DOI:10.19651/j.cnki.emt.2518813

基于改进 RT-DETR 的遥感图像目标检测算法*

肖 锋[1] 杨文豪[1] 张文娟[2] 黄姝娟[1] 周雨浩[3]
(1.西安工业大学兵器科学与技术学院 西安 710016;2.西安工业大学基础学院 西安 710016;
3.西安工业大学计算机科学与工程学院 西安 710016)

摘 要:遥感图像中的目标常呈细长、曲折等复杂形态,且伴随尺度变化大与背景干扰强等因素,导致现有检测方法易出现缺检和误检,难以满足高精度检测需求,为此,提出一种改进的遥感图像目标检测算法 TriD-DETR。首先,通过动态调整卷积核形状并优化通道适配与残差连接方式,设计了 DKFE 特征提取模块,该模块能够自适应地聚焦于细长曲折的局部区域,从而准确捕捉目标特征;其次,为了提高模型对复杂目标的定位和识别能力,提出 DATE 尺度内特征交互结构,在重构 Transformer 编码器的基础上引入可变形注意力机制,增强了模型对高级特征和深层语义信息的捕捉能力;最后,针对多尺度特征融合部分,提出 DBFB 多样性分支融合模块,通过组合不同尺度和复杂度的多样性分支使特征空间更丰富,从而增强模型的表达能力。实验结果表明,TriD-DETR 算法在 DIOR 和 RSOD 数据集上分别达到 86.8%和 94.1%的 mAP,相较于原模型 RT-DETR-R18,分别提升了 1.2%和 2.3%,充分证明了 TriD-DETR 算法的可靠性与高效性。

关键词:遥感图像;目标检测;RT-DETR;注意力机制;多尺度特征融合
中图分类号:TP751;TN98 **文献标识码**:A **国家标准学科分类代码**:520.2070

Enhanced remote sensing image target detection algorithm based on the improved RT-DETR

Xiao Feng[1] Yang Wenhao[1] Zhang Wenjuan[2] Huang Shujuan[1] Zhou Yujie[3]
(1. School of Ordnance Science and Technology, Xi′an Technological University, Xi′an 710016, China; 2. School of Sciences, Xi′an Technological University, Xi′an 710016, China; 3. School of Computer Science and Engineering, Xi′an Technological University, Xi′an 710016, China)

Abstract: Targets in remote sensing images are often elongated, zigzagging and other complex morphology, and accompanied by large scale changes and strong background interference and other factors, resulting in the existing detection methods are prone to lack of detection and misdetection, it is difficult to meet the demand for high-precision detection, in this regard, an improved remote sensing image target detection algorithm TriD-DETR. First, by dynamically adjusting the shape of convolutional kernel and optimizing the channel adaptation and residual connection methods, a DKFE feature extraction module is designed, which is able to adaptively focus on the elongated and zigzagging local regions, thus accurately capturing the target features; second, in order to improve the model′s ability of locating and identifying the complex targets. DATE in-scale feature interaction structure is proposed, which introduces a deformable attention mechanism on the basis of reconfiguring the Transformer encoder and enhances the model′s ability to capture high-level features and deep semantic information; finally, for the multi-scale feature fusion part, the DBFB diverse branch fusion block, which enriches the feature space by combining diverse branches of different scales and complexity, thus enhancing the expressive ability of the model. The experimental results show that the TriD-DETR algorithm achieves 86.8% and 94.1% mAP on the DIOR and RSOD datasets, respectively, which are 1.2% and 2.3% higher than the original model RT-DETR-R18, which fully proves the reliability and efficiency of the TriD-DETR algorithm.

Keywords: remote sensing images; target detection; RT-DETR; attention mechanism; multi-scale feature fusion

0 引 言

遥感图像目标检测是一项基本的计算机视觉任务,其目的是识别和定位光学遥感图像中的目标。现有目标检测模型有两种典型的架构:基于卷积神经网络(convolutional neural network,CNN)和基于 Transformer[1]架构。

在过去几年,人们对基于 CNN 架构的目标检测模型进行了大量研究,从刚开始的两阶段检测器[2-4]到单阶段检

收稿日期:2025-05-14
*基金项目:国家自然科学基金面上项目(62171361)、国家自然科学基金青年项目(52302505)、陕西省科技厅重点研发计划项目(2023-YBGY-027)、陕西省教育厅专项科研计划项目(22JK0412)资助

测器[5-15],从 anchor-based[5,8-10,12] 到 anchor-free[6,11,13-15],这些研究在检测速度和准确性方面都取得了明显的进步。然而,CNN 架构在处理长距离依赖和全局上下文信息时存在局限性,导致其在复杂背景和尺度差异较大的场景中往往表现不佳。

近年来,基于 Transformer 的目标检测模型因其出色的全局建模能力,逐渐引起了广泛关注。特别是 Carion 等[16]提出的 DETR(detection transformer)模型,填补了 Transformer 架构在目标检测领域的空白,它采用二分图匹配对每个真实边框做出唯一预测边框,同时消除了很多人工设计的组件,例如非极大值抑制(non-maximum suppression, NMS),极大简化了目标检测流程,实现了端到端的目标检测模型。然而,DETR 模型也存在计算复杂度高和收敛速度慢等问题,针对这些问题,多个 DETR 的变体在不同方面进行了优化。李青云等[17]提出的 RSH-RTDETR,针对特征提取及采样方式进行了优化;Zhu 等[18]提出 Deformable-DETR,通过增强注意力机制的效率,加速了多尺度特征的训练收敛过程;Meng 等[19]提出的 Conditional DETR 和 Wang 等[20]提出的 Anchor DETR 都降低了 query 的优化难度;Liu 等[21]提出的 DAB-DETR 引入了四维参考点,并逐层迭代优化预测框;Li 等[22]提出的 DN-DETR 通过引入 query 去噪,进而缩短了训练收敛时间;Chen 等[23]提出的 Group-DETR 通过引入分组式一对多分配(group-wise one-to-many assignment)来加速训练过程;Zhang 等[24]提出的 DINO 模型在以上的工作基础上进行构建优化,并得到了不错的效果;尽管这些变体在各自的方面取得一定的进展,但 Zhao 等[25]提出的 RT-DETR(real-time detection transformer)凭借其高效的混合编码器设计,展现出更为显著的优势,该设计通过将编码器解耦成尺度内特征交互和跨尺度特征融合两部分来有效地处理多尺度特征,从而有效改善了 DETR 模型的高计算量以及收敛速度慢等问题,使其可以在各种实时场景中进行实际应用。然而,尽管 RT-DETR 在多个方面表现优异,但仍存在一些局限性,例如,它在处理细长、弯曲等复杂目标时,检测性能较差,且易出现漏检或误检现象。

为了克服以上局限性,并延续 RT-DETR 的高效设计理念,提出了改进模型 TriD-DETR (triple-driven detection transformer),该方法主要贡献如下:

1)在特征提取阶段,构建了动态卷积核特征提取模块(dynamic kernel feature extractor, DKFE),基于蛇形卷积[26]在复杂几何结构方面的自适应特性,进一步设计了可动态调整卷积核形状的机制,使其能够聚焦于目标的细长、曲折等复杂结构区域。同时,优化通道适配与残差连接方式,提升了信息流动效率与特征表达能力,从而有效捕捉传统卷积操作可能忽略的细微信息;

