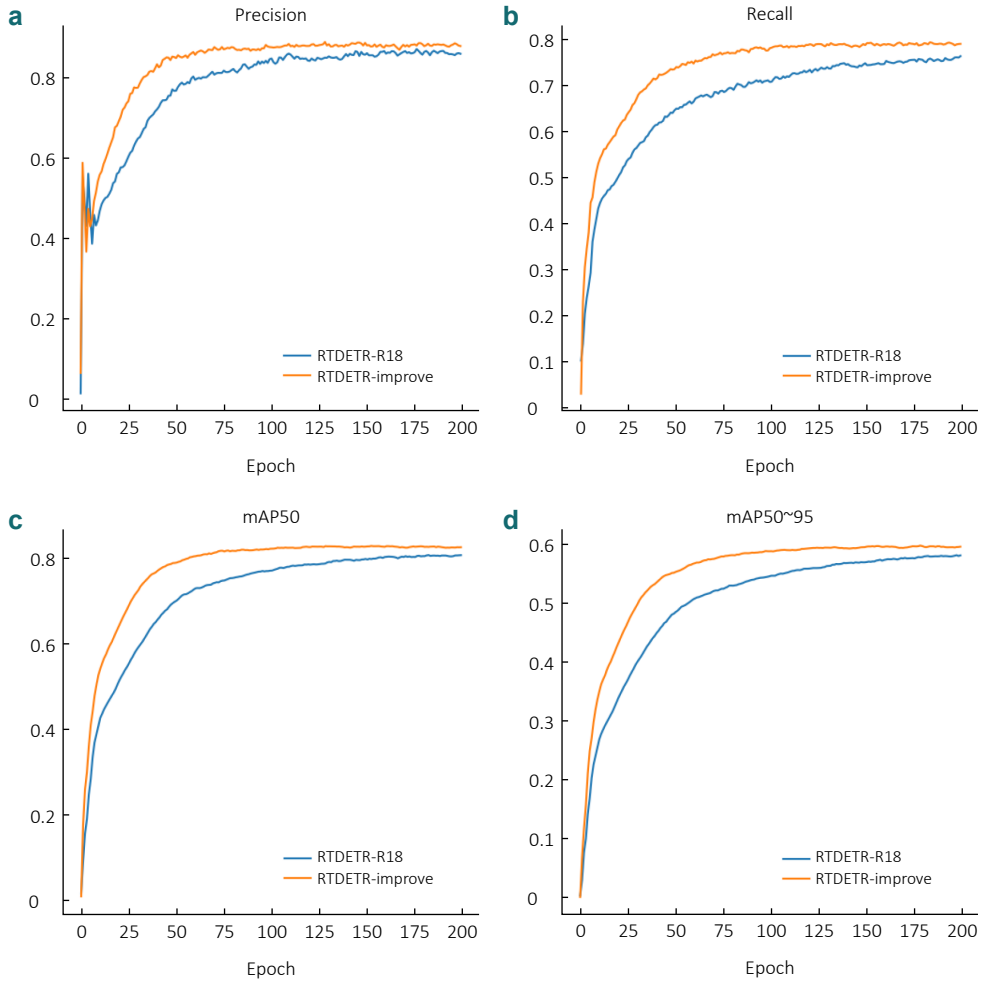


图 6 各指标变化曲线对比图。(a) 精确率曲线对比；(b) 召回率曲线对比；(c) mAP50 曲线对比；(d) mAP50-95 曲线对比

Fig. 6 Comparison chart of each index change curves. (a) Comparison of precision curves; (b) Comparison of recall curves; (c) Comparison of mAP50 curves; (d) Comparison of mAP50-95 curves

