基于 RT-DETR 的小目标检测算法

张云佐[1)] 康尧星[2),3),4)] 刘 婷[1)] 程 煜[3),4)] 任亚恒[3),4)]

[1)](石家庄铁道大学信息科学与技术学院 石家庄 050043)

[2)](石家庄铁道大学交通运输学院 石家庄 050043)

[3)](河北省科学院应用数学研究所 石家庄 050081)

[4)](河北省信息安全认证技术创新中心 石家庄 050081)

摘 要 随着低空经济的发展,小目标逐渐成为目标检测领域的难点。现有小目标检测算法无法满足实际场景需求,主要表现为特征保持能力不足、全局上下文信息利用不充分。为此,本文提出了基于 RT-DETR 的小目标检测算法 SOD-DETR。首先,本文提出了小目标特征保持网络。该网络采用双分支结构,分别提取空域特征和频域特征,通过融合空、频特征来增强特征表达,提高小目标特征保持能力。在频域特征提取时,结合频带分离与注意力机制,实现关键频域特征的自适应学习,以提升频域特征处理的效率。然后,本文构建了一个亚像素-激励模块,该模块融合亚像素处理、Focus 策略及压缩激励机制,将亚像素信息保存到通道维度、引导模块关注重要通道信息,以提升小目标检测性能。最后,通过改进混合编码策略,进一步增强了小目标提取能力。本文在 VisDrone、TT100K 和 UAVDT-2024-DET 三个数据集上进行了测试,实验结果表明:本文提出的 SOD-DETR 明显优于当前主流方法,mAP50 和 mAP50-95 指标提升分别最高达到 6%和 5.1%。

关键词 小目标检测;Transfomer 架构;特征保持;混合编码;亚像素-激励

中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2026.01046

Small Object Detection Algorithm Based on RT-DETR

ZHANG Yun-Zuo[1)] KANG Yao-Xing[2),3),4)] LIU Ting[1)] CHENG Yu[3),4)] REN Ya-Heng[3),4)]

[1)](School of Information Science and Technology, Shijiazhuang Tiedao University, Shijiazhuang 050043)

[2)](School of Traffic and Transportation, Shijiazhuang Tiedao University, Shijiazhuang 050043)

[3)](Institute of Applied Mathematics, Hebei Academy of Sciences, Shijiazhuang 050081)

[4)](Information Security Authentication Technology Innovation Center of Hebei Province, Shijiazhuang 050081)

Abstract With technological breakthroughs and sustained policy support, the low-altitude economy is undergoing explosive growth and has become a key engine for high-quality economic development. As the core carrier of the low-altitude economy, unmanned aerial vehicles (UAVs) have demonstrated enormous potential in the field of intelligent transportation, thanks to their unique advantages of flexible deployment, low operational costs and wide-area coverage. UAV object detection technology, in particular, has been widely applied in a diverse range of practical scenarios, such as road marker detection for traffic infrastructure maintenance, real-time traffic flow statistics for urban traffic management, dynamic traffic guidance to alleviate congestion during peak hours, and rapid accident handling for emergency rescue and post-disaster assessment.

收稿日期:2025-03-04;在线发布日期:2026-01-14。本课题得到国家自然科学基金(No. 61702347)、中央引导地方科技发展资金项目(No. 226Z0501G)、驻冀高校重大科技专项(No. 2512602307A)、河北省自然科学基金重点项目(F2024210008)、河北省科学院重点学科提升项目(No. 25A03)资助。**张云佐(通信作者)**,博士,教授,博士生导师,中国计算机学会(CCF)杰出会员,主要研究领域为计算机视觉、人工智能、大数据。E-mail:zhangyunzuo888@sina.com。**康尧星(通信作者)**,博士研究生,工程师,中国计算机学会(CCF)专业会员,主要研究领域为计算机视觉、人工智能。E-mail:kyaoxing@163.com。**刘 婷**,硕士研究生,主要研究领域为计算机视觉、图像处理。**程 煜**,博士,研究员,中国计算机学会(CCF)杰出会员,主要研究领域为计算机应用、人工智能。**任亚恒**,硕士,高级工程师,中国计算机学会(CCF)专业会员,主要研究领域为计算机视觉、图像处理。

However, UAV aerial images are inherently characterized by a wide field of view and small-sized targets, where small objects of interest often occupy only a few pixels in the entire image frame. Coupled with complex background interference like building occlusions, varying lighting conditions and cluttered terrain, these factors render small object detection a major technical bottleneck that restricts the further application of UAVs in intelligent transportation. Existing small objects detection algorithms still fall short of meeting the rigorous demands of practical scenarios, with their limitations mainly manifested in two key aspects: insufficient feature retention capability, which leads to the easy loss of fine-grained details of small objects during feature extraction; and inadequate utilization of global contextual information, making it hard to distinguish small objects from complex backgrounds effectively. To address these pressing issues, this paper proposes SOD-DETR, an innovative small object detection algorithm based on the RT-DETR framework. First, this paper presents a small object feature retention network, which adopts a dual-branch structure. By extracting and fusing spatial and frequency features, the network achieves complementary enhancement of features, thereby improving the feature retention capability. Specifically, the frequency feature extraction branch integrates frequency band separation and attention mechanism, enabling adaptive learning of critical frequency features and effectively enhancing the efficiency of frequency feature processing. Second, this paper designs a subpixel-excitation module, which combines subpixel processing, Focus strategy and squeeze-and-excitation mechanism. By preserving subpixel information in the channel dimension and guiding the module to focus on key channels, the module boosts feature information and reduces the loss of key features during downsampling. This module achieves improved small object detection accuracy with only a marginal increase in the number of parameters and computational complexity. Finally, this paper improves the hybrid encoding strategy. Through self-attention encoding of the spatial-frequency fused feature maps, it gives full play to the advantages of global contextual modeling in small object detection: it establishes feature correlations on a global scale, implements dynamic weight assignment, breaks the limitation of the local receptive field of Convolutional Neural Network(CNN), provides more external semantic support for small objects with scarce features, and enhances the distinguishability between small object features and background noise, thus further boosting the small object feature extraction capability. Experimental results on the VisDrone, TT100K and UAVDT-2024-DET datasets demonstrate that SOD-DETR significantly outperforms state-of-the-art methods, with notable accuracy improvements in detecting objects of different scales. Specifically, the key metrics mAP50 and mAP50-95 are increased by up to 6% and 5.1%, respectively. The significant improvement in mAP50-95 indicates that SOD-DETR exhibits excellent performance across a broader range of IoU thresholds.

Keywords　small object detection; Transformer architecture; feature preservation; hybrid encode; subpixel-excitation

1　引　言

随着技术突破与政策支持,低空经济正迎来爆发式增长。无人机作为低空经济的核心载体,在智慧交通领域展现出巨大潜力。无人机目标检测技术已经被广泛地应用在路面标志物检测、车流量统计、交通疏导、事故处理等场景中。无人机航拍图像具有视角广、目标小的特点,使得小目标检测成为主要技术难点[1]。事实上,目标检测技术一直都是计算机视觉领域的重要研究方向,提高目标检测精度是一项持续的挑战。近年来,基于深度学习的目标检测技术得到了快速发展,对于常规尺度的目标检测,很多算法都表现出了优秀的性能。但是这些算法在

小目标检测方面依然面临着巨大挑战，其检测性能仍与实际需求存在显著差距。

小目标检测的主要难点在于，小目标本身像素少，特征不明显，且容易受到背景信息干扰。针对小目标检测的这些难点，相关研究工作提出了多种优化方法，以提高小目标检测性能。主要方法包括超分辨率、多尺度融合、上下文学习、生成对抗学习等[2-4]。这些方法都在一定程度上提高了小目标检测性能，为小目标检测的研究提供了重要参考。尤其是多尺度融合策略，基本已经被当前主流的目标检测模型所采用。

