DOI:10.19651/j.cnki.emt.2519091

面向遥感图像的改进 RT-DETR 目标检测算法*

陈 辉[1] 王新蕾[1,2]
(1.南京信息工程大学电子与信息工程学院 南京 210044;2.无锡学院电子信息工程学院 无锡 214105)

摘 要:针对遥感图像目标检测中目标分布密集、背景复杂和小目标众多等问题导致检测效果不佳。本文提出一种基于 RT-DETR 的 RSD-DETR 遥感图像目标检测算法。首先,设计了轻量级多尺度特征提取 Faster-CGLU 模块,将门控机制和部分卷积融合,优化局部和全局特征信息的聚合,同时减少计算冗余。其次,结合级联分组注意力(CGA)构建 CGA-AIFI 模块,以在抑制非相关背景信息的同时关注关键特征区域,增强模型与目标特征的交互能力。最后,设计跨尺度动态特征融合(CS-DFFM)结构,通过动态尺度序列特征融合(DySSFF)模块和三重特征编码器(TFE)模块,对多尺度特征图进行尺寸对齐和动态融合,防止上下采样导致的小目标特征信息丢失,增强了网络多尺度特征融合能力。实验结果表明,在 SIMD 和 DOTA-v1.0 数据集上,所提算法在参数量较基线模型降低 22.11%的情况下,平均精度均值(mAP0.5)分别达到了 79.9%和 86.8%,较基线模型分别提高了 2.5%和 1.7%,模型实时性也得到了提高。检测效果优于其他经典模型,具有卓越的性能。

关键词:遥感图像;RT-DETR;小目标检测;特征融合;多尺度
中图分类号:TP391.4;TN914 **文献标识码**:A **国家标准学科分类代码**:520.6040

Improved RT-DETR object detection algorithm for remote sensing images

Chen Hui[1] Wang Xinlei[1,2]
(1. School of Electronic and Information Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China; 2. School of Electronic Information Engineering, Wuxi University, Wuxi 214105, China)

Abstract: Dense target distribution, complex backgrounds, and a large number of small objects often lead to suboptimal detection performance in remote sensing image object detection. To address these challenges, this paper proposes RSD-DETR, a remote sensing object detection algorithm based on RT-DETR. First, a lightweight multi-scale feature extraction module, Faster-CGLU, is designed by integrating a gating mechanism with partial convolution, which optimizes the aggregation of local and global feature information while reducing computational redundancy. Second, a CGA-AIFI module is constructed using cascaded group attention (CGA), which focuses on critical feature regions while suppressing irrelevant background information, thereby enhancing the interaction between the model and object features. Finally, a cross-scale dynamic feature fusion module (CS-DFFM) is designed, which performs spatial alignment and dynamic fusion of multi-scale feature maps through the dynamic scale-sequence feature fusion (DySSFF) module and the triple feature encoder (TFE) module. This effectively mitigates the loss of small object features caused by upsampling and downsampling, and enhances the network′s multi-scale feature fusion capability. Experimental results show that on the SIMD and DOTA-v1.0 datasets, the proposed algorithm reduces the number of parameters by 22.11% compared with the baseline model, and the mean average precision (mAP0.5) reaches 79.9% and 86.8% respectively, which are 2.5% and 1.7% higher than those of the baseline model. The real-time performance of the model is also improved. The detection effect is better than other classic models, and it has excellent performance.

Keywords: remote sensing images; RT-DETR; small object detection; feature fusion; multi-scale

0 引 言

随着技术的不断进步和智慧城市的发展建设,遥感图像的目标检测在应急救援、城市规划和资源勘探等领域应用广泛[1]。遥感图像目标检测的目的是精准识别特定地理空间目标的位置和类别,例如地面建筑和交通工具。与普

收稿日期:2025-06-12
*基金项目:国家自然科学基金(62204172)资助

通自然光学图像相比,深度学习技术在遥感图像上的应用还存在挑战,遥感图像中的目标检测存在背景混杂、目标分布密集和小目标众多等问题,导致检测精度较低[2]。以至现有检测技术往往难以有效应用,容易产生漏检和误检。因此,有必要进一步增强深度学习技术在遥感图像分析中的检测性能,以推动遥感技术的自动化、智能化和高效化发展。

近年来,基于深度学习的目标检测算法可分为两阶段目标检测和一阶段目标检测两大类,两阶段目标检测算法以 R-CNN[3]和 Faster R-CNN[4]为代表,通过一个显式的、两步骤的方式利用区域生成技术对潜在的目标区域生成检测候选框,然后通过特征提取再进行分类和回归操作。一阶段目标检测算法以 SSD[5]和 YOLO[6]为经典算法,直接对输入图像预测感兴趣目标的类别概率和位置坐标。以上算法都是基于卷积神经网络(convolutional neural network, CNN)改进完成的,尽管卷积神经网络在提取局部特征方面表现优异,但捕捉全局图像信息却面临挑战。为克服 CNN 在捕捉全局图像信息时易丢失空间特征信息的缺陷,源自自然语言处理领域的 Transformer 模型[7]中的多头自注意力机制擅长捕捉长距离依赖关系,可以更有效地利用上下文信息,但可能忽略局部特征。因此,一些研究尝试将 CNN 与 Transformer 结合,以兼顾局部特征、全局表示和长距离依赖性。

DETR[8]借鉴了 Transformer 的编码器结构和基于二分图的匹配策略,将目标检测重构为直接集合预测问题,摒弃了非极大值抑制(non-maximum suppression, NMS)后处理策略,简化了检测流程。但 DETR 的高计算成本导致训练收敛慢,阻碍了在实际场景中的应用。为解决这一问题,Deformable-DETR[9]使用局部采样的可变性注意力取代 DETR 模型的全局注意力机制,虽提高了检测效率但模型整体的参数量仍比较大。DEIM[10]通过引入密集的一对一匹配策略和匹配感知损失函数使模型更加轻量,加快模型训练效率,但模型特征层级间的信息交互不够充分。RT-DETR[11]凭借其混合编码器设计和交并比(intersection over union, IOU)感知查询选择策略实现了实时端到端的目标检测,解决了 NMS 导致的推理延迟。与其他 DETR 系列模型相比,具有相对较少的参数量和较低的计算成本。并且在速度和精度方面都优于同等尺寸的基于 YOLO 的检测模型[12]。尽管如此,RT-DETR 在遥感图像目标检测方面性能还需进一步提升。

近年来,针对遥感图像中的目标分布密集、背景复杂,以及遮挡目标和小目标难以准确检测的问题,一些研究人员设计了较好的检测算法和技术。Zhang 等[13]设计 FFCA-YOLO 算法,通过特征增强模块(feature enhancement module, FEM)、特征融合模块(feature fusion module, FFM)以及空间上下文感知模块(spatial context aware module, SCAM)等改进,增强了网络对遥感图像小目标检测的敏感性。闫钧华等[14]提出一种遥感图像弱小目标检测算法 CC-YOLO,通过重构多层级特征提取和融合模块,结合位置注意力机制,提升目标检测精度。Zhou 等[15]提出了一种基于 Transformer 的相关学习框架,通过改进特征提取和融合方法,有效提升了在遥感图像中对密集目标的检测能力。Kong 等[16]提出一种高效遥感图像目标检测算法 Drone-DETR,设计高效的小物体检测网络(effective small object detection network, ESDNet),并嵌入双路径特征融合注意力在颈部中,显著提高了小目标的检测精度。姜贸翔等[17]设计 EMRT-DETR 模型,通过替换轻量级的特征提取主干和引入 P2 级检测层,成功保持精度优势前提下,降低了模型的参数量。Wang 等[18]提出 Ship-DETR 算法,通过引入高低频注意力机制来增强对高低频特征的提取能力,并引入双向特征金字塔网络来优化跨尺度特征融合,提高了对目标的感知能力。Song 等[19]设计 DAF-DETR 模型,通过优化主干特征提取网络和引入动态感知模块,有效提升了对复杂场景中遥感目标的检测能力。