StarNet 替换原算法 ResNet18 主干网络时，mAP50 和 mAP50-95 均提升了 0.4%，参数量降低了 39.7%。相较于原算法，该改进方法在大幅降低计算量和参数量的同时提高了检测精度，这表明 StarNet 主干网络能够有效解决算法主干网络在特征提取过程中存在的信息丢失问题。单独使用 DAttention 模块替换 AIFI 中标准自注意力机制时，mAP50 和 mAP50-95 分别提升了 0.4% 和 0.2%，参数量保持不变，计算量提高了 0.2 GFLOPs。这说明 DAttention 模块能够增强编码器在融合深层高维特征与浅层低维特征方面的能力，从而优化对不同尺度目标的感知与表达能力。当同时使用以上两个模块改进算法时，相比于原算法，mAP50 和 mAP50-95 分别提升了 1.6% 和 1.5%，参数量下降了 39.7%，计算量降低了 43.7%。单独使用 BiFPN 模块来改进跨尺度特征融合部分时，mAP50 和 mAP50-95 均提升 0.4%，由于该模块增加了 P2 层，因此参数量和计算量有所增加，但并不会影响算法推理时间。这说明 BiFPN 模块可以有效解决算法在特征融合过程中特征信息的丢失问题，提高算法的特征融合能力。所提算法将 StarNet 主干网络、DAttention 和 BiFPN 三个模块进行融合之后，相较于 RTDETR 算法，其 mAP50 和 mAP50-95 分别提升了 1.8% 和 1.7%，同时参数量和计算量显著降低，进一步验证了所提算法的优越性。

不同 RTDETR 改进方法对比如表 3 所示。由表 3 数据可知，使用 FasterNet[22]、ConvNextV2[23] 等轻量化主干来改进 RTDETR，在参数量和计算量上与 StarNet 接近，但是检测精度不如 StarNet。使用 RMT (retentive networks meet vision Transformer)[24] 改进主干网络，其算法精度、参数量和计算量均不如 StarNet。使用 HiLo[25] 和 EAA (efficient additive attention)[26] 等注意力机制来改进 AIFI，其参数量和计算量与 DAttention 大体相同，但是算法精度小于 DAttention。使用 ASF (attentional scale sequence fusion)[27]、goldyolo[28]、SDI(semantics and detail infusion module)[29] 等来改进 CCFF，在检测精度方面 BiFPN 最高，参数量接近于其他改进，计算量略大于其他改进。总体来说，本文使用的改进方法会优于其他改进方法。

4.5 对比实验

为了进一步评估本文改进算法应用于遥感图像目标检测任务中的优势，将本文算法与其他主流算法在 DIOR 数据集上的检测结果进行对比，对比主流算法包括 Faster -RCNN[30]、Cascade-RCNN[31]、RetinsNet[32]、RTMDet (real-time models for object detection)[33]、DQ-DETR (dynamic query DETR)[34]、GFL (generalized focal loss)[35]、TOOD (task-aligned one-stage object detection)[36]、YOLOv5m、YOLOv8m 及 YOLOv10m。根据表 4 的实验结果所示，本文算法平均精度达到最高，其中 mAP50 和 mAP50-95 分别为 82.1% 和 59.1%，相较于其他算法有明显的优势。这主要是由于可变形注意力机制灵活的注意力分配和加权双向特征金字塔网络的双向信息流动。变形注意力机制能够在不同尺度上自适应地选择关键信息，增强算法对局部细节和多尺度特征的理解。加权双向特征金字塔网络通过引入双向信息流动，可以更充分地利用低层的细节信息与高层的语义信息，优化多尺度特征的表达能力。在轻量化方面，改进算法的参数量和

表 3 不同 RTDETR 改进方法对比

Table 3 Comparison of improvement methods for different RTDETR

Model	mAP50/%	mAP50-95/%	Parameter/M	GFLOPs
Backbone -FasterNet	78.6	55.2	10.8	28.6
Backbone -ConvNextV2	77.5	54.5	12.3	31.9
Backbone-RMT	79.9	56.9	21.4	61.2
Backbone-StarNet	**80.7**	**57.8**	**12.0**	**31.9**
AIFI-HiLo	80.1	57.0	19.9	57.2
EAA	80.2	57.1	19.9	57.2
AIFI-DAttention	**80.7**	**57.6**	**19.9**	**57.2**
Neck-ASF	80.4	57.2	20.2	61.5
Neck-goldyolo	80.1	56.8	22.3	60.0
Neck-SDI	79.7	56.8	19.9	57.2
Neck-BiFPN	**80.7**	**57.8**	**20.3**	**64.4**