但是当前的这些方法，依然遵循着卷积神经网络(Convolutional Neural Network，CNN)的范式。即，要通过多层的池化或跨步卷积等下采样方法来不断降低特征图的空间分辨率，增加特征图的感受野。这一过程，不可避免地会造成小目标特征的丢失，增加小目标识别难度。也有一些研究工作通过优化下采样方法[5-8]，在一定程度上减少了细节信息的丢失。但是，目前来看，如何在下采样过程中保留更多的小目标特征仍然是一个挑战。

近年来，随着 Transfomer 架构被引入视觉领域，基于 Transfomer 的目标检测算法得到了广泛研究。Transformer 通过自注意力机制，直接计算任意两个特征点之间的关联权重，从而构建起全局范围内的特征关联并实现动态权重分配。这一机制打破了 CNN 局部感受野限制，为特征稀缺的小目标提供了更多的外部语义支撑并增强了小目标特征和背景噪声的区分度，最终提升小目标的检测精度。Transfomer 为目标检测提供了另一种范式，其在小目检测方面具有巨大潜力[9-10]。但是由于自注意力机制的计算复杂度与输入序列长度的平方成正比。直接将图像送入 Transformer 编码器是不可行的。所以用 Transfomer 模型进行目标检测时，还是需要先用卷积神经网络完成特征提取。基于 Transfomer 的目标检测模型，从开始的直接将高层的语义特征送入编码器，发展到从不同尺度特征图上选取若干个特征点送入编码器，再到 RT-DETR(Real-Time Detection Transformer)的跨尺度特征融合机制。其检测性能不断提升[11]。

作为视觉 Transfomer 架构的集大成者，RT-DETR 结合了多种 Transformer 模型的优秀机制，其在不同尺度的目标检测上都取得了较为先进的水平。但通过对 RT-DETR 的结构进行分析，可以发现其特征融合策略是将高层语义特征图送入编码器中，再与浅层特征图进行跨尺度融合，这一策略并没能充分利用自注意力机制在小目标检测方面的优势。

为了解决小目标检测中的这些挑战，本文提出了小目标检测算法 SOD-DETR(Small Object Detection Transformer)。该算法主要从两个方面对 RT-DETR 进行改进：一、增强网络的特征保持能力；二、强化全局上下文信息在小目标检测方面的优势。针对第一点，提出双分支结构和亚像素-激励模块。双分支结构增强了网络的特征表达能力，提高了特征的利用率。亚像素-激励模块从源头上增加了特征信息，并突出了重要特征。针对第二点，提出了改进的混合编码策略。本文的主要贡献总结如下：

(1)本文提出了小目标特征保持网络，该网络采用双分支结构，通过提取并融合空、频特征，实现特征的互补增强，从而提高特征保持能力。其中频域特征提取分支，结合了频带分离及注意力机制，可以自适应地学习重要频域特征，有效提升频域特征处理效率。

(2)本文设计了一个亚像素-激励模块，该模块融合了亚像素处理，Focus 策略及压缩激励机制，在增加少量参数量及计算复杂度的条件下，实现了小目标检测精度提升。

(3)本文改进了混合编码策略，通过对空频融合特征图进行自注意力编码，充分发挥全局上下文建模在小目标检测方面的优势，进一步增强小目标特征提取能力。

(4)本文提出的 SOD-DETR 在 VisDrone 2019、TT100K 和 UAVDT-2024-DET 数据集上进行了测试，实验结果证明了算法在增强小目标检测性能上的有效性。

2 相关工作

本节提取了对论文研究内容具有直接影响的关键研究工作，并对其研究内容进行了梳理分析。

2.1 基于 CNNs 的小目标检测技术

典型的基于 CNN 的小目标检测方法，可以概括为以下几个大类别：超分辨率、多尺度特征融合、上下文学习以及下采样优化等[2]。

对图像进行超分辨重建，能够有效增加小目标特征信息。将超分辨率技术与卷积神经网络结合，是一种提高小目标检测性能的重要方法。代表性工作如，Shi 等人[12]提出利用反卷积和亚像素卷积获

得专门用于小目标检测的高分辨率特征。Wang 等人[13]使用 Real-ESRGAN[14]从低分辨率的图像中重建高分辨率图像,并与 YOLOX 结合来提高红外小目标检测能力。

多尺度特征融合技术,就是在不同分辨率和尺度上捕捉并整合信息,使模型能够更全面地理解图像内容。多尺度融合方法中,具有代表性的是基于特征金字塔的方法[15],特征金字塔思想在小目标检测领域应用十分广泛。例如,Hong 等人[16]提出了尺度选择金字塔网络,其通过学习相邻层之间的关系,在深层和浅层之间实现适当的特征共享,从而避免了不同层之间梯度计算的不一致。Xie 等人[17]针对遥感图像目标尺度差异性,提出了一种动态特征融合网络,该网络能根据输入目标尺度,动态学习融合权重。Fan 等人[18]设计了一个跨空间-通道注意力模块和压缩-激励的多尺度特征融合模块,提升了模型对含噪图像中弱小目标的有效信息关注度。Zhang 等人[19]针对无人机图像中的小目标检测问题,提出了一种基于跨层特征聚合的高效目标检测网络。

对于小目标检测来说,小目标只占图像的一小部分,只从小目标本身区域获得的信息受到很大限制。因此,不少研究工作将上下文学习应用到小目标检测,以提高小目标检测性能。Ling 等人[20]提出了基于注意力的上下文感知网络模型,其使用不同膨胀因子的膨胀卷积来聚合高分辨率特征中的上下文信息和局部细节。Chen 等人[21]在 YOlOv7 的基础上,添加了一个上下文结构感知模块和多尺度特征融合模块来提高印刷缺陷中的小目标检测性能。Liang 等人[22]提出利用隐式空间上下文信息,计算类内和类间实例之间的距离,重新检测低置信度的对象。以提高无人机图像小目标检测精度。下采样方法优化包括对池化和卷积层的改进[5]。Lu 等人[6]提出了一个鲁棒特征降采样模块,其通过融合不同降采样技术提取的多个特征图,用一组互补的特征创建了一个更鲁棒的特征图。从而提高了遥感图像中的小目标检测精确率。Hesse 等人[7]提出了一种内容自适应的卷积网络下采样算法,其通过一个预计算的降采样掩模来定义应用于每个像素的降采样操作数量。由此在下采样过程中保留更多局部细节信息。Zhang 等人[8]则提出了一个由权重分支和下采样分支组成的自适应下采样模块,利用 2×2 区域的概率分布关注重要细节,学习更丰富的特征表示。

除上述方法外,数据增强作为一种通用的数据预处理手段,也被广泛用于提高目标检测性能。例如,Kisantal 等人[23]提出了通过对图像中的小目标进行多次复制与粘贴操作来增加样本数量的方法。Wang 等人[24]采用了基于分割的方式来获得更多小目标训练样本。Bosquet 等人[25]则采用了基于生成对抗网络的方法,来生成高质量小目标合成数据。

总体而言,基于 CNN 的小目标检测优化方法已经得到了广泛的研究,这些研究在一定程度上提高了小目标检测的性能。但受限于深度神经网络的范式,其下采样过程中的信息丢失不可避免。因此,如何在下采样过程中保留更多的小目标特征,依然是一项重要挑战。

2.2　基于 Transformer 的小目标检测技术

与基于 CNN 的模型相比,基于 Transformer 的模型具有更大的接受域,由于采用了自注意机制,其更擅长捕捉全局上下文信息。而这种上下文信息在小目标检测中起着重要作用[11]。