2)提出尺度内特征交互结构(deformable attention transformer encoder,DATE),通过优化 Transformer 编码器结构,并融合可变形注意力机制[27],增强了对全局信息的捕捉能力,DATE 通过自适应调整注意力的范围和形状,能够更好地适应图像中的局部结构和形变,从而精准捕捉关键细节。尤其在复杂场景中,DATE 能够有效提高对局部特征的关注度,减少因目标形变或背景复杂性导致的漏检和误检,最终提升检测性能;

3)在跨尺度特征融合部分,受 Diverse Branch Block[28]思想的启发,提出多样性分支融合模块(diverse branch fusion block,DBFB),DBFB 在 RT-DETR 中融合模块的基础上,结合不同尺寸和复杂度的多样性分支丰富特征空间,进而增强单个卷积的表达能力,通过有效的多尺度特征融合,DBFB 能够在不增加任何推理时间成本的前提下,提高网络对不同尺度目标的检测性能,从而增强了模型的表达能力和泛化能力。

1　TriD-DETR 网络结构设计

1.1　整体网络模型结构设计

TriD-DETR 的整体网络架构如图 1 所示,主要包括 4 个模块:骨干网络、高效混合编码器、解码器和检测头。首先,采用改进后的 ResNet－18 作为主干网络,该网络由基本卷积层、最大池化层、ResNet Basic Block 和 DKFE 模块组成,主要用于从输入图像中提取初步特征。Basic Block 捕获输入图像的低尺度特征(即图像的局部细节信息),然后利用 DKFE 模块的高层特征表达能力提取高尺度特征(即图像的全局和语义信息),最终输出 3 个不同尺度的特征图 $P3$、$P4$、$P5$。其次,DATE 结构对 Transformer 编码器进行了优化,并引入可变形注意力机制,从而使模型能够更有效地捕捉全局信息,生成处理后的高尺度特征 $F5$。为了充分利用不同尺度特征图之间的互补性,采用基础卷积、上采样技术以及所提出的 DBFB 模块,对 $P3$、$P4$、$F5$ 这 3 个尺度的特征图进行有效融合,从而得到 3 个尺度的融合特征图 $Y3$、$Y4$、$Y5$。该过程不仅结合了低尺度特征图的局部细节信息和高尺度特征图的全局信息,还通过多分支结构增强了特征的多样性。最后,将这 3 个尺度的融合特征图输入解码器和检测头,用于预测目标的位置、类别及置信度。

1.2　DKFE 模块

ResNet-R18 作为一种广受欢迎的 CNN 架构,常常作为特征提取的骨干网络使用。然而,它也存在一些固有的局限性。首先,相较于大型网络,ResNet-R18 在处理具有复杂纹理和结构的图像时,其特征提取能力稍显逊色;其次,受限于其网络结构,ResNet-R18 的感受野相对较小,这在一定程度上限制了它的性能表现,特别是在需要全局上下文信息的任务中。为了突破这些局限,提出 DKFE 结构,实现动态调整卷积核的操作,并对通道适配和残差连接方式进行了优化,旨在提升网络对复杂图像特征的提取能力,从而增强模型在处理全局上下文信息任务中的性能。

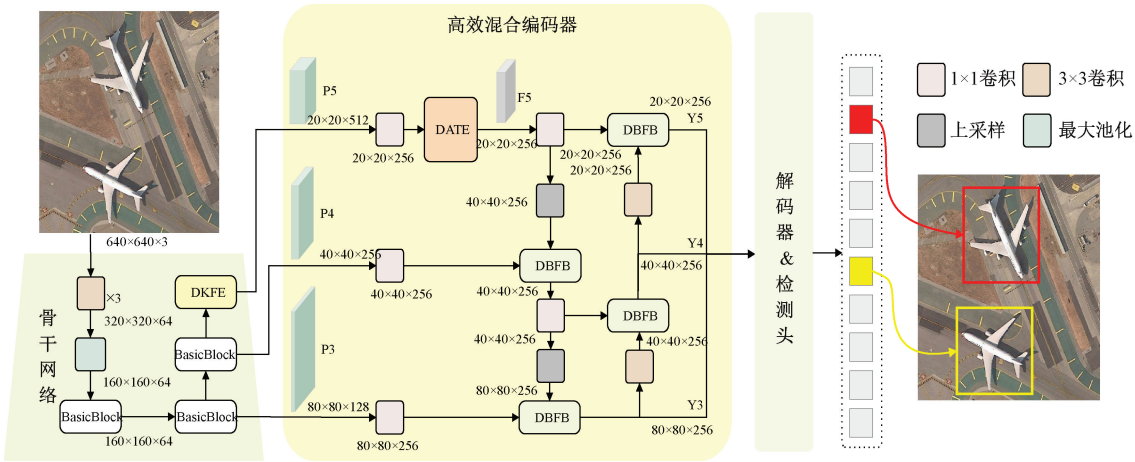


图 1　TriD-DETR 模型整体结构

Fig. 1　The general framework of TriD-DETR

与传统卷积的固定形状卷积核不同，DKFE 模块能够根据输入特征图的内容学习偏移量，从而动态调整卷积核形状。偏移量生成网络负责学习每个采样点在二维空间中的形变信息，引导卷积核自适应地贴合细节丰富的复杂结构。该网络由 3×3 卷积层、批归一化层和 Tanh 激活函数构成。对于输入特征图（$C\times H\times W$），单层 3×3 卷积层用于生成每个采样点的偏移量，批归一化层进行特征标准化，为控制变形幅度，采用 Tanh 激活函数将偏移量压缩至[−1,1]区间，最终输出偏移量张量（$2K\times H\times W$），其中 K 为卷积核大小，前 K 个通道对应 y 方向偏移，后 K 个通道对应 x 方向偏移。

给定标准的平面卷积坐标集合，以中心坐标 $K_i=(x_i,\ y_i)$ 为基准，一个 3×3 的卷积核 K 可表达为式(1)：

$$K=\{(x-1,y-1),(x-1,y),\cdots,(x+1,y+1)\} \tag{1}$$

卷积核在 x 轴和 y 轴方向的线性初始化定义如下：x 方向卷积时 y 坐标固定为 0，x 坐标线性分布于$[-K//2,\ K//2]$；y 方向卷积时 x 坐标固定为 0，y 坐标线性分布于$[-K//2,\ K//2]$。以中心坐标 K_i 为起点，K_{i+1} 相对于 K_i 增加了偏移量 $\Delta=\{\delta\mid\delta\in[-1,\ 1]\}$，通过迭代累加的方式，确保卷积核符合线性形态结构。在 x 轴方向，动态调整过程如式(2)所示。

$$K_{i\pm c}\begin{cases}(x_{i+c},y_{i+c})=(x_i+c,y_i+\sum_{i}^{i+c}\Delta y)\\(x_{i-c},y_{i-c})=(x_i-c,\ y_i+\sum_{i-c}^{i}\Delta y)\end{cases} \tag{2}$$