表 4 对比实验结果

Table 4 Contrast experiment results

Model	mAP50/%	mAP50-95/%	Parameter/M	GFLOPs
Faster-RCNN	76.7	51.7	41.4	134.0
Cascade RCNN	77.0	53.4	69.4	162.0
RetinsNet	74.9	51.4	36.7	133.0
RTMDet	79.8	56.9	4.9	8.0
DQ-DETR	81.7	55.0	-	-
GFL	76.0	52.8	32.3	128.0
TOOD	79.1	56.0	32.1	124.0
YOLOv5m	80.0	56.5	25.1	64.0
YOLOv8m	80.4	57.7	25.9	78.7
YOLOv10m	79.6	56.5	15.3	58.9
YOLOX-tiny	74.9	46.7	8.9	13.3
RTDETR	80.3	57.4	19.9	57.0
RTDETR-improve	**82.1**	**59.1**	**12.4**	**38.5**

计算量分别为 12.4 M 和 38.5 G，算法复杂度小于其他对比算法。这主要得益于 StarNet 主干网络的星操作，其通过元素乘法将输入数据映射到一个极高维的非线性特征空间，这种方式能够大大减少计算量和参数量且生成丰富的特征表示。

4.6 泛化分析实验

为进一步验证改进算法的泛化能力，本文在 RSOD 数据集上进行泛化分析实验。表 5 展示了改进算法和 RTDETR 算法在 RSOD 数据集上的检测结果，结果表明，改进算法在 mAP50 和 mAP50-95 指标均高于基准算法，参数量和计算量也均远低于基准算法。在 RSOD 数据集上的泛化实验结果进一步展示了本文改进算法应用于遥感图像目标检测中的泛化性和鲁棒性。

4.7 可视化分析

为了验证本文改进算法应用于遥感图像目标检测任务中的优势，从 DIOR 测试集上挑选了 8 张图像进行检测，RTDETR 算法和改进算法检测效果如图 7 所示。在图 7(a) 中，改进算法检测到的火车站目标具有更高的置信度，且没有出现误检现象。在图 7(b) 中，改进算法检测到的地面轨道场目标置信度达到 96%，而 RTDETR 算法检测置信度只有 31%。在图 7(c) 中，改进算法检测到的大部分船小目标具有更高的置信度。在图 7(d) 中，改进算法检测到大部分飞机小目标具有更高的置信度且检测出更多的目标。在图 7(e) 中，改进算法检测到的目标具有更高的置信度且没有出现误检现象，而 RTDETR 算法存在大量漏检的棒球场目标。在图 7(f) 中，改进算法检测到更多的风车目标，而 RTDETR 算法存在漏检现象。在图 7(g) 中，RTDETR 算法将左下角的桥误检成了船，而改进算法则没有且船和港口等目标普遍具有更高的置信度。在图 7(h) 中，RTDETR 算法将一个篮球场目标识别为了两个，而改进算法则能够正确识别出一个篮球场目标。与 RT-DETR 相比，本文改进算法在遥感图像目标检测中具有更低的漏检率和误检率，在应对复杂背景、密集目标、不同尺度目标的情况下都有比较高的检测精度。

表 5 RSOD 检测结果

Table 5 RSOD dataset detection result

Model	mAP50/%	mAP50-95/%	Parameter/M	GFLOPs
RTDETR	94.2	66.1	19.9	57.0
RTDETR-improve	**95.8**	**66.6**	**12.3**	**38.4**