作为首个将 Transfomer 引用目标检测领域的算法,DETR[26]实现了完全的端到端的检测。由于减少了超参数数量,其结构更为简洁,也更容易进行优化。但 DETR 收敛速度慢,训练时间长,对算力要求较高。而且,DETR 对小目标检测性能较差。近两年来,针对 DETR 的这些问题,研究人员在 DETR 模型基础上做了大量改进工作。基于 DETR 的改进模型,在提高收敛速度,降低算力要求,增强小目标检测性能等方面都取得了重要突破,这些研究成果,为 DETR 能真正应用于目标检测领域,并且成为一个新的范式,提供了坚实的理论依据。如 Deformable DETR[27]通过多尺度可变形注意力机制,在一定程度上提高了小目标检测精度。并且其采用稀疏采样方法,有效降低了模型的计算复杂度,提高了收敛速度。CIRA-DETR[28]则其利用大小尺度的差异性对小尺度特征图进行信息增强,在小目标检测性能上取得提升。Zhao 等人[29]认为,不同尺度之间的特征相互交互是当前 DETR 模型收敛速度慢的主要原因之一。由此,他们提出了 RT-DETR 算法,其核心是一个高效的混合编码器,通过解耦尺度内交互和跨尺度融合来高效地处理多尺度特征。RT-DETR 在不同尺度的目标上都取得了较高的检测精度。

总体而言,Transformer 架构在小目标检测方面具有巨大的潜力。但如何充分发挥自注意力机制在小目标特征检测方面优势,并平衡好检测精度和

计算复杂度，则是一个重要问题。

3 算法论述

本节，首先对提出的 SOD-DETR 算法的整体架构进行概述。然后，对小目标特征保持网络中的关键模块：亚像素-激励、频域特征提取以及混合编码三部分进行详细论述。

3.1 SOD-DETR 整体结构

本文提出的 SOD-DETR 算法整体结构如图 1 所示。模型采用两个分支来进行特征提取。一条支路，通过 ResNet 网络提取到空域特征图 S3、S4 和 S5。另一条支路，则先通过亚像素激励模块将超分辨率像素保存到通道维度中，并完成初步的通道权重学习。然后，通过频域特征提取模块提取到频域特征图，再将频域特征图与空域特征图 S4 进行特征融合得到 C4。这里 S3 表示浅层特征图，大小为 80×80，S4 表示中间层特征图，大小为 40×40，S5 表示深层特征图，大小为 20×20。接着，混合编码器通过 Transfomer 编码器和一个跨尺度融合模块，将不同尺度的特征图 S3、C4、S5 转化为图像特征序列。这一过程中，先将空频融合后的特征图 C4，输入到基于注意力机制的编码器中，以获得包含全局上下文信息的特征图。再将该特征图与 S3、S5 进行跨尺度特征融合。完成混合编码后，通过对象查询选择固定数量的图像特征作为解码器的初始对象查询。最后，带有辅助预测头的解码器完成目标类别和位置检测。跨尺度特征融合、对象查询模块及后面的解码器部分采用 RT-DETR 模型对应部分，相关内容请参考文献[29]。

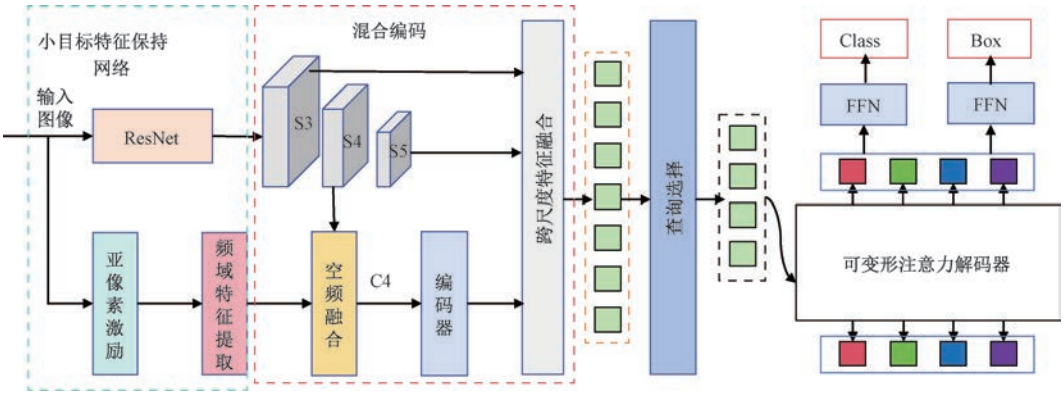

图 1 SOD-DETR 整体结构图

3.2 亚像素-激励模块

小目标本身像素少，包含的特征信息不足是造成小目标检测性能差的重要原因。而优化特征提取网络的方法，如融合频域特征，采用多尺度融合等，其本质上只是改变了特征利用的程度，并未从源头上增加小目标特征信息。所以，采用超分辨率重建的方法来提高图像的细节和清晰度，进而增加小目标特征信息，提高小目标识别定位精度是一种自然的思路，本文在相关工作部分也介绍了超分辨率在小目标检测中的应用。但超分辨率方法存在两个固有的缺点，一是超分辨率方法会增加大量的计算成本；二是超分辨率在增加目标信息的同时，也会增加大量冗余信息。为了克服现有超分辨率方法的这两个缺点，更好地平衡检测性能和计算量，并减少冗余信息，使模型尽早关注到重要特征。受到 YOLO 算法中 Focus 模块的启发，本文设计了一个新颖的亚像素-激励模块。

Focus 模块在 YOLOv5、YOLOx 等多个 YOLO 系列框架中都有应用，已经证明了其在保留空间信息，提高小目标检测性能，并减少计算量方面的有效性。但 Focus 模块是直接对原始图像进行切片处理，并没有结合超分辨率思想。本文提出的亚像素-激励模块则应用了超分辨率思想，并结合了压缩-激励策略，在增加像素信息的同时，降低冗余信息。该模块具体构成如图 2 所示。

首先，对输入的 RGB 图像进行上采样处理，这里使用双线性插值法，目标图像中的新像素值由原图像在它附近区域内的 4 个邻近像素值通过加权平均得到。假设目标图像的一个点(x,y)，其在原图中的四个临近点为 Q_{11}、Q_{12}、Q_{21}、Q_{22}，则(i,j)处的像素值 $f(x,y)$由下式计算得到：

$$f(x,y)=f(Q_{11})\omega_{11}+f(Q_{21})\omega_{21}+f(Q_{12})\omega_{12}+f(Q_{22})\omega_{22} \tag{1}$$

式中，ω_{ij} 为各点对应权重。双线性插值法计算速度快，在保持图像质量的同时，计算复杂度较低。这也是亚像素激励模块中采用双线性插值的重要原因。

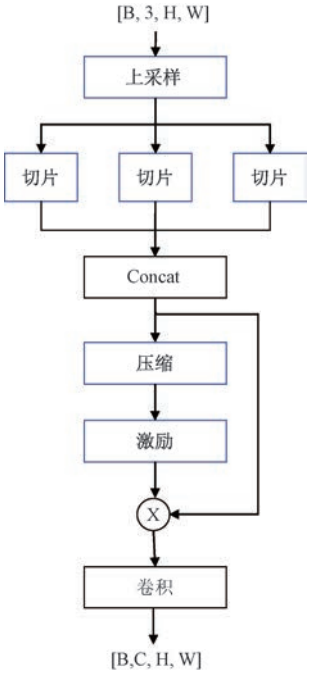

图 2　亚像素-激励模块

采用双线性插值法对图像进行 2 倍放大。得到 $2H\times2W\times3$ 的亚像素图像。然后对图像进行切片处理,即在图像的宽度和高度方向上每隔一个像素进行采样,由此 R、G、B 三个通道的特征图每一个都分解为四个独立的特征层,在通道维度上进行拼接,得到 $H\times W\times12$ 的特征层。将增加的亚像素信息保存到通道中。