上述算法在针对检测目标小以及目标尺寸多样的遥感图像时,虽然在一定程度上可以提升各类目标的检测精度,但仍然存在漏检和误检的情况,还有提升的空间。首先,部分方法在特征提取过程中对图像进行了多次下采样,导致小目标特征在浅层阶段即被压缩甚至丢失。其次,遥感图像中同一类别目标存在显著尺度差异,当前模型在不同尺度目标之间的特征适配与交互能力不足。此外,模型在融合不同层级的特征图时,缺乏有效的跨尺度建模能力,密集重叠目标容易被忽略。另一方面,部分方法引入了复杂的注意力结构或多分支融合策略,在提升检测性能的同时也带来了额外的计算负担。同时,这些方法多针对特定数据集进行结构优化,在不同遥感场景或其他下游任务中的泛化能力仍显不足。

基于上述问题,本文以 RT-DETR 模型为基础,提出一种面向遥感图像目标的 RSD-DETR 检测算法,旨在提高遥感图像目标检测的精度和速度,同时降低模型的参数量。本研究的主要贡献为:

1)设计了一种轻量级多尺度特征提取模块(faster convolutional gated linear unit, Faster-CGLU),通过卷积门控线性单元(convolutional gated linear unit, CGLU)和部分卷积(partial convolution, PConv),优化多通道信息利用,显著降低了模型计算量,同时精度不受影响,也增强模型在复杂场景下的抗干扰能力。

2)提出(cascaded group attention-based intra-scale feature interaction, CGA-AIFI)模块,引入级联分组注意力机制提高目标的注意多样性,增强模型与目标特征的交互能力,提升模型对不同尺度目标的检测和识别能力。CGA-AIFI 模块不仅提高了模型对局部细微特征的捕获能力,还加强了目标与背景的区分,使得模型在复杂环境下的鲁棒

性得到增强。

3)设计了跨尺度动态特征融合(cross-scale dynamic feature fusion module, CS-DFFM)结构,其中动态尺度序列特征融合(dynamic scale sequence feature fusion, DySSFF)模块通过动态上采样(dynamic upsampling, DySample)丰富不同尺度图像特征的语义信息,增强了模型多尺度特征融合能力,有效避免模型颈部在上下采样过程中小目标特征信息的丢失问题。同时,三重特征编码器(triple feature encoder, TFE)模块对大、中、小尺寸特征进行了拆分、调整和融合,以增强详细特征信息的表达,从而准确检测密集重叠的小目标。

1　RT-DETR 模型

RT-DETR 的架构包括 3 个主要部分:主干网络(Backbone)、高效混合编码器(efficient hybrid encoder)和带有辅助检测头的解码器(transformer decoder)。RT-DETR 主干网络采用了 CNN 结构,比如经典的 ResNet 系列或百度的 HGNet。在颈部网络中,高效的混合编码器通过多尺度内特征交互(attention-based intrascale feature interaction,AIFI)模块对主干最后一层特征解耦处理。随后利用跨尺度特征融合模块(cross-scale feature-fusion module,CCFM)将不同尺度特征进行融合处理,最终生成图像特征序列。头部网络采用了集成辅助预测头的 Transformer 解码器架构,首先通过 IoU 感知查询选择(IoU-aware query selection)技术,从编码器输出序列中选择一定数量的图像特征作为初始对象查询,然后通过迭代优化来生成预测框和置信度分数。RT-DETR 有多个型号,包括 R18、R34、R50、R50m、R101、L 和 X。针对遥感航拍图像目标检测的复杂多变的环境以及检测精度与速度平衡的问题,本文选用 R18 为基准模型进行优化,提出遥感图像目标检测 RSD-DETR 模型。

2　RSD-DETR 算法

2.1　RSD-DETR 的整体框架

针对光学遥感航拍图像背景复杂、小目标众多,以及目标分布密集混乱导致的检测精度低,同时考虑到模型参数较大的难题,本文在 RT-DETR 的基础上,提出面向遥感图像 RSD-DETR 目标检测算法,该模型的整体结构设计如图 1 所示。首先,遥感图像进入主干网络,通过设计的轻量级多尺度特征提取 Faster-CGLU 模块对其进行特征提取,最后 4 个层级生成多尺度特征图:P2、P3、P4 和 P5。然后,这些不同尺度的特征图通过跨尺度动态特征融合 CS-DFFM 结构进行特征的融合。P3、P4 和 P5 级特征图会先经过设计的动态尺度序列特征融合 DySSFF 模块进行特征融合,接着,P5 级特征图经过提出的 CGA-AIFI 模块进行内特征交互后再进一步和其他级别的特征图在 CS-DFFM 结构中进行特征融合。最后,解码器先通过 IoU 感知查询选择模块,从编码器输出序列中选择一定数量的图像特征作为初始对象查询,再通过迭代优化来生成预测框和置信度分数。

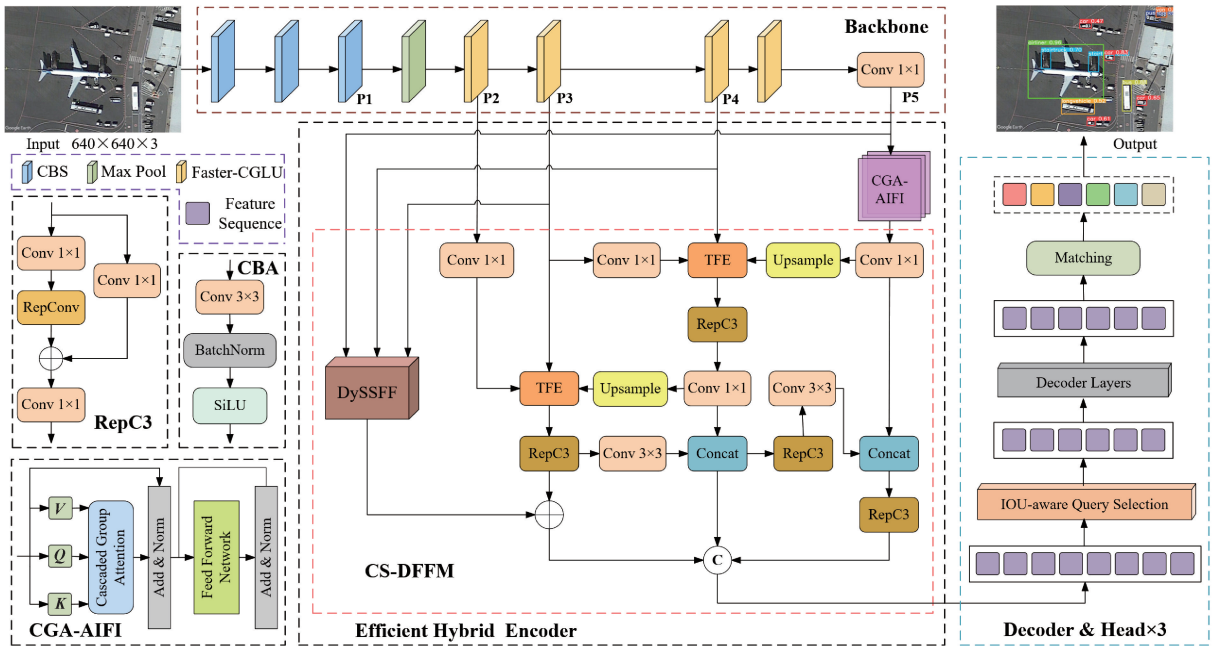


图 1　RSD-DETR 模型的整体结构

Fig. 1　Overall structure of RSD-DETR model

综上所述,设计轻量级多尺度特征提取模块 Faster-CGLU,通过 3 层残差可以显著减少模型的计算冗余,并且保留了小目标检测所需的精细特征。在编码器中提出的 CGA-AIFI 模块,通过引入级联分组注意力,增强了目标与

背景的区分度,从而提升模型对不同尺度和空间分布目标的检测能力。在此基础上,跨尺度动态特征融合 CSDFFM 结构,通过高效融合浅层与深层特征,优化了 Faster-CGLU 提取的不同尺度特征信息和 CGA-AIFI 聚合的细节特征,增强多级特征图的特征融合能力,避免上下采样过程中小目标特征信息的丢失问题。这些模块的协同作用,使得 RSD-DETR 模型可在复杂场景下更精确地识别遥感图像中的目标。以下各节将详细介绍模型的具体设计细节。