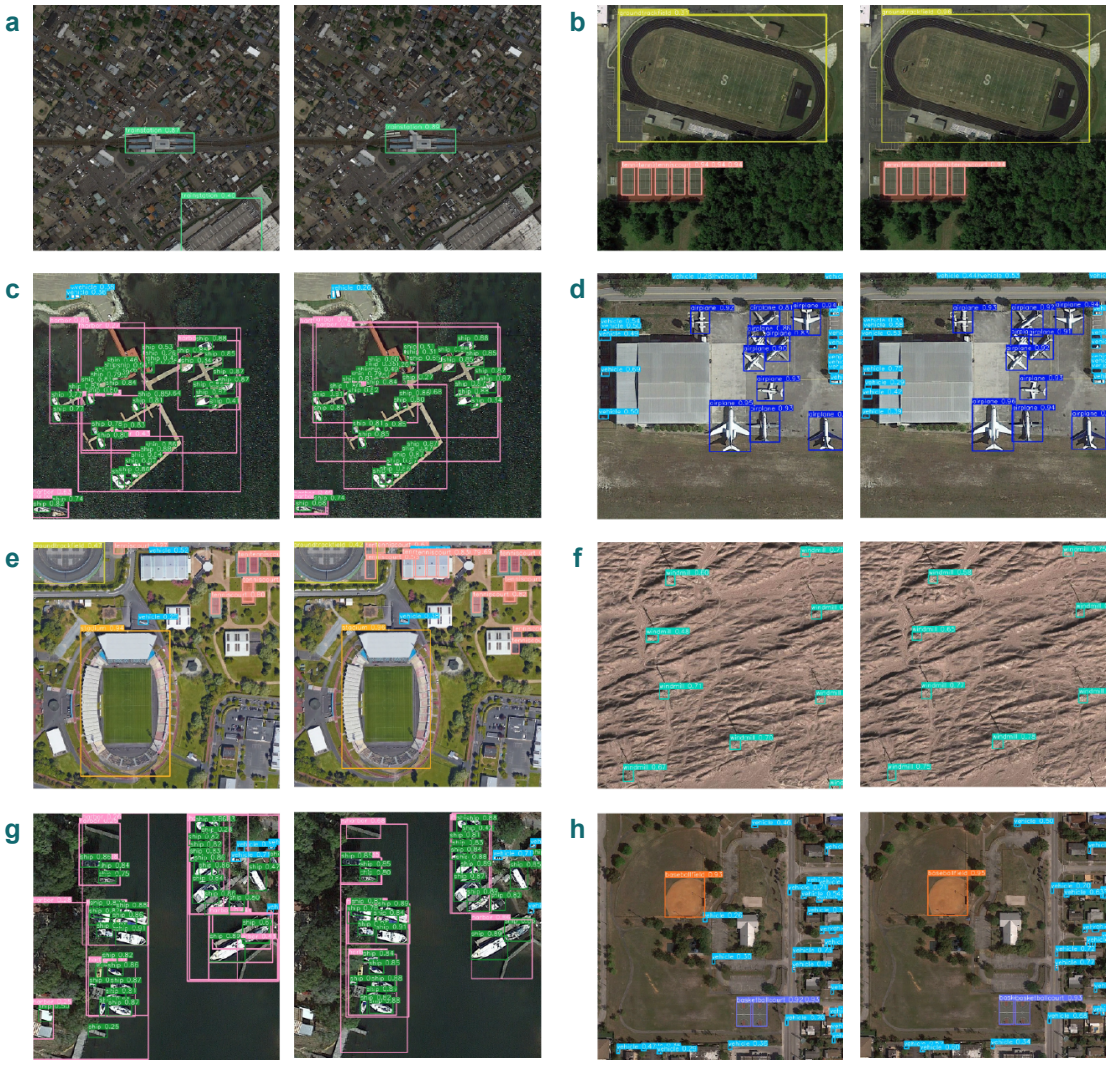


图 7 RTDETR 算法与本文算法可视化结果对比。(a-h) 从 DIOR 测试集中选取的 8 张图像检测效果

Fig. 7 Comparison of visualization results between RTDETR algorithm and improved algorithm. (a-h) The detection results of 8 images selected from the DIOR dataset