相应地，y 轴方向的变化如式(3)所示。

$$K_{j\pm c}\begin{cases}(x_{j+c},\ y_{j+c})=(x_j+\sum_{j}^{j+c}\Delta x,y_j+c)\\(x_{j-c},y_{j-c})=(x_j+\sum_{j-c}^{j}\Delta x,y_j-c)\end{cases} \tag{3}$$

由于偏移量 Δ 通常是小数，而像素坐标通常为整数形式，因此采用双线性插值来处理：

$$K=\sum\nolimits_{K'}B(K',\ K)\cdot K' \tag{4}$$

其中，K 表示由式(2)和式(3)得到的小数位置坐标，K' 列举所有整数空间位置，B 是双线性插值权重函数，可以如式(5)所示分解为两个一维核进行计算：

$$B(K,\ K')=b(K_x,\ K'_x)\cdot b(K_y,\ K'_y) \tag{5}$$

重复以上步骤，最终得到适应曲折细长结构的调整后卷积核，以更好地捕捉关键特征，从而有效保留并突出目标的细节特征。

DKFE 模块的结构如图 2 所示，其处理流程高效且直观。首先，输入特征图 $\boldsymbol{X}$ 经过 3×3 卷积层、批量归一化层以及 ReLU 激活函数，完成了初步的特征提取；紧接着，通过动态调整卷积核形状使其适应复杂几何特征，从而对特征信息进一步细化，并再次利用 ReLU 激活，随后又经过一次 3×3 卷积和 ReLU 激活，得到中间输出 out_1；接着，进行 shortcut 检查以适配通道数，如果匹配，则直接将输入 $\boldsymbol{X}$ 与 out_1 进行残差连接，若不匹配，则通过 1×1 卷积层调整 $\boldsymbol{X}$ 的维度，得到输出 out_2；之后，将输出 out_1 与 out_2 进行残差连接，这有助于网络学习输入与输出间的残差，推动网络向更深层次学习；最后，通过 ReLU 激活函数处理残差连接的结果，生成高尺度特征图 $P5$，为后续网络层提供更为精细的特征信息。

DKFE 模块通过动态调整卷积核，并优化通道适配和残差连接方式等关键环节，可以出色适应细长且弯曲的局部特征，从而更精确地捕捉和学习图像中的特征，尤其在处理图像中复杂几何形状和拓扑结构的元素，如轮廓线、道路及管状结构时表现更为优异。

1.3　DATE 结构

RT-DETR 在处理高尺度特征时，通过单个 Transformer 编码器实现尺度内的信息交互。虽然这种方法能够高效地捕捉同一尺度下的目标信息，但其在检测小目标时表现较弱，容易出现缺检和误检，尤其是在背景复杂

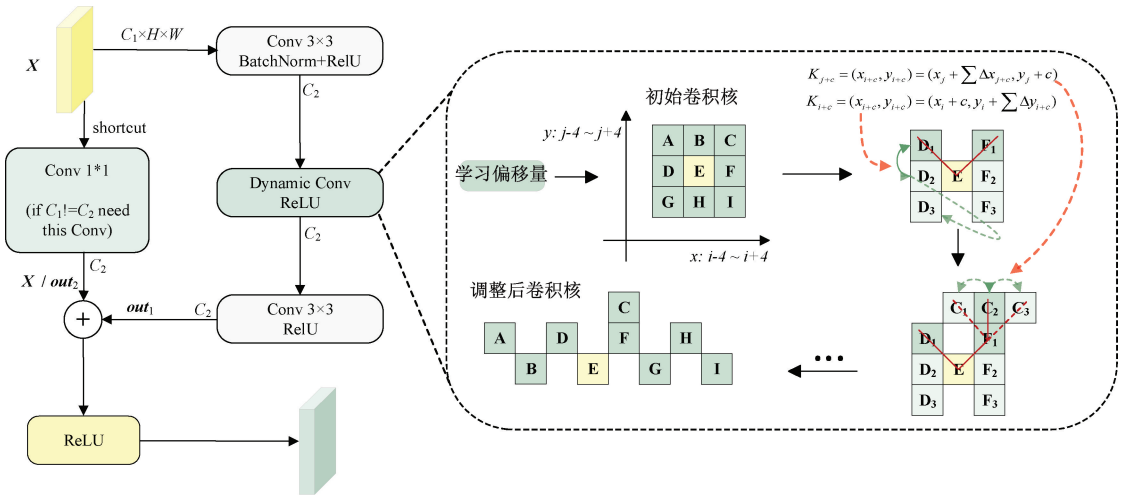


图 2　DKFE 模块结构

Fig. 2　DKFE structure

或目标尺度变化较大的场景中。

针对以上问题,提出 DATE 尺度内特征交互结构,优化了 Transformer 编码器结构,并引入可变形注意力机制。优化 Transformer 编码器结构提升了特征表示能力,从而增强了对复杂目标的识别能力。引入可变形注意力机制带来了两个主要优势:首先,它能够自适应地调整感受野的大小和形状,使模型更精准地聚焦于目标的关键区域,从而提升检测精度;其次,可变形注意力机制通过灵活选择注意力范围,有效抑制了背景干扰,减少了误检和漏检现象。通过这些优化,DATE 结构在捕捉局部细节和目标形变方面表现更为出色,尤其在小目标检测中展现了明显优势。同时,它避免了传统方法中由于固定感受野造成的局部特征丢失问题,显著提升了整体检测性能。

DATE 结构及其特征交互流程如图 3 所示。首先,输入的二维特征图 $P5$ 被转化为一系列向量,并嵌入位置编码,这些向量同时携带特征信息和位置信息。接着,这些向量进入 DATE,该部分融合了可变形注意力机制、残差连接、归一化操作和多层感知机,可以自适应地调整感受野的大小从而高效处理输入数据。最后,处理后的向量被重新转换为二维特征图 $F5$ 并输出,完成高尺度的特征交互。

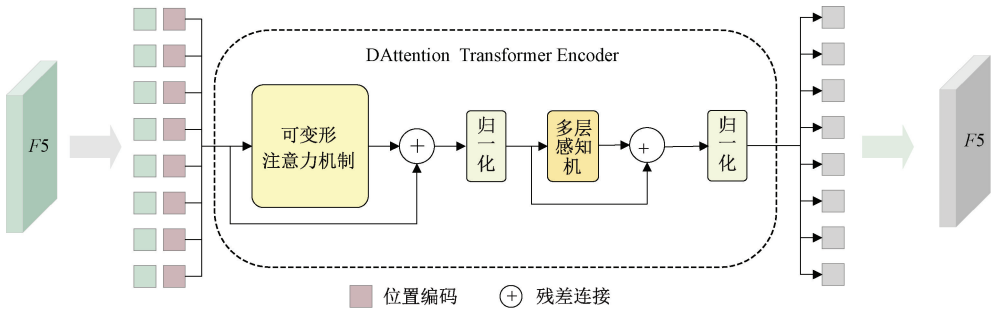


图 3　DATE 结构

Fig. 3　DATE structure

在基于注意力的特征处理过程中,输入特征图 $x \in R^{H \times W \times C}$ 通过线性投影转化为 query token $q = xW_q$,然后输入到轻量子网络 $\theta_{offset}(\cdot)$,生成偏移量 Δp;然后在变形点的位置进行特征采样,作为变形 key 和变形 value,投影矩阵及计算过程如式(6)和式(7)所示。

$$\Delta p = \theta_{offset}(q),\ \widetilde{x} = \varphi(x;\ p + \Delta p) \tag{6}$$

$$q = xW_q,\ \widetilde{k} = \widetilde{x}W_k,\ \widetilde{v} = \widetilde{x}W_v \tag{7}$$