2.2　Faster-CGLU 模块

遥感图像中存在大量的小目标和密集重叠目标导致精准检测难度大,同时对于遥感图像的拍摄和采集设备的平台部署的轻量化要求,因此需要进一步增强目标检测模型的特征提取能力。而 RT-DETR 模型的主干特征提取模块存在大量的计算冗余和无用的特征映射在网络中重复,为平衡模型的检测精度和参数量的矛盾,本文提出了 Faster-CGLU 模块,融合 FasterNet 中的部分卷积(PConv)[20] 和卷积门控线性单元(CGLU)[21]。Faster-CGLU 模块的结构如图 2 所示。

为了提取比普通残差结构更加有效的特征信息,Faster-CGLU 模块采用了一种增强的多层残差结构,在多级子模块中利用残差思想保持梯度循环。其首先将输入特

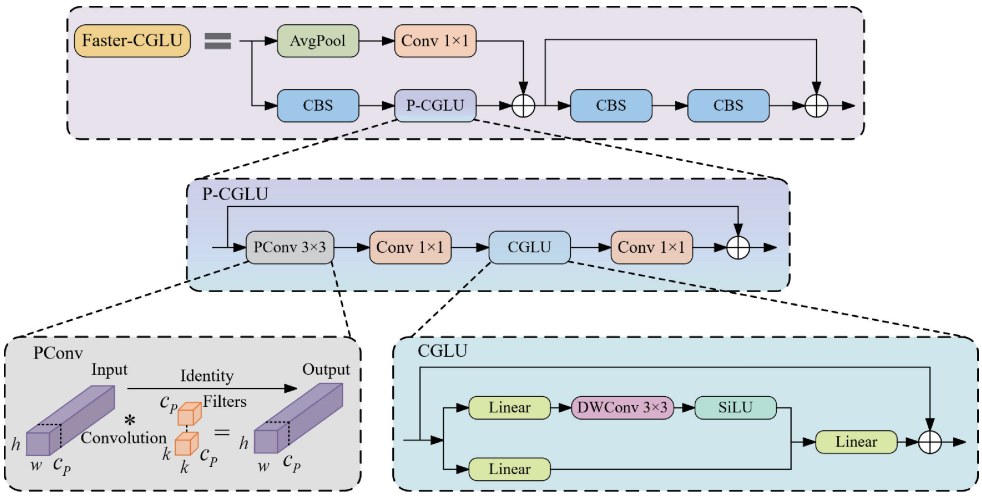


图 2　Faster-CGLU 模块结构

Fig. 2　Structure of Faster-CGLU module

征图分别通过平均池化和(partial-convolutional gated linear unit, P-CGLU)模块提取全局信息和局部信息,然后将全局和局部信息整合拼接之后输入残差网络进行下一步的特征提取。避免了目标梯度的消失,从而增强了骨干网络的特征提取能力。Faster-CGLU 内部的 P-CGLU 模块本质上也是一种残差结构,包含 PConv 和 CGLU 两个过程。PConv 在运算时会对内存访问,为了减少空间复杂度,选择特定通道而不是全部通道参与运算。考虑到底层通道包含更多的细节信息而高层通道包含更多的语义信息,模块取第一个或最后一个连续的 c_p 通道,即输入特征图中通道维度的前 c_p 个通道或最后 c_p 个通道,作为整个特征映射的代表进行计算。对部分输入通道进行卷积操作,而其他输入通道保持不变。这在一定程度上大大降低了计算复杂度和内存访问次数。PConv 的 FLOPs 和内存访问量如式(1)和(2)所示。

$$h \times w \times k^2 \times c_p^2 \tag{1}$$

$$h \times w \times 2c_p + k^2 \times c_p^2 \approx h \times w \times 2c_p \tag{2}$$

当选择通道数 c_p 作为输入通道通道数 C 的通道数的 1/4 时,计算复杂度降低到原来普通卷积的 1/16,内存访问量减少到原来普通卷积的 1/4。

在 PConv 卷积之后加入 1×1 卷积来调整特征的通道数,并引入卷积门控线性单元(CGLU)使网络的数据分布更加稳定,增强模型捕获细粒度信息的能力。CGLU 本质上也是一种残差结构,其动态门控机制保证了每个 Token 的门控信号来源于自身,避免了全局平均池化可能引入的信号共享问题,并且其随机二选一的路径选择,减少了模型参数量,加速了运算速率。左侧深度卷积(DWConv)[22] 支路使得 CGLU 可以更有效地捕获局部特征,增强模型处理细微特征的能力。而右侧线性层支路可以保留更多的全局信息。CGLU 这种动态门控机制,使模型在保留卷积局部特征提取优势的同时,也利用线性变换实现了对全局信息的有效聚合。

总之,Faster-CGLU 模块可以更有效地利用多通道信息、降低参数数量和计算开销、保持良好的梯度循环,以及增强空间特征的提取能力。这是模型的轻量化和检测精度的提升中至关重要的一环,同时提高了对遮挡和形变目标区域的敏感性。

2.3　CGA-AIFI 模块

在 RT-DETR 的高效编码器设计中,尺度内特征交互 AIFI 模块采用 transformer 中的多头自注意力机制来挖掘

P5 级特征图的深度相关性，但这种机制增加了计算复杂性和模型参数。同时，遥感图像中目标混杂，尺度多样。为此，本文提出 CGA-AIFI 模块，引入级联分组注意力[23]来增强目标注意的多样性，关注相关的特征区域，增强模型对于多尺度目标信息的敏感性。CGA-AIFI 通过切片和级联操作，使其每个注意力头可以充分捕捉本组输入的局部信息，同时可以获得前注意力头的所有信息。这不仅提高了信息的容量还恢复了特征的多样性，并减少了计算冗余。

CGA-AIFI 模块的结构如图 3 所示。在 CGA-AIFI 中，输入图像被分割成组，为每个子空间提供不同的输入，不同的子空间学习不同的切片特征，使模型可以学习更细粒度的特征表示，也有效分解了跨头部的注意力计算。与标准的多头自注意力不同，这里的分割操作是在计算 $\boldsymbol{V}$、$\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 之前，将注意力头划分为多个子注意力头。每组的输入序列首先经过线性映射以生成值（$\boldsymbol{V}$）、查询（$\boldsymbol{Q}$）和键（$\boldsymbol{K}$），然后 CGA-AIFI 应用分组注意力，并使用 $\boldsymbol{V}$、$\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 计算每个子空间中的注意力权重，以产生分组的注意力输出。

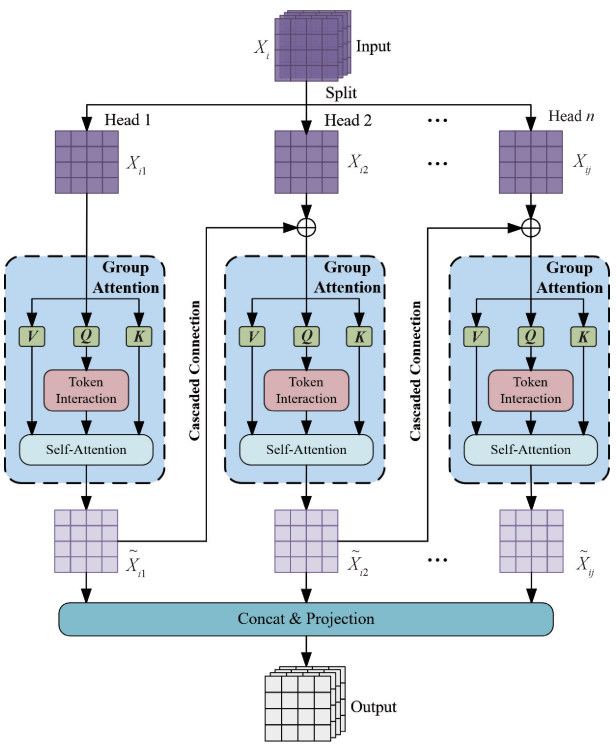


图 3　CGA-AIFI 模块结构

Fig. 3　CGA-AIFI module structure

此外，通过对不同组的输出进行级联，前头提取的物体局部特征信息，会在后续头的计算中被整合。随着头的逐层叠加，特征表示不断细化，既保留了局部特征的细节，又逐步融入了全局特征的上下文信息，进一步提升模型的容量。最后，将级联输出进一步拼接以产生 CGA-AIFI 的最终输出。这种渐进聚焦过程增强了特征在各个层次上的细化，从而提高了模型感知和准确描述特征的能力。具体计算过程如式(3)～(5)所示。

$$\tilde{X}_{ij} = Attn(X_{ij}P_{ij}^{V}, X_{ij}P_{ij}^{Q}, X_{ij}P_{ij}^{K}) \tag{3}$$

$$X'_{ij} = X_{ij} + \tilde{X}_{i(j-1)},\ 1 < j \leqslant n \tag{4}$$

$$\tilde{X}_{i+1} = Concat[\tilde{X}_{ij}]_{j=1:n} W_i^p \tag{5}$$

其中，X_{ij} 是输入特征 X_i 的第 j 次切片投影，$\tilde{X}_{ij}$ 为 X_{ij} 经过第 j 层注意力头后的输出，P_{ij}^{V}、P_{ij}^{Q}、P_{ij}^{K} 分别表示输入特征映射投影不同层的 V、Q 和 K，X'_{ij} 表示当前层输入与上一层输出的求和结果，W_i^p 是一个线性层，其将连接的输出特征重新投影回与输入特征相同的维度，n 是总头数。