5 结 论

遥感图像存在背景复杂、尺度多样性以及小目标众多等问题，当前遥感图像目标检测算法不能够很好地解决这些问题，因此本文提出了一种复杂背景下的改进 RTDETR 遥感图像目标检测算法。本文改进算法相较于 RT-DETR 算法，在 DIOR 遥感数据集上的检测精度大幅度提升，在一定程度上减少了遥感图像目标检测任务中容易出现的误检、漏检情况，同时较好地实现了算法轻量化，使得算法能够满足遥感图像实时检测的需求。此外，与当前主流算法相比，本文算法在检测精度上表现更优且具有更小的算法复杂度。在 RSOD 数据集上的检测结果更加充分地展示了所提算法泛化性和鲁棒性。本文提出的改进算法在目标检测精度上显著提升，同时大幅度降低了算法复杂度，优化了算法的实时性，充分验证了该算法在遥感图像目标检测中的应用潜力和实际价值。未来将通过优化损失函数和进一步改进特征融合部分来减少算法复杂度，探索模型在资源受限环境下的轻量化与推理加速策略，从而在保证检测精度的前提下提升算法的实时性。

作者贡献

沈汉滨：研究资料的收集、分析，论文写作；梅婷、林珊玲、

林志贤：论文选题指导、技术指导、论文修改；林坚普、吕珊红、郭太良：理论指导、论文修改。

利益冲突

所有作者声明无利益冲突

参考文献

1. 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.
2. Redmon J, Farhadi A. YOLOv3: an incremental improvement[Z]. arXiv: 1804.02767, 2018.
https://doi.org/10.48550/arXiv.1804.02767.
3. Wang C Y, Bochkovskiy A, Liao H Y M. YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[C]//*Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022: 7464–7475.
https://doi.org/10.1109/CVPR52729.2023.00721.
4. Wang A, Chen H, Liu L H, et al. YOLOv10: real-time end-to-end object detection[C]//*Proceedings of the 38th Conference on Neural Information Processing Systems*, 2024.
5. Liu W, Anguelov D, Erhan D, et al. SSD: single shot MultiBox detector[C]//*Proceedings of the 14th European Conference on Computer Vision*, 2016: 21–37.
https://doi.org/10.1007/978-3-319-46448-0_2.
6. Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//*Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014: 580–587.
https://doi.org/10.1109/CVPR.2014.81.
7. 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.
8. Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers[C]//*Proceedings of the 16th European Conference on Computer Vision*, 2020: 213–229.
https://doi.org/10.1007/978-3-030-58452-8_13.
9. Zhu X Z, Su W J, Lu L W, et al. Deformable DETR: deformable transformers for end-to-end object detection[C]//*Proceedings of the 9th International Conference on Learning Representations*, 2021.
10. Meng D P, Chen X K, Fan Z J, et al. Conditional DETR for fast training convergence[C]//*Proceedings of 2021 IEEE/CVF International Conference on Computer Vision*, 2021: 3631–3640.
https://doi.org/10.1109/ICCV48922.2021.00363.
11. Zhang H, Li F, Liu S L, et al. DINO: DETR with improved DeNoising anchor boxes for end-to-end object detection[C]//*Proceedings of the 11th International Conference on Learning Representations*, 2023.
12. Zhao Y A, Lv W Y, Xu S L, et al. DETRs beat YOLOs on real-time object detection[C]//*Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024: 16965–16974.
https://doi.org/10.1109/CVPR52733.2024.01605.
13. Jin R J, Wang K, Liu M H, et al. DETR with improved DeNoising training for multi-scale oriented object detection in optical remote sensing images (invited)[J]. *Laser & Optoelectronics Progress*, 2024, **61**(2): 0211023.
金睿蛟, 王堃, 刘敏豪, 等. 基于 DETR 和改进去噪训练的光学遥感图像多尺度旋转目标检测 (特邀)[J]. 激光与光电子学进展, 2024, **61**(2): 0211023.
14. Xiao Z J, Zhang J H, Lin B H. Feature coordination and fine-grained perception of small targets in remote sensing images[J]. *Opto-Electron Eng*, 2024, **51**(6): 240066.
肖振久, 张杰浩, 林渤翰. 特征协同与细粒度感知的遥感图像小目标检测[J]. 光电工程, 2024, **51**(6): 240066.