其中,$\widetilde{k}$ 和 $\widetilde{v}$ 分别表示变形之后的 key 与 value;接下来对 q ,$\widetilde{k}$,$\widetilde{v}$ 进行多头注意操作,一个注意力头的输出如式(8)所示。

$$z^{(m)} = \sigma\left(q^{(m)}\widetilde{k}(m)^{\mathrm{T}}/\sqrt{d} + \varphi(\hat{B};\ R)\right)\widetilde{v}^{(m)} \tag{8}$$

其中,$\boldsymbol{R}$ 为相对位置偏移量,每个注意力头利用 $q^{(m)}$,$\widetilde{k}(m)$ 计算注意力得分,并根据这些得分加权求和变形 $\widetilde{v}^{(m)}$ 来获取最终的输出 $z^{(m)}$。这一过程可以确保能够根

据目标的局部特征动态聚焦于关键区域，从而提升小目标的检测能力，显著减少缺检和误检现象。

1.4　DBFB 模块

跨尺度特征融合是高效混合编码器的关键部分，它通过结合不同尺度的信息来增强模型的性能和鲁棒性，不同尺度的特征分别捕捉到图像的全局信息和局部细节，从而使得模型能够更好地平衡这两方面信息，进而提高图像识别的准确性。提出的 DBFB 多样性分支融合模块，旨在实现跨尺度特征的有效整合，该设计不仅能融合不同来源、不同尺度和不同层次的特征，还能转化为更丰富和全面的特征表示，且在推理过程中不会增加额外的时间开销。

DBFB 融合块中的 Diverse Branch Block 运用网络结构重参数化的思想，即构造一系列结构（用于训练），并将其参数等价转换为另一组参数（用于推理），从而将这一系列结构等价转换为另一系列结构。Diverse Branch Block 其中包含了多个不同感受野、不同复杂度的分支，每个分支都可以独立地进行运算，然后将各个分支的结果进行整合，以生成最终的特征表达。其优势在于，它可以显著提升原有模型的精度，同时在推理阶段，通过网络结构重参数化的方法，可以将 Diverse Branch Block 等价地转换为一个普通的卷积层，从而在不改变模型结构、计算量和推理时间的前提下，实现无损的性能提升。

DBFB 结构示意图如图 4 所示，Diverse Branch Block 采用多分支拓扑结构，包含了多尺度的卷积、$1\times1-K\times K$ 卷积、全局平均池化、批量归一化、多分支相加等操作，其目的就在于增强原来的 $K\times K$ 卷积，这种具有不同感受野和不同复杂度的操作可以极大的丰富特征空间。具体来说，首先，对于 $1\times1-K\times K$ 分支，将其内部通道数设置为与输入通道相同，并将 1×1 卷积核初始化为单位矩阵，其他卷积核则正常初始化；其次，每一个卷积或平均池化层操作（average pooling，AVG）后面都衔接着一个批量归一化（batch normalization，BN），这在训练时提供了非线性，也有助于加速训练过程并提高模型的收敛速度；最后，在 Diverse Branch Block 中加入非线性层（nonlinearity），因为非线性层可以帮助模型学习更复杂的函数映射以及更复杂的非线性关系，以帮助模型更好地适应各种类型的数据，同时可以提高模型的表达能力和灵活性。

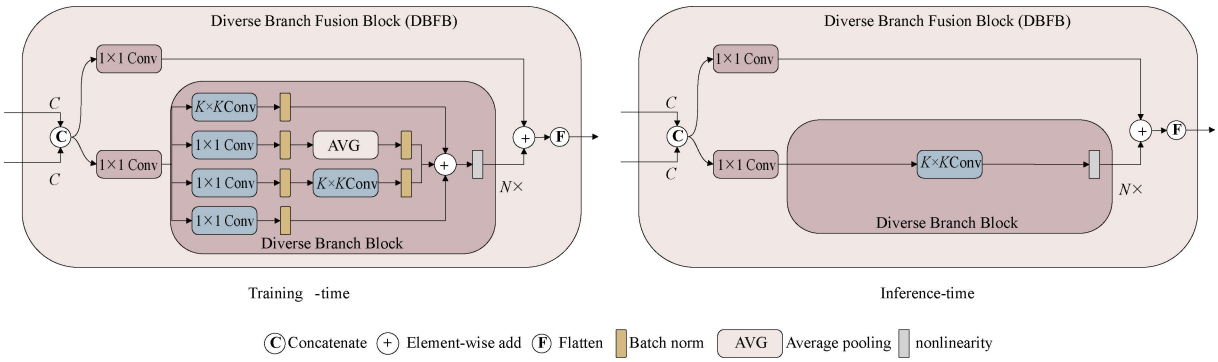


图 4　DBFB 结构

Fig. 4　DBFB structure

综上所述，DBFB 模块通过在训练阶段增加模型复杂度，在推理阶段将其转化为原始结构，从而解耦了训练和推理时的网络结构。这一策略在提升模型性能的同时，避免了额外增加推理时间。DBFB 融合块的优势在于，它能够在不牺牲推理速度的前提下，充分利用训练阶段的数据多样性和模型复杂性，从而有效提升多尺度特征的交互能力。此外，DBFB 的出色性能和强大可扩展性使其能够轻松集成到现有网络架构中，进一步增强原有模型的性能。

2　实验设计与结果分析

2.1　实验环境

实验训练阶段使用的硬件平台和环境参数配置如表 1 所示，利用 CUDA 11.3 和 Cudnn 8.2.0 对实验进行加速，选用 PyTorch（1.11.0＋cu113）作为深度学习框架。在训练时，设置输入图片为 640×640，训练周期 epoch 为 200，批次大小 batchsize 为 4，初始学习速率取 0.000 1。

表 1　训练环境和参数配置表

Table 1　Training environment and parameter configuration table

环境	参数配置
操作系统	Ubuntu 18.04
CPU	Intel Xeon Platinum 8255C
GPU	NVIDIA GeForce RTX 3090
显存	24 G
内存	43 G
开发环境	Pycharm
编程语言	Python

2.2　实验数据集

DIOR 数据集由 Li 等[29]构建，共有 23 463 张图像，采集自谷歌地球（Google Inc.），图像大小均为 800 pixel×800 pixel，空间分辨率为 0.5 ～ 30 m。DIOR 数据集共标