2.4　跨尺度动态特征融合 CS-DFFM 结构

在多尺度特征融合技术不断发展的过程中，Neck 部分的设计起着至关重要的作用，从最初的自上而下的特征金字塔融合方案，到适应多样化应用场景的各种改进，颈部的设计经历了一个渐进的演变过程[24]。RT-DETR 中的 CCFF 模块通过跨尺度特征融合方式提高了传统 PAFPN[25]的融合效率。然而，遥感图像中存在众多小目标，CCFF 模块仅对主干网络提取到的 P3、P4 和 P5 级别的特征图进行融合，并不能充分有效地利用所有金字塔特征图之间的相关性，容易造成小目标的漏检和误检。对此，本文设计了 CS-DFFM 结构，核心组件为动态尺度序列特征融合 DySSFF 模块和三重特征编码器 TFE 模块[26]，如图 1 所示。将 P2 级特征图也融入到颈部特征融合网络中，CS-DFFM 通过在更早的阶段获取高分辨率的小目标特征信息，并充分融合不同尺寸的特征图，模型能够更好地感知小目标的细节信息，也避免了不同尺度目标因在颈部中的上下采样导致特征信息丢失的问题。

1）DySSFF 模块

RT-DETR 中的 CCFF 模块采用简单的拼接或加和来融合遥感图像中不同尺度的特征信息时，忽略了不同尺度特征表示的差异。为了更有效地整合深层特征图的高维信息与浅层特征图的细节信息，受动态学习采样思想[27]的启发，提出了动态尺度序列特征融合 DySSFF 模块，结构如图 4 所示。该模块放弃了传统的静态融合策略，转而采用了动态融合方法。可以根据当前输入数据的特征，自适应调整不同尺度特征之间的融合权值和方式，从而实现更精细有效的特征融合。

DySSFF 模块充分利用图像的尺度空间特性，鉴于高分辨率特征图 P3 级别包含了各尺寸目标检测所需的重要信息，DySSFF 对 P4 和 P5 级特征图利用 1×1 卷积保持尺度不变改变其通道数，并通过 DySample 动态上采样模块改变特征图尺寸以提取关键信息。再将三路相同分辨率的特征图进行拼接堆叠，并利用三维卷积提取它们的尺度序列特征。DySSFF 有效整合了不同特征图中的信息，增强了特征融合能力。

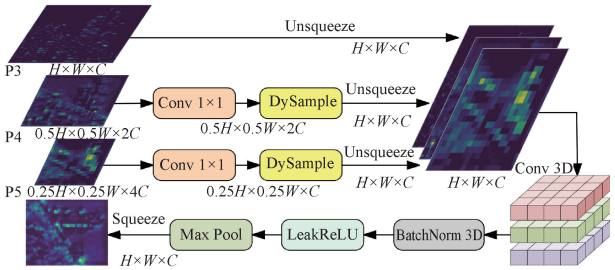


图 4　DySSFF 模块结构

Fig. 4　Structure of DySSFF module

DySSFF 中的动态上采样 DySample 模块结构如图 5 所示,大小为 $H\times W\times C$ 的输入特征图 x 先通过动态点采样生成器生成点采样集 S,以确定对其输入的特征图采样区域获取采样点坐标,动态点采样集 S 大小为 $sH\times sW\times 2g$,其中 $2g$ 为 x 轴与 y 轴的坐标。然后利用 Grid sample 线性插值函数来使得点采样集对输入特征图重新采样,生成新的大小为 $sH\times sW\times C$ 的特征图 X,计算过程如式(6)所示。

$$X = Grid\ sample(x,S) \tag{6}$$

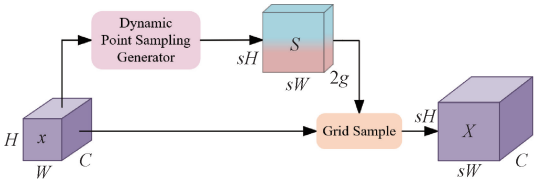


图 5　DySample 结构

Fig. 5　DySample structure

动态点采样集生成过程如图 6 所示,特征图 x 通过两个并行线性层 Linear 输出两个尺寸为 $H\times W\times 2gs^2$ 的偏移量 O_1 和 O_2,将 O_1 的采样因子设置为动态因子 0.5σ,σ 表示 sigmoid 激活函数。再通过像素重组操作将偏移量 O_1 和 O_2 相结合,重塑成 $sH\times sW\times 2g$ 的偏移量 O,其中偏移范围由 0.5 和激活函数 σ 决定。最后,通过结合偏移量 O 与原始采样网格 g 生成采样集 S。具体表达如式(7)和(8)所示。

$$O = 0.5sigmoid(Linear_1(x))\cdot Linear_2(x) \tag{7}$$

$$S = O + g \tag{8}$$

与传统的静态插值上采样方法不同,DySample 动态上采样根据输入特征内容自适应地生成采样点的分布,提高了特征图分辨率恢复和扩展的准确性,促进了更稳定的特征融合。这种根据图像内容动态调整采样策略的方式,在复杂场景下尤为有利,提高了模型对复杂地形条件的适应性,同时也避免了上采样过程中可能削弱小目标特征信息的问题。

2)TFE 模块

RT-DETR 的 CCFM 模块未能充分考虑大尺寸特征层中的细节信息,针对遥感图像中密集重叠的小目标物体,需要扩展图像,以便准确地检测这些目标。为此,本文引入三重特征编码器 TFE 模块。TFE 模块结构如图 7 所示。

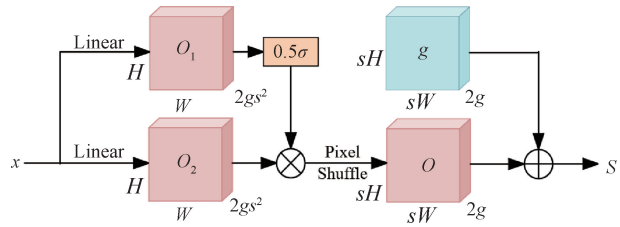


图 6　动态点采样集生成过程

Fig. 6　Dynamic point sampling set generation process

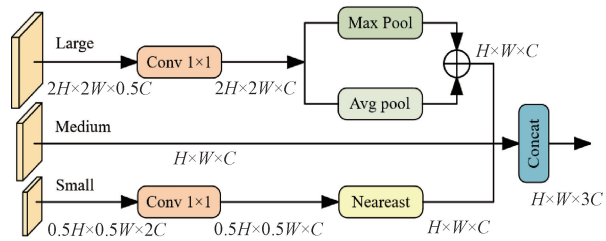


图 7　TFE 模块结构

Fig. 7　TFE module structure

TFE 模块能够对大、中、小尺寸特征图进行调整和融合,以增强详细特征信息。首先通过 1×1 卷积调整大和小尺寸特征层的通道数,以使其与中尺度特征保持一致。然后利用最大池化和平均池化的组合对大尺寸特征图进行下采样,有助于保留遥感图像中不同物体的高分辨率特征和语义信息的多样性;对小尺寸特征图采用最近邻插值法进行上采样,既能保持低分辨率图像局部特征的丰富性,又能防止小目标特征信息的丢失。最后,将 3 个相同尺寸的大、中和小特征图在通道维度上连接融合。计算如式(9)所示。

$$F_{TFE} = Concat(F_l,F_m,F_s) \tag{9}$$

其中,F_{TFE} 表示 TFE 模块输出的特征图。F_l、F_m、F_s 分别表示大、中和小尺寸的特征图。F_{TFE} 由 F_l、F_m、F_s 拼接得到。F_{TFE} 具有与 F_m 相同的分辨率,并且通道数是 F_m 的 3 倍。