15. Bai J Y, Jiang T, Wei Y M, et al. Remote sensing image detection algorithm based on improved RT-DETR[J]. *Radio Eng*, 2025, **55**(2): 334−342.
白金燕, 江涛, 魏玉梅, 等. 基于改进 RT-DETR 的遥感图像检测算法[J]. 无线电工程, 2025, **55**(2): 334−342.
16. Zhou H L, Ye Q, Liu W Q. Lightweight remote sensing military aircraft target detection in complex backgrounds[J]. *Opto-Electron Eng*, 2025, **52**(2): 240270.
周韩莲, 叶青, 刘文祺. 复杂背景下的轻量级遥感军用飞机目标检测[J]. 光电工程, 2025, **52**(2): 240270.
17. Ma X, Dai X Y, Bai Y, et al. Rewrite the stars[C]//*Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024: 5694–5703.
https://doi.org/10.1109/CVPR52733.2024.00544.
18. Xia Z F, Pan X R, Song S J, et al. Vision transformer with deformable attention[C]//*Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022: 4784–4793.
https://doi.org/10.1109/CVPR52688.2022.00475.
19. Tan M X, Pang R M, Le Q V. EfficientDet: scalable and efficient object detection[C]//*Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 10778–10787.
https://doi.org/10.1109/CVPR42600.2020.01079.
20. He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]//*Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 770–778.
https://doi.org/10.1109/CVPR.2016.90.
21. Liu S Y, Gao K, Yong L Q. Improved RT-DETR algorithm for aerial small object detection[J]. *Comput Eng Appl*, 2025, **61**(4): 272−281.
刘思元, 高凯, 雍龙泉. 改进 RT-DETR 的航拍小目标检测算法[J]. 计算机工程与应用, 2025, **61**(4): 272−281.
22. Chen J R, Kao S H, He H, et al. Run, don't walk: chasing higher FLOPS for faster neural networks[C]//*Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023: 12021–12031.
https://doi.org/10.1109/CVPR52729.2023.01157.
23. Woo S, Debnath S, Hu R H, et al. ConvNeXt v2: co-designing and scaling ConvNets with masked autoencoders[C]//*Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023: 16133–16142.
https://doi.org/10.1109/CVPR52729.2023.01548.
24. Fan Q H, Huang H B, Chen M R, et al. RMT: retentive networks meet vision transformers[C]//*Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024: 5641–5651.
https://doi.org/10.1109/CVPR52733.2024.00539.
25. Pan Z Z, Cai J F, Zhuang B H. Fast vision transformers with HiLo attention[C]//*Proceedings of the 36th Conference on Neural Information Processing Systems*, 2022: 14541–14554.
26. Shaker A, Maaz M, Rasheed H, et al. SwiftFormer: efficient additive attention for transformer-based real-time mobile vision applications[C]//*Proceedings of 2023 IEEE/CVF International Conference on Computer Vision*, 2023: 17379–17390.
https://doi.org/10.1109/ICCV51070.2023.01598.
27. Kang M, Ting C M, Ting F F, et al. ASF-YOLO: a novel YOLO model with attentional scale sequence fusion for cell instance segmentation[J]. *Image Vision Comput*, 2024, **147**: 105057.
28. Wang C C, He W, Nie Y, et al. Gold-YOLO: efficient object detector via gather-and-distribute mechanism[C]//*Proceedings of the 37th Conference on Neural Information Processing Systems*, 2023: 51094–51112.
29. Peng Y P, Sonka M, Chen D Z. U-Net v2: rethinking the skip connections of U-Net for medical image segmentation[Z]. arXiv:

2311.17791, 2024. https://doi.org/10.48550/arXiv.2311.17791.
30. 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.
31. Cai Z W, Vasconcelos N. Cascade R-CNN: delving into high quality object detection[C]//*Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: 6154–6162. https://doi.org/10.1109/CVPR.2018.00644.
32. Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//*Proceedings of 2017 IEEE International Conference on Computer Vision*, 2017: 2999–3007. https://doi.org/10.1109/ICCV.2017.324.
33. Lyu C Q, Zhang W W, Huang H A, et al. RTMDet: an empirical study of designing real-time object detectors[Z]. arXiv: 2212.07784, 2022. https://doi.org/10.48550/arXiv.2212.07784.
34. Huang Y X, Liu H I, Shuai H H, et al. DQ-DETR: DETR with dynamic query for tiny object detection[C]//*Proceedings of the 18th European Conference on Computer Vision*, 2024: 290–305. https://doi.org/10.1007/978-3-031-73116-7_17.
35. Li X, Wang W H, Wu L J, et al. Generalized focal loss: learning qualified and distributed bounding boxes for dense object detection[C]//*Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020: 1763.
36. Feng C J, Zhong Y J, Gao Y, et al. TOOD: task-aligned one-stage object detection[C]//*Proceedings of 2021 IEEE/CVF International Conference on Computer Vision*, 2021: 3490–3499. https://doi.org/10.1109/ICCV48922.2021.00349.

作者简介



沈汉滨 (2001-)，男，硕士研究生，主要研究方向为图像处理、深度学习目标检测。

E-mail：2098767048@qq.com



【通信作者】林珊玲 (1991-)，女，福建泉州人，博士，讲师，硕士研究生导师，福州大学先进制造学院电子信息系教师，主要研究方向为新型显示技术、平板显示器驱动系统图像处理技术等。

E-mail：sllin@fzu.edu.cn

Improved RTDETR based remote sensing image object detection algorithm for complex backgrounds

Shen Hanbin[1,2], Mei Ting[1,2], Lin Shanling[2,3*], Lin Zhixian[1,2,3], Lin Jianpu[2,3], Lv Shanhong[2,3], Guo Tailiang[1,2,3]

[1] School of Physics and Information Engineering, Fuzhou University, Fuzhou, Fujian 350116, China;
[2] Fujian Science & Technology Innovation Laboratory for Optoelectronic Information of China, Fuzhou, Fujian 350116, China;
[3] School of Advanced Manufacturing, Fuzhou University, Quanzhou, Fujian 362200, China

Abstract

Objective Remote sensing images are formed by capturing the reflection or radiation information of electromagnetic waves from ground objects, reflecting the spatial distribution and semantic features of these objects. Remote sensing image target detection holds significant application value in fields such as environmental monitoring, military defense, and land resource management. However, due to the complex background of remote sensing images, significant variations in target scales, and a large number of small targets, existing target detection algorithms still exhibit deficiencies in feature extraction, multi-scale feature fusion, and model complexity control, making it difficult to simultaneously achieve both detection accuracy and computational efficiency. To address the aforementioned issues, this paper proposes an improved RTDETR remote sensing image target detection method based on an analysis of the structure and performance characteristics of the RTDETR algorithm. The aim is to enhance the model's ability to detect multi-scale targets and reduce model complexity while maintaining detection accuracy.

Methods This article proposes an improved RTDETR remote sensing image object detection method to address the problems of insufficient feature extraction ability, limited multi-scale object modeling ability, and high model complexity in remote sensing image object detection tasks. Firstly, in terms of backbone network design, a lightweight backbone network StarNet is adopted to replace the original ResNet18. StarNet maps input features to a high-dimensional nonlinear feature space through element level multiplication, which

Foundation item: National Key Research and Development Program of China (2021YFB3600603)
*Correspondence: Lin S L, E-mail: sllin@fzu.edu.cn