记了 192 472 个检测对象,囊括了 20 种遥感图像常见的类别,包括飞机、机场、桥梁、大坝、车辆等 20 个类别。为了保证各类样本的均衡分布,采用按类别进行分层抽样的方式,以 7∶2∶1 的数量比例分层抽样,划分为训练集、验证集以及测试集,分别包含 16 424、4 692、2 347 张图像。DIOR 数据集的显著优势体现在以下几个方面:首先,其图像和目标类别的数量庞大,提供了丰富的样本;其次,目标尺寸的变化范围广泛,检验模型的泛化能力;最后,数据集具有高类间相似性和类内相似性,增大了精确识别的难度。其部分示例如图 5 所示。

RSOD 数据集由武汉大学团队 Long 等[30]标注,数据来源于 Google Earth 和天地图,涵盖四大类别:飞机、油箱、操场和立交桥,总计 976 幅影像,共标注 6 950 个目标区域。数据集中的影像尺寸为 920×1 050,空间分辨率为 0.3～3 m。采用基于类别的分层抽样方法,并按照 8∶1∶1 的比例划分为训练集、验证集与测试集,分别包含 782、97 和 97 幅图像,从而在各子集中维持类别分布的一致性。

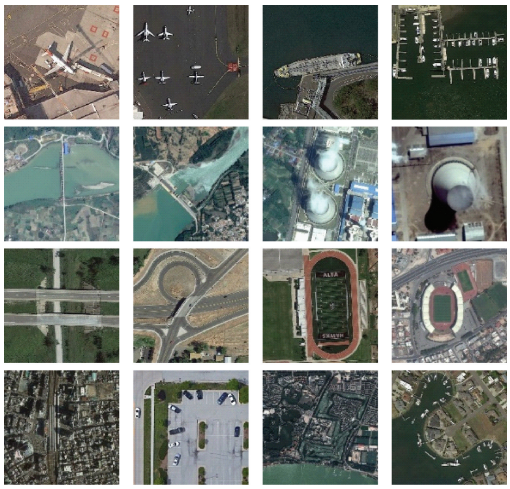


图 5　DIOR 数据集部分示例图

Fig. 5　DIOR dataset partial sample diagram

2.3　评价指标

本次针对遥感目标检测任务,实验采用精确度(precision)、召回率(recall)、平均精度 mAP(mean average precision, mAP)、模型参数量(params)、测试时每秒浮点运算次数(floating point operations per second, FLOPs)指标作为评判算法的性能。Precision、recall、AP、mAP 计算如式(9)所示。

$$\begin{cases} P = \dfrac{TP}{TP + FP} \\ R = \dfrac{TP}{TP + FN} \\ AP_c = \dfrac{1}{N_c} \displaystyle\sum_{r_c \in R_c} p(r_c) \\ mAP = \dfrac{1}{N} \sum AP_c \end{cases} \tag{9}$$

其中,TP 代表真正例(true positives, TP),即模型正确地将实际为正样本的实例预测为正;FP 代表假正例(false positives, FP),即模型错误地将实际为负样本的实例预测为正;FN 代表假反例(false negatives, FN),即模型错误地将实际为正样本的实例预测为负。精度(P)衡量的是模型在所有预测为正例的样本中,真正例所占的比例;而召回率(R)则反映了模型找到所有正例的能力,即预测为正例的样本中真正例所占实际正例总数的比例。为了全面评估模型性能,引入 AP(平均精度),而 mAP(平均精度均值)则是对多个类别平均精度的汇总。在评估模型效率方面,使用了 FLOPs(浮点运算次数)来衡量模型的计算复杂度,Params(参数数量)来评估模型的轻量化程度。

2.4　DKFE 模块特征提取效果分析

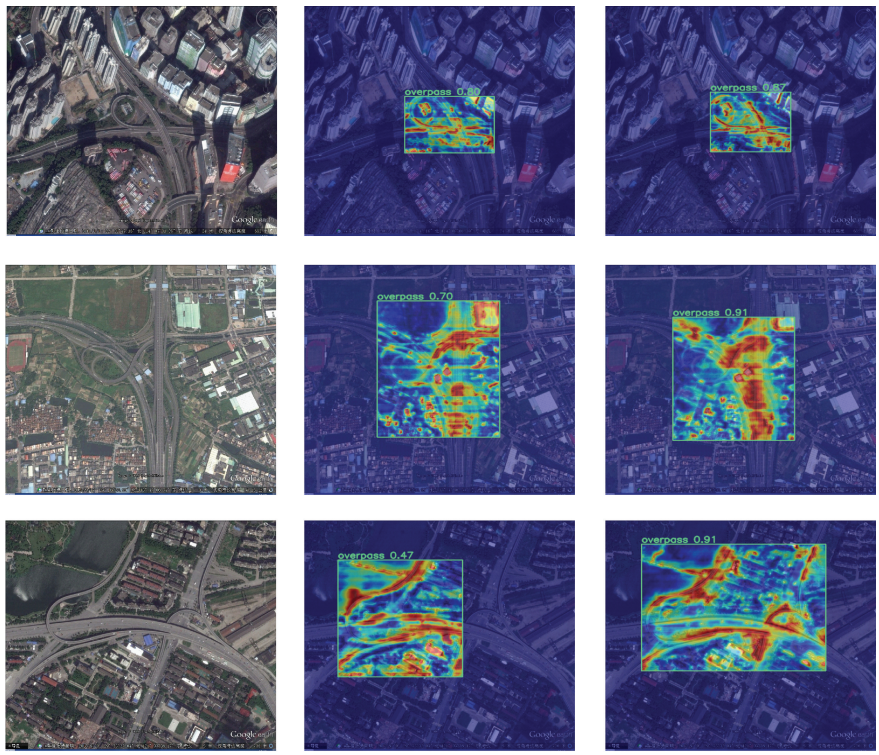
为验证 DKFE 结构在细长曲折目标特征提取中的有效性,采用 Grad-CAM＋＋技术对骨干网络中基础模块 Basic Block 和 DKFE 模块输出的特征图中目标区域进行热力图可视化分析,为更直观展示模型的关注区域,热力图仅在检测到目标的区域生成,并对颜色强度归一化处理(范围为 0～1),红色表示高响应,蓝色表示低响应,结果如图 6 所示。

从热力图对比可以观察到,DKFE 在目标区域产生更高的响应强度,显著改善特征图的空间分布,使关注更加集中。例如,在第 1 组中,DKFE 对弯曲立交桥的响应区域更紧凑,热力图准确覆盖桥体轮廓,展现了对细长目标的精准定位能力;第 2 组中,DKFE 对复杂交叉路口捕捉全面,热力集中于关键位置,而 Basic Block 响应分散,难以突出目标细节。进一步分析发现,DKFE 的动态适应性使其能根据目标几何形态灵活调整卷积路径,从而增强对曲折目标的特征表达能力。尤其在第 3 组中,DKFE 精确捕捉了蛇形目标的局部细节,同时保留整体结构,展现了全局特征表示能力,弥补了 Basic Block 在边缘与关键区域响应不足的缺陷。

综上所述,图 6 的热力图对比充分证明 DKFE 在高复杂度遥感目标检测任务中的卓越表现。通过动态调整卷积核,DKFE 可以高效提取细长、曲折的目标特征,增强网络对关键区域的响应,并可以精准覆盖目标轮廓与细节。