3　实验与结果分析

3.1　实验环境与实验参数配置

实验采用操作系统为 Windows11(专业版),处理器为 Intel core i9-13900K、64 G 内存和 GPU 为 RTX4090(24 G)。实验框架环境是 Pytorch 1.31.1,python 解释器版本 3.8.18,CUDA 版本为 11.7。

为保证实验的公平性,本文均在相同实验环境下训练,所有消融实验和比较实验中的各个模型训练过程均未使用任何预训练权重。实验超参数设置为:总训练轮数 epoch 设置为 300,初始学习率设置为 0.000 1,一次输入模

型的样本数 batch size 设置为 8，进程数为 4，输入图像尺寸为 640 pixel×640 pixel。模型训练时使用 AdamW 优化器对模型学习率进行调整。模型使用权重衰减策略防止过拟合，权重衰减值设为 0.000 1。

3.2　实验数据集

考虑到遥感图像中物体的种类和小目标数量要众多，本文在 SIMD[28] 和 DOTA-v1.0[29] 两个公开数据集上进行实验，以对 RSD-DETR 进行性能评估。本文实验结果表格中的加粗数据表示同一指标或同一目标类别下的最优数据。

SIMD 数据集是由巴基斯坦国立科学技术大学提出的在 2020 年发布的公开遥感图像目标检测数据集，包含 5 000 幅尺寸为 1 024 pixel×768 pixel 遥感图像和 45 096 个实例。数据集中的实例分为 15 类：汽车、卡车、厢货、长车、公共汽车、客机、螺旋桨飞机、教练机、包机、战斗机、其他、楼梯车、拖车、直升机和船。本文将其按 7∶2∶1 划分训练集、验证集和测试集。该数据集中存在大量的密集遮挡的小目标，场景从密集到稀疏及目标物体类别多样，很好的满足了本文对遥感图像目标的验证需求。

DOTA-v1.0 数据集是由武汉大学提出的在 2017 年发布的一个用于航空影像中目标检测的大规模数据集，包含 2 806 幅航拍图像，每幅图像像素尺寸在 800 pixel×800 pixel 到 4 000 pixel×4 000 pixel 的范围内，不同尺度和形状的实例 188 282 个。数据集中的实例包括飞机、船舶、储罐、港口、桥梁、小型车辆、直升机等 15 类。为了在特征提取下采样时小目标实例不被漏检，采用图像分割的方式降低图像的分辨率，以 824 的滑窗步长，切分为 21 046 张分辨率为 1 024 pixel×1 024 pixel 的图像，将训练集、验证集和测试集按 7∶2∶1 划分。该数据集中的图像类别多样、背景复杂、目标尺度变化大及小目标物体占比过半，满足了本文所提模型的验证条件。

3.3　评价指标

为了评估 RSD-DETR 的性能。本文实验以准确率（Precision，P）、召回率（Recall，R）、平均精度（average precision，AP）、平均精度均值（mean average precision，mAP）、参数量（parameters，Params）、每秒检测帧数（frame per second，FPS）作为评价指标。其中，本文在 mAP 中设置了两个阈值，分别是 mAP0.5 和 mAP0.5∶0.95，表示 IoU 为不同值时 mAP 的具体数值。Params 和 FPS 指标用来衡量模型的轻量化和推理速度。相关计算公式为：

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

$$AP = \int_0^1 P(R)\mathrm{d}R \tag{12}$$

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{13}$$

3.4　消融实验

为评估每个模块对模型的贡献，在 SIMD 数据集上进行了消融实验，对比基线模型 RT-DETR。在基线模型基础上，依次添加不同改进模块及其组合，量化评估每组实验对模型性能的影响。实验结果如表 1 所示，其中“√”表示使用了该模块，空白则表示未使用该模块。

表 1　消融实验结果

Table 1　Results of ablation experiment

序号	RT-DETR	Faster-CGLU	CGA-AIFI	CS-DFFM	P/%	R/%	Params/M	FPS	mAP0.5/%	mAP0.5∶0.95/%
1	√				77.9	76.2	19.9	69.6	77.4	62.0
2	√	√			78.1	77.1	15.4	78.5	78.3	62.4
3	√		√		78.5	76.9	19.7	72.4	78.5	62.8
4	√			√	81.3	78.0	20.3	68.8	79.4	63.1
5	√	√	√		81.1	77.8	**15.1**	**79.8**	79.1	62.9
6	√	√	√	√	**82.7**	**78.5**	15.5	78.3	**79.9**	**63.6**

在同等实验条件下，从表 1 可以看出，实验 1 采用基线模型 RT-DETR 进行遥感图像目标检测时，各目标的准确率和召回率的平均值分别为 77.9% 和 76.2%，mAP0.5 为 77.4%，参数量高达 19.9 M。实验 2 首先引入 Faster-CGLU 模块，模型参数量较基线模型锐减了 4.5 M，FPS 提高了 8.9，同时检测精度也得到了小幅度的提高，有效增强了主干网络的特征提取能力，也凸显出 Faster-CGLU 的轻量化和快速性优势。实验 3 使用 CGA-AIFI 模块后，准确率和检测精度 mAP0.5 较基线模型分别提高了 0.6% 和 1.1%，参数量减少了 0.2 M，说明 CGA-AIFI 的分组和级联学习策略有效提高了模型与多尺度目标特征的交互能力。实验 4 引入设计的 CS-DFFM 结构替换掉基线模型的 CCFF，结果显示，参数量和 FPS 较实验 1 变化不大，表明引入 CS-DFFM 模块并没有过多增加模型的复杂度。检测精度 mAP0.5 和 mAP0.5∶0.95 较基线模型分别提高了 2.0% 和 1.1%，充分表明 CS-DFFM 模块增强了模型对多尺度特征的融合能力，避免了小目标特征信息的丢失。实验 5 共同融入 Faster-

CGLU 和 CGA-AIFI 模块后,参数量减少到 15.1 M,说明两个模块对降低模型的参数量有良好的协同性;同时,准确率和召回率较基线模型分别提高了 3.2%和 1.6%,两个模块对模型性能可达到增益叠加的结果,提高了对遥感图像目标的注意多样性。实验 6 为本文的 RSD-DETR 模型,召回率和 mAP0.5 较实验 5 分别提高了 0.7%和 0.8%,准确率和 mAP0.5 较基线模型分别上涨了 4.8%和 2.5%;相比于实验 5,FPS 虽比最优的 79.8 略低,但相比于实验 1 的基线模型 FPS 提升了 8.7;参数量与最优数据仅差了 0.4 M,却比实验 1 减少了 4.4 M,其他指标均达到最优。进一步表明了 CS-DFFM 模块在提升模型性能与模型复杂度之间达到良好平衡。消融实验总体证明,三方面改进策略可以达到良好的协同效应,最终可达到增益叠加的效果。

此外,为评估 RSD-DETR 中每个模块对每类检测目标对象的影响,本文在 SIMD 数据集上挑选汽车(C)、卡车(T)、厢货(V)、长车(LV)、公共汽车(B)、螺旋桨飞机(PA)、教练机(TA)、战斗机(FA)、楼梯车(ST)、拖车(PT)、直升机(H)和船舶(S)12 类不同尺度的经典遥感目标进行验证,结果如表 2 所示。实验 2 中 Faster-CGLU 有效地利用多通道信息,使公共汽车(B)的检测精度达到最优。实验 3 引入 CGA-AIFI 主要对具有较大感受野的高层特征进行依赖建模,螺旋桨飞机(PA)的精度提高了 4.8%。实验 4 得益于 DySSFF 模块避免了小目标特征信息的丢失,使直升机(H)目标的精度提高了 19.2%。整体来看,随着 3 个改进点的融入,除汽车(C)、公共汽车(B)、螺旋桨飞机(PA)和直升机(H)4 类目标的检测精度和最优精度几乎接近,其他各类别目标的检测精度均达到了最优。其中,基线模型中检测精度最低的楼梯车(ST)和拖车(PT)两类易遮挡目标,RSD-DETR 对其分别提高了 7.8%和 8.6%。证明了本文 3 个改进点可以良好的协同作用,达到了增益叠加的效果。