此时的通道信息中包含了大量冗余信息。由此引入压缩-激励机制,来建立通道之间的相互依赖关系,自适应地学习各通道权重响应,强调重要信息抑制无关信息。

压缩步骤使用全局平均池化来生成通道级统计量。计算公式如下:

$$z_c=\frac{1}{H\times W}\sum_{x=1}^{H}\sum_{y=1}^{W}x_c(x,y) \tag{2}$$

式中,$x_c(x,y)$代表第 c 个通道的灰度图中位置(x,y)的值。z_c 表示第 c 个通道的统计量。

激励步骤使用 sigmoid 激活门控机制来实现。这里由于只是对冗余信息做初步筛选,且通道数比较少。所以使用一个单层的全连层,公式表达如下:

$$s=\sigma(W_1*z+b_1) \tag{3}$$

式中,σ 表示 sigmoid 激活函数。

之后将通道权重标量 s 加权到每个通道灰度图上,再进行一次 1×1 的卷积操作,得到最终输出。公式表达如下:

$$x_{out}=Conv(x\otimes s) \tag{4}$$

式中,⊗表示按通道相乘。由于 s 的取值范围为[0,1],通过压缩-激励模块,将强化部分重要信息,抑制部分无关信息,这一过程是可以学习的。

3.3　频域特征提取

提取语义特征的同时,不可避免地会丢失掉部分重要特征,如边缘、纹理等。这对小目标检测的影响尤为严重。因此,本文设计了一条频域特征提取支路,将提取的频域特征融入空域特征图中,以补偿下采样中的特征损失,提高小目标特征保持能力。其总体思路是通过频带分离及卷积来提取图像的频域特征,再结合注意力机制筛选出关键的频域特征。其具体实现如图 3 所示。

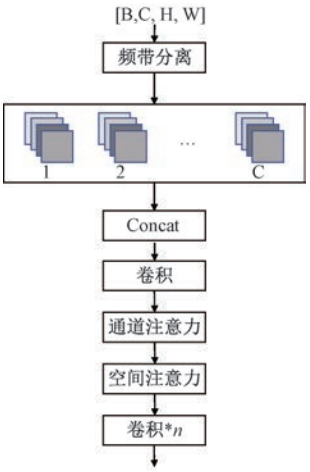

图 3　频域特征提取

首先,对各通道的灰度图进行频带分离处理。频带分离流程如图 4 所示。其实现过程如下:

第一步,对灰度图进行离散傅里叶变换(Discrete Fourier Transform,DFT),如公式(5)所示。

$$F(u,v)=\sum_{x=0}^{H-1}\sum_{y=0}^{W-1}f(x,y)e^{-j2\pi\left(\frac{ux}{H}+\frac{vy}{W}\right)} \tag{5}$$

式中,$f(x,y)$表示原图,$F(u,v)$为对应的频谱图。

第二步,构造两个滤波器,高斯低通滤波器和高斯高通滤波器,分别见公式(6)和公式(7)。

$$H_L(u,v)=e^{-D^2(u,v)/{D_0}^2} \tag{6}$$

$$H_H(u,v)=1-e^{-D^2(u,v)/{D_0}^2} \tag{7}$$

式中,$D(u,v)$表示点(u,v)到频率平面原点的距离;D_0 表示截止频率。$H(u,v)$为高斯滤波器。

第三步,将频谱图与滤波器相乘,后进行反傅里叶变换(Inverse Discrete Fourier Transform,iDFT),即得到分别包含低频信息和高频信息的灰度图,表达式见式(8)。

$$g(x,y)=\frac{1}{hw}\sum_{u=0}^{h-1}\sum_{v=0}^{w-1}F(u,v)H(u,v)e^{j2\pi\left(\frac{ux}{h}+\frac{vy}{w}\right)} \tag{8}$$

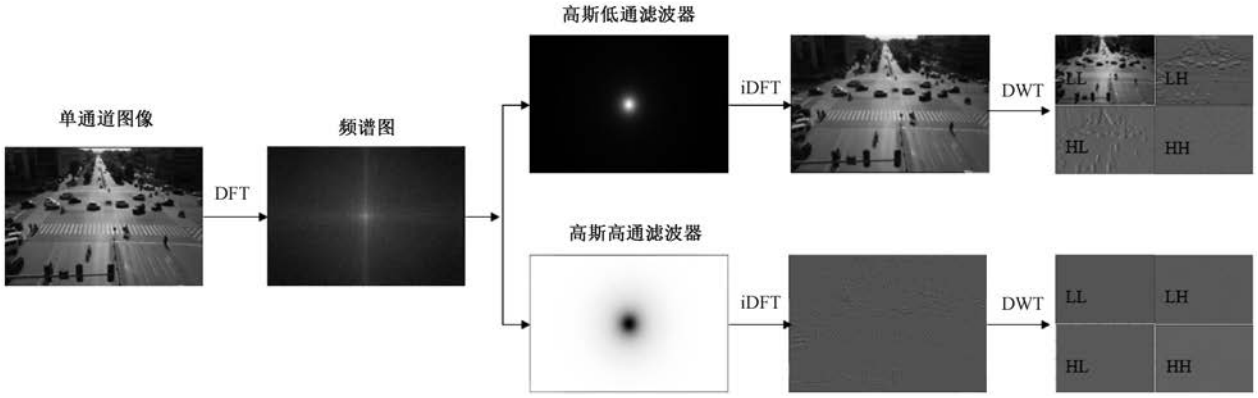

图 4 频带分离流程

式中，$H(u,v)$为 $H_L(u,v)$时，即得到低频灰度图，为 $H_H(u,v)$时，即得到高频灰度图。

第四步，采用二维离散小波变换（Two-Dimensional Discrete Wavelet Transform，2D DWT）分别对高频灰度图和低频灰度图进行再次分解，得到四个频率子带：水平细节（LH）、垂直细节（HL）、对角线细节（HH）和近似子带（LL）。这一部分用四个滤波器，进行步幅为 2 的卷积来实现。四个滤波器分别为 $f_{LL}=\frac{1}{2}\begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix}$、$f_{LH}=\frac{1}{2}\begin{bmatrix}1 & -1\\ 1 & -1\end{bmatrix}$、$f_{HL}=\frac{1}{2}\begin{bmatrix}1 & 1\\ -1 & -1\end{bmatrix}$和 $f_{HH}=\frac{1}{2}\begin{bmatrix}1 & -1\\ -1 & 1\end{bmatrix}$。

则对于每一个输入通道，卷积的输出为

$$[G_{LL},G_{LH},G_{HL},G_{HH}]=\mathrm{Conv}([f_{LL},f_{LH},f_{HL},f_{HH}],g(x,y)) \quad (9)$$

由此完成对图像的频带分离，将每个单通道图，分解为 8 个具有不同频带的子图。并且每个子图的分辨率被下采样到原图的一半。

接着，对 C 个通道的所有子图，在通道维度上进行拼接，拼接后的特征图用 $\boldsymbol{U}_1$ 表示。这一下采样过程，将不同频带的信息保存在了通道维度，并未造成信息的丢失。

然后，采用卷积注意力机制，完成关键频域特征的提取，这一过程中首先应用通道注意力机制，筛选出重要的频带通道。再应用空间注意力机制突出关键区域。通道注意力和空间注意力结构如图 5 和图 6 所示。

通道注意力具体流程是，将输入的特征图分别进行全局最大池化和全局平均池化，将空间维度压缩为 1，保留通道信息。将两个池化后的特征送入共享的多层感知机（Multilayer Perceptron，MLP）提取特征。将经过 MLP 的池化特征相加，经过 sigmoid 激活得到最终的通道注意力权重矩阵 $\boldsymbol{M}_c$。公式表达如下：