2.5　DATE 结构可视化及性能分析

为了验证 DATE 结构在减少漏检和误检方面的改进效果,采用可视化热力图对比分析 RT-DETR-R18 中尺度内特征交互层(intra-scale feature interaction, AIFI)和 TriD-DETR 中 DATE 层中输出特征图,重点评估其在目标区域的检测表现,热力图颜色强度已归一化至[0,1]范围,其中红色表示高响应区域,蓝色表示低响应区域,相关可视化结果如图 7 所示。在 AIFI 的可视化结果中,红色实线框表示漏检区域,绿色虚线框表示误检区域。结果显示,AIFI 在复杂背景下难以有效区分目标与干扰区域,漏检现象主要集中在背景边缘,且部分背景区域发生明显误

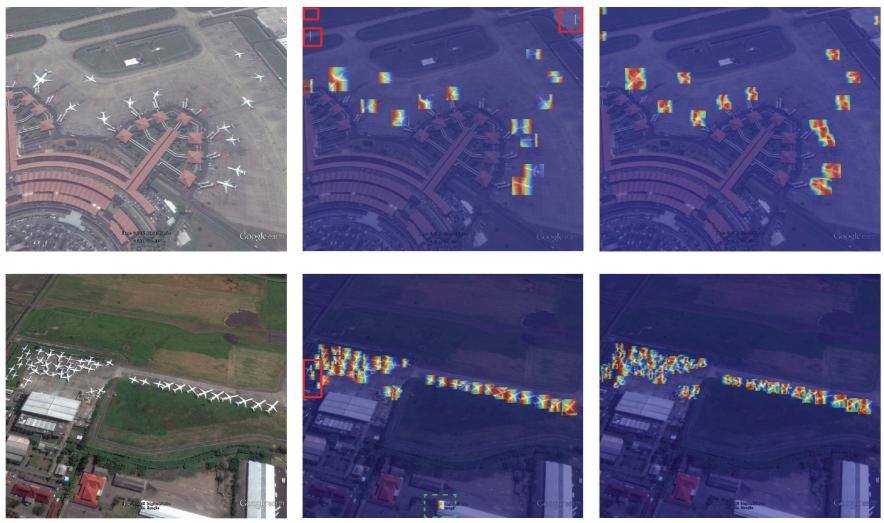


(a) 原图　　(b) Basic Block　　(c) DKFE

(a) Original images

图 6　DKFE 与 Basic Block 热力图可视化对比

Fig. 6　Heatmap visualization comparison of DKFE and basic block



(a) 原图　　(b) Basic Block　　(c) DKFE

(a) Original images

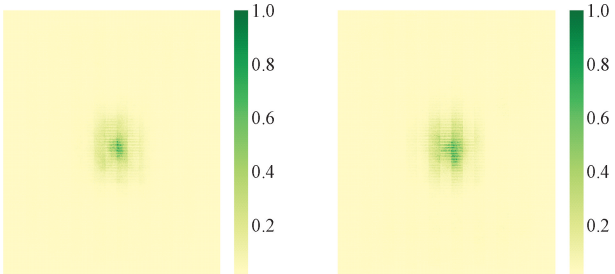
图 7　DATE 与 AIFI 热力图可视化对比

Fig. 7　Heatmap visualization comparison of DATE and AIFI

检，例如建筑被误标为飞机。而 DATE 结构通过引入可变形感受野，自适应调整注意力范围和形状，有效增强了对目标区域的响应，在高背景干扰和目标形态复杂的场景中展现出明显优势，不仅能够精准捕获局部细节，还显著增强了对目标整体的关注能力。

有效感受野(effective receptive field，ERF)指网络中每个神经元能够感知的图像区域，它直接影响模型对局部与全局特征的捕捉能力。为进一步揭示 DATE 结构的内

在优势,分析其有效感受野的分布并进行可视化对比,如图 8 所示。AIFI 层的有效感受野(图 8(a))表现出固定且局限的特性,响应主要集中在较小范围,难以全面覆盖目标的全局特征。在形态多变或背景复杂的场景下,这种刚性分布容易导致感受野过小或过度聚焦非目标区域。相比之下,DATE 层的有效感受野(图 8(b))展示了更广阔且自适应的分布特性,能够根据目标需求进行动态扩展,覆盖更大范围,同时增强对局部细节的捕捉能力,并有效抑制背景干扰,在复杂目标形态和复杂背景检测中能够实现更优的检测性能。



(a) AFI　　(b) DATE

图 8　AIFI 与 DATE 层有效感受野分布对比图

Fig. 8　Effective receptive field distribution comparison of AIFI and DATE layer

最后对不同覆盖比例(t 值)下的有效感受野进行了统计对比,如表 2 所示。实验结果表明,DATE 在高覆盖率($t=99\%$)下的有效感受野范围显著优于 AIFI,显示了其更强的全局关注能力,能够有效关注到目标的整体信息。同时,在中低覆盖率(如 $t=20\%$、$t=50\%$)下,DATE 的有效感受野分布更均匀,体现了全局与局部特征提取的良好平衡性。

综上所述,DATE 结构不仅增强了对目标区域的聚焦能力,还在适应目标尺度变化、形态复杂性以及背景干扰方面展现了显著优势,进一步验证了其在目标检测任务中有效减少误检与漏检的能力。

表 2　AIFI 与 DATE 不同覆盖率下有效感受野定量对比

Table 2　Quantitative comparison of ERF under different coverage ratios for AIFI and DATE

结构	$t=20\%$	$t=30\%$	$t=50\%$	$t=99\%$
AIFI	1.3	2.5	7.3	92.3
DATE	1.5	2.8	8.5	93.5

2.6　DBFB 模块实验结果分析

为验证 DBFB 模块在目标检测任务中的性能提升效果及可扩展性,将其集成到多种主流目标检测模型中进行实验。选取 RT-DETR-R18、RT-DETR-R34、YOLOv8n、YOLOv8s、YOLOv9s 和 YOLOv10 s 作为基准模型,并基于这些模型构建引入 DBFB 模块的改进版本(标记为+DBFB)。所有实验均在相同的实验条件下进行,使用遥感图像目标检测数据集 RSOD 进行训练和评估,实验结果如表 3 所示。

表 3　RSOD 数据集上各模型添加 DBFB 模块的实验结果

Table 3　Experimental results of adding DBFB module to models on RSOD dataset

模型	Params/MB	FLOPs/GFLOPs	Inference time/ms	mAP50/%
YOLOv8n	3.1	8.1	17.7	90.2
YOLOv8n+**DBFB**	4.5	11.5	11.3	**90.6**(+0.4)
YOLOv8s	13.1	28.5	21.7	92.5
YOLOv8s+**DBFB**	17.1	42.2	16.0	**92.7**(+0.2)
YOLOv9s	6.8	26.7	13.5	92.3
YOLOv9s+**DBFB**	8.1	30.7	14.2	**93.5**(+1.2)
YOLOv10s	7.6	24.5	8.2	87.9
YOLOv10s+**DBFB**	10.1	34.9	8.4	**92.9**(+5.0)
RT-DETR-R34	29.6	88.8	31.8	93.2
RT-DETR-R34+**DBFB**	29.6	88.9	32.1	**94.1**(+0.9)
RT-DETR-R18	18.9	57.0	33.0	91.8
RT-DETR-R18+**DBFB**	18.9	57.0	34.2	**92.5**(+0.7)