表 2　RSD-DETR 各模块对不同检测目标的精度评估

Table 2　Precision evaluation of RSD-DETR modules for different detection objects

序号	RT-DETR	Faster-CGLU	CGA-AIFI	CS-DFFM	AP/%											
					C	T	V	LV	B	PA	TA	FA	ST	PT	H	S
1	√				89.4	77.1	78.4	83.3	89.4	71.9	94.1	98.4	50.3	40.2	64.8	95.7
2	√	√			90.2	77.3	77.6	83.2	**92.8**	75.8	93.4	99.2	53.1	44.5	65.9	96.2
3	√		√		90.1	77.6	78.5	83.4	91.1	**76.7**	93.5	98.7	53.6	43.2	77.3	96.4
4	√			√	90.3	78.1	79.1	83.4	92.6	76.4	94.3	99.5	56.4	46.4	**84.0**	96.3
5	√	√	√		**90.6**	78.6	79.2	83.5	92.5	76.2	94.7	98.8	55.7	47.6	78.3	96.3
6	√	√	√	√	90.5	**78.8**	**80.4**	**83.7**	92.7	76.5	**94.9**	**99.6**	**58.1**	**48.8**	83.9	**96.5**

3.5　对比实验

1)不同模型对比实验

为了进一步验证 RSD-DETR 在遥感图像目标检测任务中的优越性,在 SIMD 和 DOTA-v1.0 数据集上,分别考虑模型大小和检测性能,将本文提出的 RSD-DETR 与经典模型进行对比,选择的算法包括 Faster-RCNN、SSD、EfficientDet[30]、YOLOv3[31]、YOLOv5、YOLOv8、YOLOv9[32]、YOLOv10[33]、YOLOv11[34]、YOLOv12[35]、文献[19]、Deformable-DETR、DEIM 和文献[36]。

实验结果如表 3 所示。从中可以看出 Faster-RCNN、SSD 和 EfficientDet 算法精度表现较差,主要是因为遥感图像中小目标像素低,特征提取不足,目标检测定位分类困难。YOLOv3 模型较 SSD 和 EfficientDet 检测精度有显著的提高,但是高额的参数量限制了在实际场景的应用。被广泛应用的 YOLOv5 模型在两个数据集中对应的检测精度也不是很高,分析原因是 YOLOv5 在训练过程中使用了 anchor-free 的方式,在遥感小目标检测方面还需进一步优化。检测精度较高的 YOLOv8m 和 YOLOv8l 模型的表现相对较好,但其参数量高于本文所提模型,且检测精度低于本文的 RSD-DETR 模型,缺乏对形变目标的建模与泛化能力。与采用了梯度流更加丰富的特征提取和融合结构的 YOLOv9c 和 YOLOv10m 相比,在 DOTA-v1.0 数据集上,RSD-DETR 的 mAP0.5 分别比其高了 8.2%和 13.1%。YOLOv11m 和 YOLOv12m 两个 SOTA 模型在 SIMD 数据集上的检测精度分别为 78.8%和 77.9%,在 DOTA-v1.0 数据集上的检测精度分别为 82.1%和 86.2%;检测性能还是略低于本文模型,且模型参数量高于 RSD-DETR。文献[36]是改进 YOLOv8 模型,虽参数量略低于本文模型,但在 DOTAv1.0 数据集上仅取得了 74.6%的 mAP0.5,比本文的 RSD-DETR 低了 12.2%,其并没有很好的平衡模型参数量与检测精度。

Deformable-DETR 引入可变形注意力增强了特征建模能力,但 40 M 的参数量远高于本文模型,同时在两个数据集上的检测精度较低。DEIM 模型是采用 transformer 架构的实时检测 SOTA 模型,RSD-DETR 的参数量和检测精度均优于 DEIM,在两个数据集上 mAP0.5 分别提高了 0.8%和 1.5%。RT-DETR(r18)和 RT-DETR(r50)的检测精度较高,但因内部的多头自注意力,参数量远高于

表 3　不同模型在 SIMD 和 DOTA-v1.0 数据集上的性能比较

Table 3　Performance comparison of different model on SIMD and DOTA-v1.0 datasets

模型	Params/M	SIMD		DOTA-v1.0	
		mAP0.5/%	mAP0.5:0.95/%	mAP0.5/%	mAP0.5:0.95/%
Faster-RCNN	136.9	62.4	42.1	42.3	25.4
SSD	26.3	59.8	40.8	59.6	37.8
EfficientDet	**12.0**	56.1	40.2	55.7	34.7
YOLOv3	103.7	60.3	41.8	68.2	45.6
YOLOv5m	21.6	69.3	44.3	69.4	46.2
YOLOv8m	25.8	77.2	61.3	71.2	49.8
YOLOv8l	43.6	77.9	61.8	74.8	53.6
YOLOv9c	25.3	78.2	62.3	78.6	56.3
YOLOv10m	16.4	76.3	61.2	73.7	53.3
YOLOv11m	20.1	78.8	62.4	82.1	60.2
YOLOv12m	20.2	77.9	62.0	86.2	62.0
文献[36]	12.5	78.6	62.5	74.6	52.4
Deformable-DETR	40.0	75.2	59.4	78.4	56.5
DEIM	19.2	79.1	63.0	85.3	61.2
文献[19]	18.6	78.7	62.3	85.4	62.1
RT-DETR(r50)	42.9	78.3	62.7	85.9	61.4
RT-DETR(r18)	19.9	77.4	62.0	85.1	60.7
RSD-DETR (ours)	15.5	**79.9**	**63.6**	**86.8**	**62.5**

本文所提模型，不易在实际场景中运用。与当前改进 RT-DETR 模型的文献[19]相比，RSD-DETR 的参数量比其低了 3.1 M，并且检测精度 mAP0.5 在两个数据集上比文献[19]分别提高了 1.2%和 1.4%。RSD-DETR 模型在参数量仅为 15.5 M 的前提下，在 SIMD 和 DOTA-v1.0 和数据集上 mAP0.5 精度分别达到了 79.9%和 86.8%，综合精度 mAP0.5:0.95 分别达到了 63.6%和 62.5%。这得益于轻量级多尺度特征提取模块增强了模型在复杂场景下的抗干扰能力，级联分组注意力对目标的注意多样性，以及 CS-DFFM 结构对多尺度目标的高效跨尺度融合。通过与以上模型的对比验证实验，充分证明了 RSD-DETR 模型以较低参数量设计实现了更高精度的优势，也表明了本文模型的先进性。

2)不同注意力机制对比实验

为了进一步验证 CGA-AIFI 中级联分组注意力的优越性，本文将其与几种常见的注意力机制在具有复杂背景和大量密集小物体的 DOTA-v1.0 数据集上进行了 6 组对比实验，包括 MHSA(multi-head self-attention)、DA(deformable attention)[37]、M2SA(multi-scale multi-head self-attention)[38]、EAA(efficient additive attention)[39]以及 HiLo(HiLo attention)[40]，实验结果如表 4 所示。可以看出，RSD-DETR 引入 CGA-AIFI 模块后，各个指标较其他五种注意力机制均达到了最优且参数量最少。同时 CGA-AIFI 的 FPS 达到了 80.1，远高于 EAA 和 DA，说明其在保持高精度的同时具有更低的计算开销。CGA-AIFI 的级联和分组策略实现了特征信息从局部到全局的渐进融合，显著提升了模型对复杂背景中小目标和的感知能力。

表 4　不同注意力的性能比较

Table 4　Performance comparison of different attention

注意力	Param/M	P/%	R/%	mAP0.5/%	FPS
MHSA	15.8	87.6	80.0	84.8	74.5
DA	15.9	87.7	80.1	85.5	72.6
M2SA	15.9	86.0	81.4	85.3	72.6
EAA	16.0	87.8	80.4	85.5	69.4
HiLO	15.7	87.5	81.1	86.0	76.6
CGA-AIFI	**15.5**	**89.4**	**82.3**	**86.8**	**80.1**