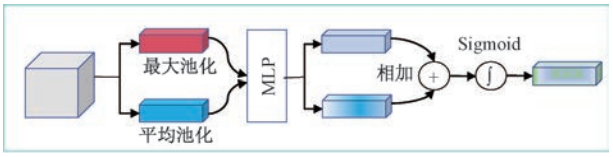

图 5 通道注意力结构图

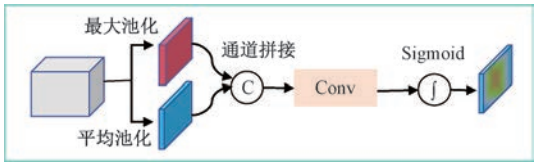

图 6 空间注意力结构图

$$\boldsymbol{M}_c(\boldsymbol{U}_1)=\sigma(MLP(AvgPool(\boldsymbol{U}_1))+MLP(MaxPool(\boldsymbol{U}_1)))=\sigma(\boldsymbol{W}_1(\boldsymbol{W}_0(F^c_{\mathrm{avg}}))+\boldsymbol{W}_1(\boldsymbol{W}_0(F^c_{\max}))) \quad (10)$$

式中，σ 代表 sigmoid 函数，$\boldsymbol{W}_1$ 和 $\boldsymbol{W}_0$ 代表 MLP 的权重矩阵。将权重值与特征图 $\boldsymbol{U}_1$ 进行点乘，得到加权后的特征图 $\boldsymbol{U}_2$。

然后，对 $\boldsymbol{U}_2$ 特征层进行空间注意力处理，以进一步突出其在空间维度上的特征。空间注意力模块具体实现流程如下：将特征图（经过通道注意力计算后的）在通道维度分别进行最大值池化和平均池化，将通道维度压缩为 1，保留空间信息。将池化特征在通道维度拼接，再经过一个卷积层提取特征，同时将通道维度降至 1。经过 sigmoid 激活，得到最终的空间注意力权重。空间注意力权重矩阵表达式如下：

$$\boldsymbol{M}_s(\boldsymbol{U}_2)=\sigma(f^{3\times3}Concat(F^s_{\mathrm{avg}},F^s_{\max})) \quad (11)$$

式中，f 代表卷积运算，这里卷积核取 3。

再将权重值乘到原特征图上，就得到最终的频域特征图。

最后,再通过 n 次卷积进行下采样,并使特征图大小与 ResNet 提取的 S4 特征图相一致。同样在通道维度上对两条支路提取的特征图进行拼接,完成空频特征的融合。

3.4　改进混合编码模块

百度研发的 RT-DETR 目标检测模型[26],最早提出了混合编码器结构。RT-DETR 论文作者分析了多尺度 Transformer 编码器中存在的计算冗余,并设计了一组变体来证明尺度内和跨尺度特征的同时交互在计算上是低效的。由此,他们重新思考了编码器的结构,提出了一种新型的高效混合编码器。所提出的混合编码器由两个模块组成,一个是 Transfomer 编码器,用于尺度内特征交互。另一个则是基于 CNN 的跨尺度特征融合模块。而且,他们认为相较于浅层的 S3 特征图和 S4 特征图,S5 拥有更深、更高级、更丰富的语义特征,这些语义特征是 Transformer 更加感兴趣的和需要的,对于区分不同目标的特征是更加有用的,而浅层特征因缺少较好的语义特征而起不到什么作用。所以仅在最后一层的语义特征图 S5 上进行了尺度内交互。RT-DETR 的混合编码器结构如图 7(a)所示。图中 Enc 代表 Transfomer 编码器,CCFM 代表跨尺度融合模块。RT-DETR 通过混合编码器有效融合了多尺度特征和自注意力机制,同时又解决了多尺度 Transformer 架构的计算冗余问题,使得 RT-DETR 与其他实时检测器和类似大小的端到端检测器相比,在速度和精度方面都达到了当时最先进的性能。但是,深层语义特征层 S5 会丢失掉小目标特征,虽然跨尺度特征融合模块融合了浅层特征图信息,一定程度上弥补了小目标检测方面的不足,但整体来看 RT-DETR 混合编码器模块并没能充分发挥出这一机制对小目标检测的能力。有研究证明,自注意力机制的全局上下文建模对小目标特征识别是十分有利的。由此,本文提出了改进的混合编码模块。我们沿用了 RT-DETR 混合编码器中的编码器模块和基于 CNN 的跨尺度特征融合模块,但对其整体融合策略进行了改进,以进一步增强模型的小目标检测能力。本文设计的混合编码模块如图 7(b)所示。

新的融合策略如下:基于注意力机制的尺度内特征交互模块输入的是由小目标特征提取网络得到的空频融合特征图 C4。C4 特征图固定大小为 40×40,比较直接从 ResNet 提取的语义特征图 S5,其包含了不同频谱的频域特征,特征信息更为丰富。将

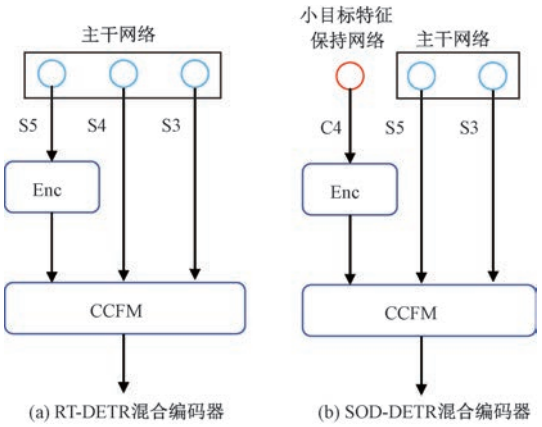

图 7　特征融合策略比较

C4 送入到自注力编码器中,可以得到包含全局上下文信息的特征图。然后,再将 C4 编码后的特征图与 ResNet 网络的提取特征图 S3、S5 进行跨尺度融合。通过这种策略,可以进一步提高小目标特征提取能力。

4　实　验

模型构建采用 Python3.9 和 Pytorch1.13.0。实验运行环境是 Windows 10 系统,GPU 为 NVDIA GeForce RTX 3060。关键的超参数设定如下:批处理量(batch size)为 2,实验使用 AdamW 优化器,初始学习率为 0.0001,模型训练前未使用预训练权重。

实验采用的数据集是 VisDrone 2019、TT100K 和 UAVDT-2024-DET。VisDrone 数据集是一个专为无人机拍摄的日常场景中目标检测而设计的数据集[30]。该数据集聚焦于小目标检测,60%的目标实例大小小于 20 像素,25%的目标实例大小在 20-30 像素之间,这对算法提出了更高的要求。适合用于测试小目标检测算法的性能。

TT100K(Tsinghua-Tencent 100K)是一个专为交通标志识别设计的大型数据集,由清华-腾讯联合实验室提出[31],涵盖 100 类常见交通标志,总计 30000 多个精细标注实例。该数据集中的交通标志通常较小,且存在光照变化、遮挡、模糊等问题,是一个典型的适用于小目标检测的数据集。因标签类别不是特别均衡,因此本文只处理了其中标签数量大于 100 的图片,共 45 类。

UAVDT 是由中国科学院大学牵头提出的大规模无人机检测与跟踪基准数据集,聚焦无人机视角下车辆的检测与跟踪任务[32]。UAVDT-2024-DET 是基于 UAVDT 的目标检测专项版本,对目标检测