实验结果显示,引入 DBFB 模块后,各模型的 mAP50 指标均有不同程度的提升,其中 YOLOv8n 和 YOLOv8s 分别提升了 0.4%和 0.2%,YOLOv9s 和 YOLOv10 s 分别提升了 1.2%和 5.0%,RT-DETR-R18 和 RT-DETR-R34 分别提升了 0.7%和 0.9%。这表明 DBFB 模块在不同量级的网络架构上均能稳定适配并显著提升性能。尽管引入 DBFB 模块增加了模型的参数量与计算复杂度,但这种增加相对可控,对推理时间的影响较小。DBFB 模块通过增强了卷积的特征表达能力,进而丰富了特征空间表示,同时在保持计算效率的前提下,各模型性能得到了稳定提

升，充分验证了 DBFB 模块在遥感图像目标检测任务中的高效性和实用价值。

2.7　消融实验

为了验证本文提出的每种改进模块的有效性，采用 RT-DETR-R18 作为基线网络，在数据集 DIOR 和 RSOD 上进行消融实验，所有实验均在相同的实验条件和参数设置下进行。

消融实验结果如表 4 所示，其中“√”表示使用了相应方法。在基线模型 RT-DETR-R18 中加入结构后，DIOR 和 RSOD 数据集的 mAP50 分别提升了 0.3%和 0.9%；进一步引入 DATE 结构后，mAP50 分别再提升 0.4%和 0.5%。最终，综合引入 DKFE、DATE 和 DBFB 结构的 TriD-DETR 模型在两数据集上的 mAP50 分别提升了 1.2%和 2.3%，证明了改进措施的有效性。综上所述，这一系列改进在实现网络结构优化的同时，显著提升了检测精度。

表 4　在 DIOR、RSOD 数据集上的消融实验结果

Table 4　Results of ablation experiments on DIOR and RSOD datasets

数据集	RT-DETR-R18	DKFE	DATE	DBFB	Params/MB	FLOPs/GFLOPs	mAP50/%
DIOR	√				**18.9**	**57.0**	85.6
	√	√			26.5	60.9	85.9
	√	√	√		26.6	61.1	86.3
	√	√	√	√	26.6	61.1	**86.8**
RSOD	√				**18.9**	**57.0**	91.8
	√	√			26.5	60.8	92.7
	√	√	√		26.6	61.0	93.2
	√	√	√	√	26.6	61.0	**94.1**

2.8　对比实验

为准确评估 TriD-DETR 算法性能，本文在 DIOR 和 RSOD 数据集上与 8 种常见检测算法进行对比，包括 YOLO 系列、Faster R-CNN、MobileNetV2 和基准模型 RT-DETR-R18。实验结果如表 5 所示。

在 DIOR 数据集上，TriD-DETR 的 mAP50 显著超过 Faster R-CNN 和 MobileNetV2，且分别比 YOLOv3、YOLOv4、YOLOv5s、YOLOv8n、YOLOv8s 和基准模型 RT-DETR-R18 提高了 4.8%、1.5%、1.3%、1.6%、1.0%和 1.2%，在所有对比算法中处于领先地位，展现出卓越的检测精度。在 RSOD 数据集上，TriD-DETR 同样表现优异，mAP50 达到 94.1%。与传统算法 Faster R-CNN 和 MobileNetV2 相比，TriD-DETR 的 mAP50 分别提高了 12.4%和 20.7%，此外，TriD-DETR 在 YOLO 系列和基准模型 RT-DETR-R18 中的表现也有所提升，分别提高了 7.9%、5%、2.8%、3.9%、1.6%和 2.3%。综合来看，TriD-DETR 在保证参数量和计算量适中的同时，显著提升了检测精度，展现了出色的检测能力。

表 5　不同算法的对比实验结果

Table 5　Comparative experimental results of different algorithms

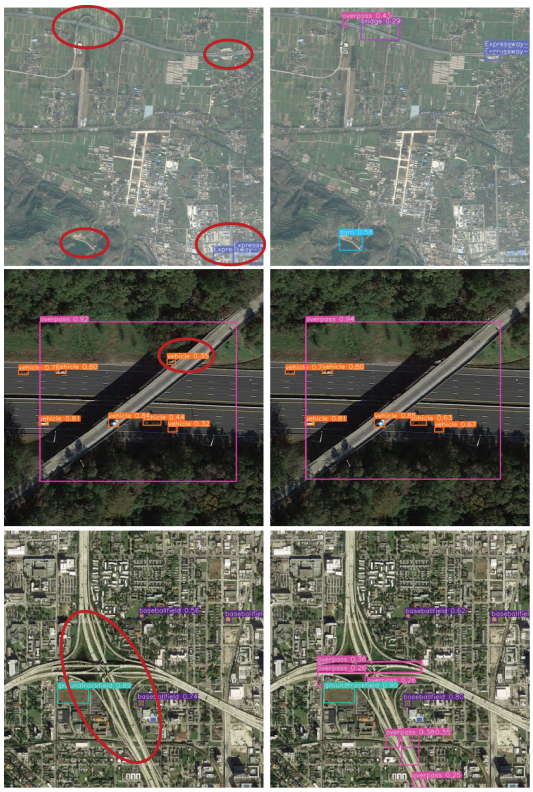
算法	Params/MB		FLOPs/GFLOPs		mAP50/%		$FPS_{bs=1}$/fps	
	DIOR	RSOD	DIOR	RSOD	DIOR	RSOD	DIOR	RSOD
Faster R-CNN	60.2	60.2	182.2	182.1	63.1	81.7	18.5	21.2
MobileNetV2	10.3	10.3	76.1	76.0	58.2	73.4	58.3	62.1
YOLOv3	61.9	61.9	122.2	122.1	82.0	86.2	42.8	46.5
YOLOv4	63.9	63.9	97.4	97.2	85.3	89.1	48.6	52.9
YOLOv5s	9.1	9.1	24.0	23.8	85.5	91.3	72.1	78.4
YOLOv8n	3.1	3.1	8.1	8.1	85.2	90.2	78.5	84.3
YOLOv8s	11.3	11.3	28.5	28.4	85.8	92.5	69.1	75.8
RT-DETR-R18	18.9	18.9	57.0	57.0	85.6	91.8	65.8	71.2
TriD-DETR(our)	26.6	26.6	61.1	61.0	**86.8**	**94.1**	62.5	67.3

TriD-DETR 在 DIOR 和 RSOD 数据集上 FPS(每秒帧率)分别达到 62.5 和 67.3。从推理效率分析，TriD-DETR 在保持高检测精度的同时，展现出良好的实时性能。与同级别 YOLO 系列算法如 YOLOv5s、YOLOv8s

相比,TriD-DETR 在速度上进行了合理权衡,但其在精度方面的显著提升进一步凸显了整体性能优势,综合而言,TriD-DETR 在精度与速度之间实现了良好平衡,为遥感图像目标检测提供了一个高效实用的解决方案。