3.6　可视化分析

1)模型热力图

为了直观地展示 RSD-DETR 模型的检测性能，本文采用梯度加权类激活图谱(Grade-CAM)技术进行可视化操作，生成基线模型和 RSD-DETR 模型的热力图。如图 8 所示。图中通过颜色由浅至深表明网络对目标关注程度由低到高的变化，其中红色代表目标特征点最为密集的区域。

从图 8 可以看出，RT-DETR 基线模型对小型目标的

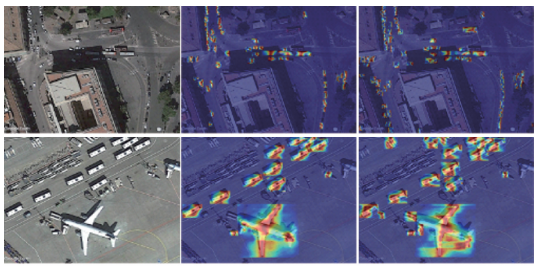


图 8　热力图可视化对比

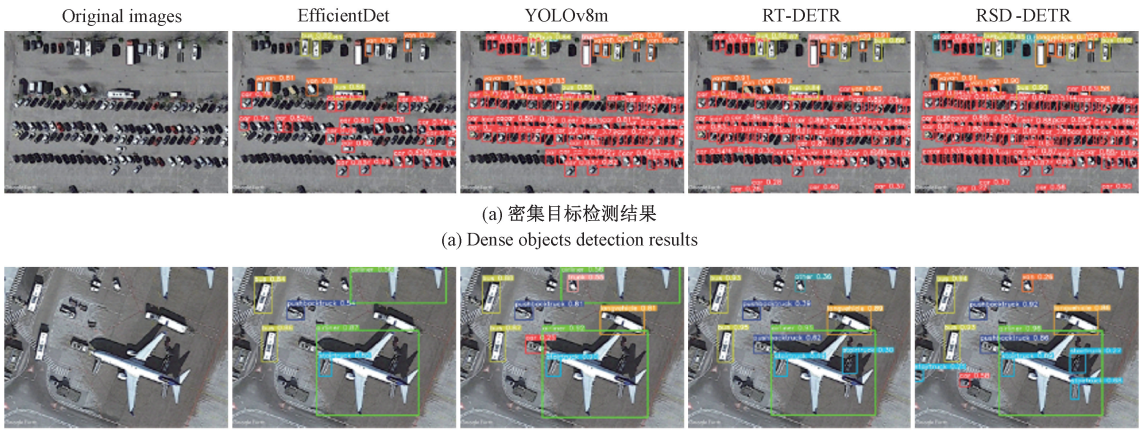
Fig. 8　Heatmap visualization comparison

关注度较低,对复杂环境中的物体不敏感。而 RSD-DETR 模型展现出了显著的优势,能够检测出基线模型未能发现的遮掩目标和小目标,对多尺度和不同种类目标都能准确检测,这进一步证明了改进方法的有效性。

2)检测结果图可视化

为更加清晰地感受到本文算法对遥感图像中不同干扰环境和不同尺度及众多小目标的检测性能,选择经典模型 EfficientDet、当前被广泛应用的 YOLOv8m 模型和基线模型 RT-DETR,与本文 RSD-RTDETR 模型进行了定性对比实验。

从 SIMD 数据集中挑选具有密集目标和多尺度目标的两类经典图像,检测结果如图 9 所示。图 9(a)的密集目标场景图像中可以看出,EfficientDet 模型检测出的目标数量最少,其中在第一排车辆中外形较大的卡车目标也未检测出来。YOLOv8m 模型对于图像下方部分遮挡目标漏检率较高。RT-DETR 模型较 YOLOv8m 准确检测出的目标数量增多,但是仍部分汽车目标未被识别。RSD-DETR 模型对图中所有密集排列的目标车辆几乎都能准确识别,且检出目标的置信度评分整体高于其他模型。对于图 9(b)中多尺度目标场景,EfficientDet 模型对于多尺度目标的漏检率仍然最高。EfficientDet 和 YOLOv8m 模型都没有识别出飞机尾部的两台楼梯车,YOLOv8m 模型将飞机头前的拖车误检为汽车以及将右上方的厢货车误检为卡车。RT-DETR 模型误检和漏检了图像上方的厢货车和左下方的汽车。RSD-DETR 模型凭借级联分组注意力的辅助,提高了模型对于多尺度目标的注意多样性,对多尺度目标的检测效果最优。



(a) 密集目标检测结果

(a) Dense objects detection results

(b) 多尺度目标检测结果

(b) Multi-scale objects detection results

图 9　不同模型在 SIMD 数据集中不同场景下检测结果

Fig. 9　Detection results of different models in various scenarios on SIMD dataset

从 DOTA-v1.0 数据集中挑选具有复杂背景和众多小目标的两类经典图像,检测结果分别如图 10(a)和(b)所示。对于两组场景中的检测效果,EfficientDet 漏检的目标较多且被检测出的目标分类置信度评分最低;YOLOv8m 和 RT-DETR 模型对两类场景中的小目标较敏感,但 YOLOv8m 对图 10(a)中在狭小房屋间的大型车辆和储罐,以及图 10(b)中马路上的小型车辆目标漏检比较多。RT-DETR 对于图 10(b)图像左下角的大型车辆和水中重叠遮挡的船舶目标未识别出来,并且对目标的分类准确率不如 RSD-DETR 模型。相比之下,无论是水面上还是陆地上的小目标,RSD-DETR 模型充分展示了对复杂背景中小目标的敏感性,漏检率最低。说明了模型颈部中的跨尺度动态特征融合结构充分发挥了动态上采样和特征融合的优势,不仅增强了模型的语义建模和上下文感知能力而且避免了小目标特征信息的丢失。

3.7　模型泛化性讨论

为进一步验证本文提出的 RSD-DETR 模型在不同数据集上泛化性,选用 VisDrone2019 数据集进行实验。该数据集是无人机航拍视角下的大型目标检测数据集,其中小目标物体过半,包含行人、人、汽车、面包车、巴士、卡车、摩托车、自行车、雨篷三轮车和三轮车 10 个类别。可以较好测试 RSD-DETR 在无人机场景下目标检测的迁移能力,实验结果如表 5 所示,可视化结果如图 11 所示。

从表 5 中可以看出,RSD-DETR 的各个指标均高于

(a) 复杂背景下目标检测结果

(a) Objects detection results under complex backgrounds

(b) 众多小目标检测结果

(b) Detection results of numerous small objects

图 10　不同模型在 DOTA-v1.0 数据集中不同场景下检测结果

Fig. 10　Detection results of different models in various scenarios on DOTA-v1.0 dataset

YOLOv8m 模型。与 RT-DETR 相比，RSD-DETR 在无人机航拍图像任务中，参数量减低到基线模型 22.11%前提下，准确率和召回率分别达到了 63.8%和 47.5%，mAP0.5 提升了 2.2%，FPS 提高了 9.2。

表 5　在 VisDrone2019 数据集上的泛化性评估

Table 5　Generalization evaluation on VisDrone2019 dataset

模型	Param/M	P/%	R/%	mAP0.5/%	FPS
YOLOv8m	25.8	54.2	42.1	43.5	62.8
RT-DETR	19.9	61.5	46.3	47.6	72.3
RSD-DETR	**15.5**	**63.8**	**47.5**	**49.8**	**81.5**

(a) 原图　(b) 基线模型　(c) RSD-DETR模型

(a) Original images　(b) Baseline model　(c) RSD-DETR model

图 11　VisDrone2019 数据集检测效果

Fig. 11　Detection performance on VisDrone2019 dataset

图 11 展示了 RT-DETR 和 RSD-DETR 模型在 VisDrone2019 数据集的图像检测效果，从中可以看出 RSD-DETR 模型可以检测出距离较远的行人、自行车和摩托车等密集遮挡的小目标，且整体的目标分类置信度高于基线模型。证明了本文模型在高密度复杂交通领域，具有稳健的泛化性。