任务相关标注进行了校验与优化，适合评估算法在动态视角、尺度变化、复杂光照等条件下的鲁棒性。

本文采用 COCO 风格的评价指标。用 AP 来衡量训练好的模型在每个类别上的好坏，用 mAP 衡量模型在所有类别上的好坏。其中 mAP50 指 IOU 阈值取 0.5 时，各个类别上平均精度。mAP50-95 则是指，以 0.05 为步长，从 0.5 到 0.95 取 10 个 IoU 阈值，计算各阈值下的 AP，然后取平均值。AP_S 代表小目标，AP_M 代表中等尺度目标，AP_L 代表大尺度目标。为了进一步衡量模型的计算复杂度、推理时间等性能，本文采用了参数量、每秒千兆浮点运算(GFLOPs)和每秒帧数(FPS)这几种评价指标。

4.1 实验分析

为了展示 SOD-DETR 相对于基准模型 RT-DETR 对目标的检测效果，本文将训练过程中 mAP 随训练轮数增加的变化情况进行了可视化，结果如图 8 和图 9 所示，这里的 mAP 是在验证数据集上的检测值。图 8 是在 VisDrone 2019 数据集上的训练情况，从图中可以看出，模型在训练早期，收敛速度都比较快，但 SOD-DETR 的变化更为平稳，20 轮后，mAP50 的增长越来越平缓，在 100 轮左右时，mAP50 值已经达到最优值，模型收敛。此时 SOD-DETR 的 mAP50 值明显高于 RT-DETR。图 9 是在 TT100K 数据集上的训练情况，在 30 个 epoch 后，mAP50-95 的增长趋于平缓，在 80 个 epoch 左右时，模型收敛。由图中可以看出，在该数据集上，SOD-DETR 的 mAP50-95 值也明显高于 RT-DETR。综合这两个数据集上的实验结果来看，SOD-DETR 在训练阶段表现出更强的学习能力和更高的检测精度。基于 mAP-epochs 曲线，本文选取了 VisDrone 数据集上的第 100 个 epoch 的模型，和 TT100K 数据集上的第 80 个 epoch 的模型。来对 SOD-DETR 算法进行进一步的测试和评估。UA-VDT-2024-DET 数据集上 mAP-epochs 曲线整体变化趋势与前两者比较接近，这里不再给出。

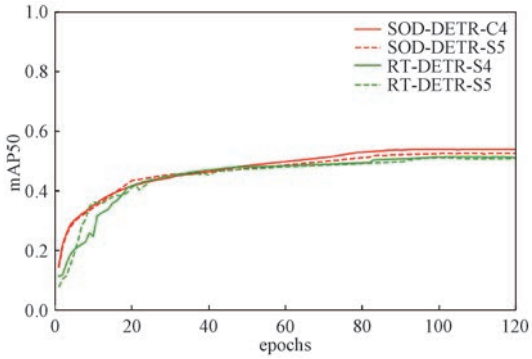

图 8 VisDrone 2019 验证集上的 mAP-epochs 曲线

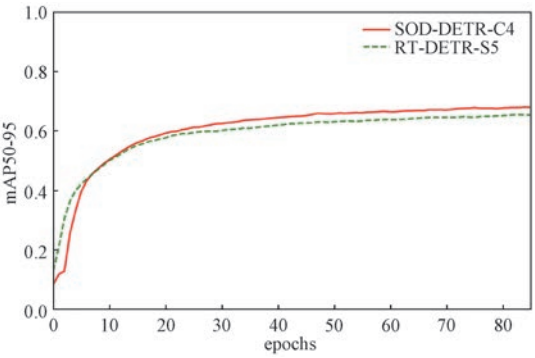

图 9 TT100K 验证集上的 mAP-epochs 曲线

表 1 给出了典型算法在 VisDrone 2019 数据集上的比较结果。与 RT-DETR 比较，SOD-DETR 的 mAP50 提高了 3.2%，mAP50-95 提高了 5.1%，AP_S 提高了 2.7%，AP_M 提高了 2.8%，AP_L 提高了 5.7%。而相比 Drone-DETR，mAP50 虽然提升不多，但在 mAP50-95 上依然有 1.5%的提升。SOD-DETR 的空频特征融合机制有效增强了目标特征表达，减少了下采样过程中的特征损失，提高了不同尺度目标的识别定位精度。

表 1 典型算法在 VisDrone 2019 数据集上的比较结果

网络模型	mAP50 (%)	mAP50-95 (%)	AP_S (%)	AP_M (%)	AP_L (%)	FPS	参数量(M)	GFLOPs
YOLOx[33]	40.5	26.2	12.3	35.4	50.2	130.0	9.0	26.8
SWIN-T[34]	28.9	17.6	9.3	254.0	28.1	—	29.0	—
Deformable-DETR[27]	33.7	19.0	11.1	28.3	37.1	24.0	40.0	196.0
DINO[35]	52.3	34.4	20.6	44.8	56.7	—	47.0	279.0
Drone-DETR[36]	53.9	33.9	—	—	—	—	—	128.3
SOD-YOLO[37]	50.7	30.0	21.0	40.7	—	72.5	30.3	83.5
RT-DETR_r50[29]	50.9	30.3	20.2	42.8	58.6	73.0	42.7	136.0
SOD-DETR	**54.1**	**35.4**	**22.9**	**45.6**	**64.3**	64.0	44.3	145.9

表 2 给出了 SOD-DETR 和 RT-DETR 在 TT100K 数据集上的比较结果。在该数据集上,SOD-DETR 的 mAP50 提高了 2.2%,mAP50-95 提高了 3.7%,AP_S 提高了 2.4%,AP_M 提高了 2.2%,AP_L 提高了 2.8%。

表 2　RT-DETR 与 SOD-DETR 在 TT100K 数据集上的比较结果

网络模型	mAP50 (%)	mAP50-95 (%)	AP_S (%)	AP_M (%)	AP_L (%)
RT-DETR	89.1	65.0	43.7	74.2	85.7
SOD-DETR	**91.3**	**68.7**	**46.1**	**76.4**	**88.5**

表 3 给出了 SOD-DETR 和 RT-DETR 在 UAVDT-2024-DET 数据集上的比较结果。在该数据集上,SOD-DETR 的 mAP50 提高了 6%,mAP50-95 提高了 3.5%,AP_S 提高了 4.7%,AP_M 提高了 6%,AP_L 提高了 1.5%。

表 3　RT-DETR 与 SOD-DETR 在 UAVDT-2024-DET 数据集上的比较结果

网络模型	mAP50 (%)	mAP50-95 (%)	AP_S (%)	AP_M (%)	AP_L (%)
RT-DETR	62.2	39.1	30.7	48.6	82.1
SOD-DETR	**68.2**	**42.6**	**35.4**	**54.6**	**83.6**

综合来看,模型在不同数据集上的表现具有一定差异性,这是受到数据集本身的影响。但整体来看,相对于 RT-DETR,SOD-DETR 的各项检测精度都有所提升:mAP50 和 mAP50-95 指标提升分别最高达到 6%和 5.1%,而 mAP50-95 是比 mAP50 更严格的评估指标,这就说明 SOD-DETR 在多个 IoU 阈值下的检测性能更为出色,尤其是在高精度阈值下表现更好,且模型具有较好的泛化性;与此同时,该模型在小目标检测任务上的精度提升同样明显,充分证明 SOD-DETR 在小目标检测方面具有明显技术优势。

当然,结合计算复杂度和推理速度来看,相比 Drone-DETR 和 SOD-YOLO,SOD-DETR 的参数量和计算复杂度都有所增加,但增加的幅度并不大。模型推理速度也能维持在每秒 64 帧。SOD-DETR 在对检测精度要求较高的任务中是更为适合的。