2.9　可视化分析

为了验证本文提出的改进算法在复杂实际场景中的有效性,在 DIOR 测试集中选取了具有密集、曲折及尺度差异较大的代表性图像进行检测,检测结果如图 9 所示。



(a) RT-DETR　　(b) TriD-DETR

图 9　改进前后检测效果对比

在第 1 组图像中,RT-DETR 在处理细长且曲折的目标时表现不佳,未能检测到图像上方的桥梁、立交桥、高速公路收费站以及左下角的大坝,右下角区域被误识别为收费站,而改进算法准确检测出了这些目标;在第 2 组图像中,RT-DETR 将立交桥阴影误识别为车辆,而改进算法成功避免了这一误检;在第 3 组图像中,RT-DETR 未能检测到细长且曲折的立交桥,而改进算法能够准确识别该目标。与 RT-DETR 相比,本文提出的 TriD-DETR 算法在处理细长、曲折等复杂目标时表现出显著优势,同时减少了漏检和误识别的发生,提升了检测精度。

3　结　　论

本文提出了一种遥感图像目标识别算法——TriD-DETR,旨在解决遥感图像中细长曲折等复杂目标的检测问题,并有效减少漏检和误检。具体而言,在特征提取阶段,通过动态调整卷积核形状,同时优化通道适配与残差连接方式,设计了 DKFE 结构,可以有效捕捉目标的复杂几何特征;接着,提出了 DATE 尺度内特征交互结构,优化了 Transformer 编码器结构,并引入可变形注意力机制,提升了对局部特征的关注度;最后,在跨尺度特征融合阶段,构建了 DBFB 多样性分支融合模块,实现了不同尺度特征的高效整合。实验结果表明,本文算法在 DIOR 和 RSOD 两个遥感数据集上均取得了较高的检测精度。然而,这些改进措施不可避免地增加了模型的参数量和计算复杂度,因此,后续工作将会在保证检测精度和检测速度的前提下,继续探索新的模型剪枝和蒸馏等轻量化技术,以实现更高效、更轻量级的遥感图像目标检测模型。

参考文献

[1] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Neural Information Processing Systems, 2017, 30, DOI:10.48550/arXiv.1706.03762.

[2] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014: 580-587.

[3] REN SH Q, HE K M, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 39(6): 1137-1149.

[4] CAI ZH W, VASCONCELOS N. Cascade R-CNN: Delving into high quality object detection[C]. IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2018:6154-6162.

[5] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017: 2980-2988.

[6] TIAN ZH, SHEN CH H, CHEN H, et al. Fcos: Fully convolutional one-stage object detection[C]. IEEE/CVF International Conference on Computer Vision, 2019: 9627-9636.

[7] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. YOLOv4: Optimal speed and accuracy of object detection[J]. ArXiv preprint arXiv:2004.10934, 2020.

[8] ZHANG Y, GUO ZH Y, WU J Q, et al. Real-time vehicle detection based on improved YOLOv5[J]. Sustainability, 2022, 14(19): 12274.

[9] LONG X, DENG K P, WANG G ZH, et al. PP-YOLO: An effective and efficient implementation of object detector[J]. ArXiv preprint arXiv:

2007.12099, 2020.

[10] 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]. IEEE/CVF Conference on Computer Vision and Pattern Recognition,2023: 7464-7475.

[11] SOHAN M, SAI R T, REDDY CH V R. A review on YOLOv8 and its advancements [C]. International Conference on Data Intelligence and Cognitive Informatics, 2024: 529-545.

[12] HUANG X, WANG X X, LYU W Y, et al. PP-YOLOv2: A practical object detector [J]. ArXiv preprint arXiv:2104.10419, 2021.

[13] GE ZH, LIU S T, WANG F, et al. YOLOX: Exceeding YOLO series in 2021[J]. ArXiv preprint arXiv:2107.08430, 2021.

[14] XU SH L, WANG X X, LYU W Y, et al. PP-YOLOE: An evolved version of YOLO[J]. ArXiv preprint arXiv:2203.16250, 2022.

[15] LI CH Y, LI L L, GENG Y F, et al. YOLOv6 v3.0: A full-scale reloading [J]. ArXiv preprint arXiv: 2301.05586, 2023.

[16] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers [C]. European Conference on Computer Vision, 2020: 213-229.

[17] 李青云, 魏佳. 改进 RT-DETR 的密集行人检测算法[J]. 电子测量技术,2025,48(21):148-156.
LI Q Y, WEI J. Improvement of the dense pedestrian detection algorithm of RT DETR [J]. Electronic Measurement Technology, 2025,48(21):148-156.

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

[19] MENG D P, CHEN X K, FAN Z J, et al. Conditional detr for fast training convergence [C]. IEEE/CVF International Conference on Computer Vision(ICCV), 2021:3651-3660.

[20] WANG Y M, ZHANG X Y, YANG T, et al. Anchor DETR: Query design for transformer-based object detection[J]. ArXiv preprint arXiv:2109.07107, 2021.

[21] LIU SH L, LI F, ZHANG H, et al. DAB-DETR: Dynamic anchor boxes are better queries for detr[J]. ArXiv preprint arXiv:2201.12329, 2022.

[22] LI F, ZHANG H, LIU SH L, et al. DN-DETR: Accelerate detr training by introducing query denoising[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022: 13609-13617.

[23] CHEN Q, CHEN X K, WANG J, et al. Group DETR: Fast detr training with group-wise one-to-many assignment [C]. IEEE/CVF International Conference on Computer Vision (ICCV), 2023: 6633-6642.

[24] ZHANG H, LI F, LIU SH L, et al. DINO: DETR with improved denoising anchor boxes for end-to-end object detection [J]. ArXiv preprint arXiv: 2203.03605, 2022.

[25] ZHAO Y, LYU W Y, XU SH L et al. DETRs beat YOLOs on real-time object detection[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), 2024:16965-16974.

[26] QI Y L, HE Y T, QI X M, et al. Dynamic snake convolution based on topological geometric constraints for tubular structure segmentation[C]. IEEE/CVF International Conference on Computer Vision(ICCV), 2023:6070-6079.

[27] XIA ZH F, PAN X R, SONG SH J, et al. Vision transformer with deformable attention[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), 2022:4794-4803.

[28] DING X H, ZHANG X Y, HAN J G, et al. Diverse branch block: building a convolution as an Inception-like unit[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021: 10886-10895.

[29] LI K, WAN G, CHENG G, et al. Object detection in optical remote sensing images: A survey and a new benchmark[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2020, 159: 296-307.

[30] LONG Y, GONG Y P, XIAO ZH F, et al. Accurate object localization in remote sensing images based on convolutional neural networks[J]. IEEE Transactions on Geoscience and Remote Sensing, 2017, 55(5): 2486-2498.

作者简介

肖锋,博士,教授,博士生导师,主要研究方向为智能信息处理、模式识别和深度学习等。
E-mail:xffriends@163.com

杨文豪,硕士,主要研究方向为目标检测、图像处理。

张文娟(通信作者),博士,副教授,硕士生导师,主要研究方向为图像处理、计算机视觉及机器学习。
E-mail:zhangwenjuan@xatu.edu.cn

黄姝娟,博士,副教授,硕士生导师,主要研究方向为人工智能、多核计算和嵌入式等。