4　结　　论

本文提出了一种用于遥感图像目标检测 RSD-DETR 模型，旨在解决遥感图像目标检测中目标分布密集、尺度多样和小目标众多等问题导致检测效果不佳的问题。设计了 Faster-CGLU 轻量级特征提取模块，通过优化通道信息的利用，减少模型参数量；构建了 CGA-AIFI 模块，提高了模型对不同尺度目标的感知能力；设计了跨尺度动态特征融合 CS-DFFM 结构，内部的 DySSFF 模块和 TFE 模块，有效增强了模型对多尺度特征融合的稳定性，避免了小目标特征信息的丢失。大量实验结果表明，在 SIMD 和 DOTA-v1.0 数据集上，RSD-DETR 模型在参数量较基线模型降低 22.11%的情况下，mAP0.5 分别达到了 79.9%和 86.8%，较基线模型分别提高了 2.5%和 1.7%，并且推理速度也得到了改善。具有优异的准确性和泛化性。

本文模型虽提高了遥感图像目标检测的精度，降低了模型的参数量。但模型不足够轻量化，未来将会进一步设计更加轻量化的模型结构。尽量保证良好的检测精度前提下进一步提升网络的检测速度，以便应用在其他复杂的检测场景。

参考文献

[1]　LIU C Y, CHEN K Y, QI Z P, et al. Pixel-level change detection pseudo-label learning for remote sensing change captioning[C]. IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium, 2024: 8405-8408.

[2]　王海群，武泽锴，晁帅. 改进 YOLOv8n 的遥感图像目标检测算法[J]. 电子测量与仪器学报，2025，

39(4): 84-94.
WANG H Q, WU Z K, CHAO SH. Improve the YOLOv8n object detection algorithm for remote sensing images[J]. Journal of Electronic Measurement and Instrumentation, 2025, 39(4): 84-94.

[3] 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, 2014: 580-587.

[4] 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.

[5] LIU W, ANGUELOV D, ERHAN D, et al. SSD: Single shot multibox detector[C]. Computer Vision-ECCV 2016, 2016: 21-37.

[6] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: Unified, real-time object detection[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2016: 779-788.

[7] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. ArXiv preprint arXiv: 1706.03762, 2017.

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

[9] 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.

[10] HUANG SH H, LU ZH CH, CUN X D, et al. DEIM: DETR with improved matching for fast convergence[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2025: 15162-15171.

[11] ZHAO Y A, 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, 2024: 16965-16974.

[12] 孙海青, 杨传颖, 敖乐根. 基于改进 RT-DETR 的公路路面交通标识检测算法[J]. 电子测量技术, 2025, 48(8): 187-195.
SUN H Q, YANG CH Y, AO L G. Highway traffic sign detection algorithm based on improved RT-DETR[J]. Electronic Measurement Technology, 2025, 48(8): 187-195.

[13] ZHANG Y, YE M, ZHU G Y, et al. FFCA-YOLO for small object detection in remote sensing images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 1-15.

[14] 闫钧华, 张琨, 施天俊, 等. 融合多层级特征的遥感图像地面弱小目标检测[J]. 仪器仪表学报, 2022, 43(3): 221-229.
YAN J H, ZHANG K, SHI T J, et al. Multi-level feature fusion based dim small ground target detection in remote sensing images[J]. Chinese Journal of Scientific Instrument, 2022, 43(3): 221-229.

[15] ZHOU Y, CHEN S L, ZHAO J Q, et al. CLT-Det: Correlation learning based on transformer for detecting dense objects in remote sensing images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-15.

[16] KONG Y N, SHANG X F, JIA SH J. Drone-DETR: Efficient small object detection for remote sensing image using enhanced RT-DETR model[J]. Sensors, 2024, 24(17): 5496.

[17] 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]. 计算机工程与应用, 2025, 61(1): 98-108.
JIANG M X, SI ZH J, WANG X ZH. Improved target detection algorithm for UAV images with RT-DETR[J]. Computer Engineering and Applications, 2025, 61(1): 98-108.

[18] WANG Y, LI X. Ship-DETR: A transformer-based model for efficient ship detection in complex maritime environments [J]. IEEE Access, 2025, 13: 66031-66039.

[19] SONG B Y, ZHAO SH H, WANG Z D, et al. DAF-DETR: A dynamic adaptation feature transformer for enhanced object detection in unmanned aerial vehicles[J]. Knowledge-Based Systems, 2025, 323: 113760.

[20] CHEN J R, KAO SH H, HE H, et al. Run, don't walk: Chasing higher FLOPS for faster neural networks[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 12021-12031.

[21] SHI D. Transnext: Robust foveal visual perception for vision transformers[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 17773-17783.

[22] CHOLLET F. Xception: Deep learning with depthwise separable convolutions[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2017: 1251-1258.

[23] LIU X Y, PENG H W, ZHENG N X, et al. Efficientvit: Memory efficient vision transformer with

cascaded group attention[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 14420-14430.

[24] 李泽胤，李栋，房建东，等. 改进的 YOLOv8n 遥感图像轻量化检测模型[J]. 电子测量技术，2025，48(6)：130-142.
LI Z Y, LI D, FANG J D, et al. Improved YOLOv8n lightweight detection model for remote sensing images[J]. Electronic Measurement Technology, 2025, 48(6): 130-142.

[25] LIU SH, QI L, QIN H F, et al. Path aggregation network for instance segmentation[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2018: 8759-8768.

[26] 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 and Vision Computing, 2024, 147: 105057.

[27] LIU W Z, LU H, FU H T, et al. Learning to upsample by learning to sample[C]. IEEE/CVF International Conference on Computer Vision, 2023: 6027-6037.

[28] HAROON M, SHAHZAD M, FRAZ M M. Multisized object detection using spaceborne optical imagery[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2020, 13: 3032-3046.

[29] XIA G S, BAI X, DING J, et al. DOTA: A large-scale dataset for object detection in aerial images[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2018: 3974-3983.

[30] TAN M X, PANG R M, LE Q V. Efficientdet: Scalable and efficient object detection[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 10781-10790.

[31] REDMON J, FARHADI A. YOLOv3: An incremental improvement[J]. ArXiv preprint arXiv: 1804.02767, 2018.

[32] WANG C Y, YEH I H, MARK LIAO H Y. YOLOv9: Learning what you want to learn using programmable gradient information[C]. European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024: 1-21.

[33] WANG A, CHEN H, LIU L H, et al. YOLOv10: Real-time end-to-end object detection[J]. Advances in Neural Information Processing Systems, 2024, 37: 107984-108011.

[34] KHANAM R, HUSSAIN M. YOLOv11: An overview of the key architectural enhancements[J]. ArXiv preprint arXiv: 2410.17725, 2024.

[35] TIAN Y J, YE Q X, DOERMANN D. YOLOv12: Attention-centric real-time object detectors[J]. ArXiv preprint arXiv: 2502.12524, 2025.

[36] 秦伦明，梅温泉，崔昊杨，等. 改进 YOLOv8s 的遥感图像目标检测算法[J]. 激光与光电子学进展，2025，62(10)：317-325.
QIN L M, MEI W Q, CUI H Y, et al. Improved YOLOv8s object detection algorithm for remote sensing image[J]. Laser & Optoelectronics Progress, 2025, 62(10): 317-325.

[37] 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, 2022: 4794-4803.

[38] WU H L, HUANG P, ZHANG M, et al. CMTFNet: CNN and multiscale transformer fusion network for remote-sensing image semantic segmentation[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 1-12.

[39] SHAKER A, MAAZ M, RASHEED H, et al. SwiftFormer: Efficient additive attention for transformer-based real-time mobile vision applications[C]. IEEE/CVF International Conference on Computer Vision, 2023: 17425-17436.

[40] PAN Z ZH, CAI J F, ZHUANG B H. Fast vision transformers with hilo attention[J]. Advances in Neural Information Processing Systems, 2022, 35: 14541-14554.

作者简介

陈辉，硕士，主要研究方向为深度学习、目标检测。
E-mail：202312490400@nuist.edu.cn

王新蕾（通信作者），副教授，硕士生导师，主要研究方向为目标检测、图像超分辨率重建、传感器技术等。
E-mail：wangxinlei@nuist.edu.cn