为了更直观地说明 SOD-DETR 算法性能,本文对 VisDrone 2019、TT100K 和 UAVDT-2024-DET 数据集上的测试进行了可视化,部分检测结果如图 10、图 11 和图 12 所示。图中最左侧为原图,中间为 RT-DETR 算法检测结果,最右边则是 SOD-DETR 算法检测结果。图中的虚线框及箭头指向部分,代表对图中指定区域进行放大,图中实线方框表示检测出的目标。不同颜色则代表不同类别。实验中设定置信度为 0.6,图中只保留大于该置信度的目标框。由图 10 和图 12 可以直观看出,SOD-DETR 模型比基准模型 RT-DETR 检测出了更多的远距离处的车辆、行人等物体。由图 11 给出的三组图片可以看出,SOD-DETR 对于小目标交通标志的误检和漏检数量明显少于 RT-DETR。总体而言,由以上实验结果可以看出,SOD-DETR 在目标检测性能上要优于 RT-DETR。

图 10　可视化结果对比图 1(VisDrone 2019)

图 11 可视化结果对比图 2(TT100K)

图 12 可视化结果对比图 3(UAVDT-2024-DET)

4.2 消融实验

为了进一步测试 SOD-DETR 中各模块的有效性,继续在 VisDrone 2019 数据集上做了一系列消融实验。

首先,对模型的命名方式进行说明,消融实验部分,本文采用了算法名称加融合策略的命名方式,如 SOD-DETR-S5{S5|S3,C4},{S5|S3,C4}是对融合策略的说明,表示将 ResNet 网络提取的 S5 特征图送入编码器中,再与浅层特征图 S3 以及小目标特征保持网络提取到的空频融合特征图 C4 进行融合。部分图表中采用了简化的命名方式,去掉了融合策略说明部分,如 SOD-DETR-S5。其他模型的表征方式与此相同。实验结果统计在表 4 中,以下将结合实验数据对重要模块的有效性进行分析。

4.2.1 亚像素-激励模块分析

小目标特征保持网络中,亚像素激励模块能够提高图像的细节和清晰度,进而增加特征信息。

为验证该模块的有效性,本文设计了两组对照实验,第一组:SOD-DETR-C4{C4|S3,S5}和 SOD-DETR-C4(去掉亚像素激励模块);第二组:SOD-DETR-S5{S5|S3,C4}和 SOD-DETR-S5(去掉亚像素激励模块)。由表 4 中数据可知,SOD-DETR-S5 去掉亚像素激励模块后,mAP50 降低了 1.4%,mAP50-95 降低了 1.6%,AP_S 降低了 0.7%。SOD-DETR-C4 去掉亚像素激励模块后,mAP50 降低了 1.6%,mAP50-95 降低了 1.7%,AP_S 降低了 0.8%。这表明亚像素激励模块能有效提高检测性能,而由表 4 中的数据可以看到,加入亚像素激励模块后,其参数量和计算复杂度只有少量的增长,相比于检测性能的提升,这些代价几乎是可以忽略的。

4.2.2　频域特征提取模块分析

SOD-DETR-S5 和 SOD-DETR-C4 都去掉亚像素激励模块后,相比 RT-DETR-S5 和 RT-DETR-S4,其性能的变化主要归因于频域特征的融合。由表 4 数据,SOD-DETR-S5(去掉亚像素激励模块)相比 RT-DETR-S5,mAP50 提高了 1%,mAP50-95 提高了 2.9%,AP_S 提高了 0.8%,AP_M 提高了 1.3%,AP_L 提高了 2.7%。SOD-DETR-C4(去掉亚像素激励模块)相比 RT-DETR-S4,mAP50 提高了 1.3%,mAP50-95 提高了 2.8%,AP_S 提高了 1.2%,AP_M 提高了 1.7%,AP_L 提高了 2.7%。

由以上两组数据可以得出,融合频域特征对大中小不同尺度的目标检测性能都有明显提升。而在这些性能指标中,mAP50-95 的提升尤为显著。这说明,SOD-DETR 在更广泛的 IoU 阈值范围内的出色性能主要可以归因于其有效融合了频域特征。

4.2.3　改进混合编码器模块分析

不同融合策略对检测性能的影响,可以由 RT-DETR-S5 与 RT-DETR-S4、SOD-DETR-S5(去掉亚像素激励模块)与 SOD-DETR-C4(去掉亚像素激励模块),SOD-DETR-S5 与 SOD-DETR-C4,这三组对照实验分析得出。RT-DETR-S5 就是基础模型 RT-DETR-r50,采用的是{S5|S3,S4}融合策略。RT-DETR-S4 则采用了{S4|S3,S5}融合策略。在仅改变融合策略的情况下,RT-DETR-S4 相比 RT-DETR-S5,mAP50 提高了 0.3%,mAP50-95 提高了 0.6%,AP_S 提高了 0.7%,AP_M 提高了 1.1%,AP_L 提高了 0.4%。SOD-DETR-S5 与 SOD-DETR-C4 都去掉亚像素激励模块后,其区别也只在于融合策略不同,采用{C4|S3,S5}融合策略的模型与采用{S5|S3,C4}融合策略的模型比较,mAP50 提高了 0.6%,mAP50-95 提高了 0.5%,AP_S 提高了 1.1%,AP_M 提高了 1.5%,AP_L 提高了 0.4%。

SOD-DETR-S5 和 SOD-DETR-C4 的数据见表 4。虽然各精度提升具体数值有微小差异,但各精度的相对提升比例基本与上一组模型接近。由这三组数据可以看出,融合策略的改变对检测精度具有直接影响。{C4|S3,S5}融合策略,对中小尺度目标的检测精度提升更为明显,对大尺度目标的影响相对较小。而将中层尺度特征图(S4/C4)送入到编码器,并没有改变模型的参数量,但在计算复杂度上会有一定的提升。

表 4　消融实验结果(VisDrone 2019)

网络模型	mAP50 (%)	mAP50-95 (%)	AP_S (%)	AP_M (%)	AP_L (%)	FPS	参数量 (M)	GFLOPs
RT-DETR-S5{S5|S3,S4}	50.9	30.3	20.2	41.7	58.6	73.0	42.7	136.2
RT-DETR-S4{S4|S3,S5}	51.2	30.9	20.9	42.8	59.0	72.0	42.7	137.4
SOD-DETR-S5(去掉亚像素激励模块)	51.9	33.2	21.0	43.0	61.3	68.0	44.29	141.7
SOD-DETR-C4(去掉亚像素激励模块)	52.5	33.7	22.1	44.5	61.7	67.0	44.29	142.9
SOD-DETR-S5{S5|S3,C4}	53.3	34.8	21.7	43.9	63.6	65.0	44.3	144.7
SOD-DETR-C4{C4|S3,S5}	54.1	35.4	22.9	45.6	64.3	64.0	44.3	145.9

如果将更大尺度的 S3 特征图送入编码器,其计算复杂度是不可接受的。综合考虑小目标检测精度和计算复杂度之间的平衡,本文最终采用了{C4|S3,S5}的融合策略。

5　结　论

本文提出了基于 RT-DETR 的小目标检测算法 SOD-DETR。针对现有算法存在的特征保持能力不足问题,提出了小目标特征保持网络并设计了一个独立的亚像素-激励模块。其中,小目标特征保持网络通过融合空、频特征,实现特征的互补增强,提高了网络的特征保持能力。实验结果证明,该网络对不同尺度的目标检测性能都有明显提升;而亚像素-激励模块通过从源头上增加特征信息,并突出重要特征,进一步强化了网络的特征保持能力。实验

结果证明,加入亚像素-激励模块能有效提升模型的检测性能。

针对全局上下文信息利用不充分问题,提出了改进的混合编码策略,通过对空频融合特征图进行自注意力编码,充分发挥全局上下文建模在小目标检测方面的优势。实验结果证明,改进的混合编码策略能进一步提高小目标检测精度。

整体来看,相比于 RT-DETR,本文设计的 SOD-DETR 模型,在不同尺度的目标检测精度上都取得了明显提高。此外,融合频域特征后,mAP50-95 的显著提升,也使得 SOD-DETR 在更广泛的 IoU 阈值范围内具有出色的性能。后续,将开展模型轻量化研究,在保持模型性能的前提下,尽可能减少模型参数量和计算量。