can generate richer feature representations and enhance the model's ability to express complex remote sensing scenes. Secondly, to enhance the modeling ability of the model for targets of different scales, a deformable attention mechanism is introduced in the AIFI module of the efficient hybrid encoder to replace the original standard attention mechanism. The deformable attention mechanism combines the idea of sparse attention with dynamic shift strategy. By learning variable sampling positions, attention allocation becomes more flexible, allowing for adaptive attention to target regions of different spatial positions and scales. In addition, a bidirectional feature pyramid network module is introduced in the feature fusion stage of the efficient hybrid encoder. This module achieves full interaction and fusion of feature information at different levels through a bidirectional feature transfer mechanism from top to bottom and from bottom to top, effectively enhancing the model's ability to comprehensively utilize multi-scale features and further improving the detection performance of targets at different scales in remote sensing images. To verify the effectiveness of the proposed method, this paper conducted ablation experiments and comparative experiments on the DIOR remote sensing image dataset. At the same time, in order to evaluate the generalization ability of the algorithm in practical remote sensing application scenarios, further generalization analysis experiments were conducted on the RSOD dataset to reduce the risk of overfitting and verify the robustness of the model.

Results and Discussions The experimental results show that the proposed improved algorithm has achieved significant performance improvement on the DIOR dataset. Among them, the accuracy and recall of the algorithm increased by 1.8% and 3.0% respectively, and the mAP50 and mAP50-95 indicators increased by 1.8% and 1.7% respectively. While improving detection performance, the parameter and computational complexity of the model were reduced by 37.7% and 32.5%, respectively, indicating that the proposed method significantly reduces model complexity while ensuring detection accuracy, achieving an effective balance between accuracy and efficiency. To further verify the generalization ability of the algorithm, this paper conducted generalization analysis experiments on the RSOD dataset. The experimental results showed that the improved algorithm improved the mAP50 and mAP50-95 metrics by 1.6% and 0.5%, respectively, while reducing the parameter count by 37.9%. The above results indicate that the method has good robustness and generalization performance on different remote sensing datasets, and can effectively alleviate overfitting problems. In addition, the detection performance of the algorithm proposed in this article was compared with various mainstream object detection algorithms on the DIOR dataset. The experimental results showed that the algorithm proposed in this article achieved the best performance in terms of average detection accuracy, showing significant advantages compared to other comparative algorithms. The main reason for the performance improvement lies in the synergistic effect of the introduced deformable attention mechanism and the weighted bidirectional feature pyramid network. The deformable attention mechanism can adaptively focus on key information regions at different scales through flexible attention allocation strategies, thereby enhancing the modeling ability of the model for local details and multi-scale features; The weighted bidirectional feature pyramid network achieves full integration of low-level detail features and high-level semantic features through a bidirectional information flow mechanism, effectively enhancing the ability to express multi-scale features. In terms of lightweight performance, the parameter and computational complexity of the improved algorithm are 12.4 M and 38.5 G, respectively, and the overall algorithm complexity is lower than other compared algorithms. This is mainly due to the adoption of the StarNet backbone network, which maps input features to a high-dimensional nonlinear feature space through element level multiplication in star operations. While generating rich feature representations, it significantly reduces model parameters and computational overhead, thereby enhancing the practicality and deployment value of the algorithm in remote sensing image object detection tasks.

Conclusions This paper addresses the challenge of balancing accuracy and efficiency in remote sensing image object detection by improving the RTDETR algorithm and proposing a target detection method that considers both detection performance and model lightweighting. By introducing the StarNet backbone network, deformable attention mechanism, and bidirectional feature pyramid network, the model's ability to model and fuse features of complex scenes and multi-scale targets is enhanced. Experimental results show that this method achieves good detection performance on multiple remote sensing datasets and significantly reduces model complexity, validating its effectiveness and generalization ability. This research provides a reference for the deployment of remote sensing object detection algorithms in practical applications and resource-constrained platforms. In the future, we will further explore more efficient feature modeling methods to enhance the algorithm's applicability in complex remote sensing scenes.

Keywords: remote sensing image target detection; RTDETR; lightweight; deformable attention mechanisms

Shen H B, Mei T, Lin S L et al. Improved RTDETR based remote sensing image object detection algorithm for complex backgrounds[J]. *Opto-Electron Eng*, 2026, **53**(1): 250064; DOI: 10.12086/oee.2026.250064

Opto-Electronic Journals Group
www.oejournal.org

扫描二维码，获取 PDF 全文