致 谢 我们感谢编辑部和所有审稿人给本论文提出的宝贵意见。

参 考 文 献

[1] Wei Wei, Cheng Yu, He Jiafeng, et al. A review of small object detection based on deep learning. Neural Computing and Applications, 2024, 36(12): 6283-6303

[2] Chen Guang, Wang Haitao, Chen Kai, et al. , A survey of the four pillars for small object detection: Multiscale representation, contextual Information, super-resolution, and region proposal. IEEE Transactions on Systems Man and Cybernetics Systems, 2022,52(2): 936-953

[3] Zhang Yunzuo, Wu Cunyu, Zhang Tian, et al. Full-scale feature aggregation and grouping feature reconstruction based UAV image target detection. IEEE Transactions on Geoscience and Remote Sensing, 2024. , 62: 1-11

[4] Zhang Yunzuo, Zhen Jiawen, Liu Ting, et al. Adaptive differentiation siamese fusion network for remote sensing change detection. IEEE Geoscience and Remote Sensing Letters, 2025, 22: 1-5

[5] Rippel O, Snoek J, Adams R P. Spectral representations for convolutional neural networks//Proceedings of the Advances in Neural Information Processing Systems. Montreal, Canada, 2015: 2449-2457

[6] Lu Wei, Chen Sibao, Tang Jin, et al. A robust feature downsampling module for remote-sensing visual tasks. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 1-12

[7] Hesse R, Schaub-Meyer S, Roth S. Content-adaptive downsampling in convolutional neural networks//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada,2023: 4544-4553

[8] Zhang Yunzuo, Liu Ting, Zhen Jiawen, et al. Adaptive downsampling and scale enhanced detection dead for tiny object detection in remote sensing image. IEEE Geoscience and Remote Sensing Letters, 2025, 22: 1-5

[9] Li Zheng, Wang Yongcheng, Feng Hao, et al. Local to global: A sparse transformer-based small object detector for remote sensing images. IEEE Transactions on Geoscience and Remote Sensing, 2025,63: 1-16

[10] Han Zixian, Jia Dongli, Zhang Lei, et al. FNI-DETR: Real-time DETR with far and near feature interaction for small object detection. Engineering Research Express, 2025, 7(1): 015204

[11] Amjoud A B, Amrouch M. Object detection using deep learning, CNNs and vision transformers: A review. IEEE Access, 2023, 11: 35479-35516

[12] Shi W Z, Caballero1 J , Huszar F, et al. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 1874-1883

[13] Wang Zhiyong, Xiang Xuefu, Zeng Kan, et al. Infrared small target detection based on the combination of single image super-resolution reconstruction and YOLOX//Proceedings of the 2023 2nd Asia Conference on Algorithms, Computing and Machine Learning. Shanghai, China, 2023: 547-552

[14] Wang Xintao, Xie Liangbin, Dong Chao, et al. Real-esrgan: Training real-world blind super-resolution with pure synthetic data//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 1905-1914

[15] Lin T Y, Dollar P, Girshick R, et al. Feature pyramid networks for object detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu,USA, 2017: 936-944

[16] Hong Mingbo, Li Shuiwang, Yang Yuchao, et al. SSPNet: Scale selection pyramid network for tiny person detection from uav images. IEEE Geoscience and Remote Sensing Letters, 2021, 19: 1-5

[17] Xie Xingxing, Cheng Gong, Yao Yanqing, et al. Dynamic feature fusion for object detection in remote sensing images. Chinese Journal of Computers, 2022,45(4):735-747(in Chinese)
(谢星星, 程塨, 姚艳清等. 动态特征融合的遥感图像目标检测. 计算机学报,2022,45(4):735-747)

[18] Fan Mingkai, Xue Danna, Yan Qingsen, et al. Dim and small space target detection method based on ehanced information representation. Chinese Journal of Computers, 2025, 48(3):537-555(in Chinese)
(范铭楷, 薛丹娜, 闫庆森等. 基于信息表征增强的空间弱小目标检测方法. 计算机学报,2025,48(3):537-555)

[19] Zhang Yunzuo, Wu Cunyu, Guo Wei, et al. CFANet: Efficient detection of UAV image based on cross-layer feature aggregation. IEEE Transactions on Geoscience and Remote

Sensing, 2023, 61: 1-11

[20] Ling Siyao, Chen Lunfeng, Wu Yujie, et al. ACANet: Attention-based context-aware network for infrared small target detectionsk. The Journal of Supercomputing, 2024, 80(12): 17068-17096

[21] Chen Wenqian, Zheng Yuanlin, Liao Kaiyang, et al. Small target detection algorithm for printing defects detection based on context structure perception and multi-scale feature fusion. Signal, Image and Video Processing, 2024, 18(1): 657-667

[22] Liang Xi, Zhang Jing, Zhuo Li, et al. Small object detection in unmanned aerial vehicle images using feature fusion and scaling-based single shot detector with spatial context analysis. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 30(6):1758-1770

[23] Kisantal M, Wojna Z, Murawski J, et al. Augmentation for small object detection (arXiv), https://arxiv. org/abs/1902.07296 2019

[24] Wang Xiaobin, Zhu Dekang, Yan Ye. Towards efficient detection for small objects via attention-guided detection network and data augmentation. Sensors, 2022, 22(19): 7663

[25] Bosquet B, Cores D, Seidenari L, et al. A full data augmentation pipeline for small object detection based on generative adversarial networks. Pattern Recognition, 2023, 133: 108998

[26] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers//Proceedings of the European Conference on Computer Vision. Online,2020: 213-229

[27] Zhu Xizhou, Su Weijie, Lu Lewei, et al. Deformable DETR: Deformable transformers for end-to-end object detection//Proceedings of the International Conference on Learning Representations. Vienna, Austria, 2021: 894-910

[28] Ning Xin, Tian Weijuan, Yu Lina, et al. Brain-inspired CIRA-DETR full Inference model for small and occluded object detection. Chinese Journal of Computers, 2022, 45(10): 2080-2092(in Chinese)
(宁欣, 田伟娟, 于丽娜等. 面向小目标和遮挡目标检测的脑启发 CIRA-DETR 全推理方法. 计算机学报, 2022,45(10): 2080-2092)

[29] Zhao Yan, Lv Wenyu, Xu Shangliang, et al. Detrs beat yolos on real-time object detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2024: 16965-16974

[30] Du Dawei, Zhu Pengfei, Wen Longyin, et al. VisDrone-DET2019: The vision meets drone object detection in image challenge results//Proceedings of the IEEE International Conference on Computer Vision Workshop. Seoul, Republic of Korea, 2019: 213-226

[31] Zhu Zhe, Liang Dun, Zhang Songhai, et al. Traffic-sign detection and classification in the wild//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 2110-2118

[32] Du Dawei, Qi Yuankai, Yu Hongyang, et al. The unmanned aerial vehicle benchmark: Object detection and tracking//Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018: 370-386

[33] Ge Zheng, Liu Songtao, Wang Feng, et al. YOLOX: Exceeding YOLO series in 2021, https://arxiv. org/2107.08430 2021

[34] Liu Ze, Lin Yutong, Cao Yue, et al. Swin transformer: Hierarchical vision transformer using shifted windows//Proceedings of the IEEE International Conference on Computer Vision. 2021: 10012-10022

[35] Zhang Hao, Li Feng, Liu Shilong, et al. DINO: DETR with improved denoising anchor boxes for end-to-end object detection//Proceedings of the International Conference on Learning Representations. Kigali,Republic of Rwanda, 2023,8587-8606

[36] Kong Yaning, Shang Xiangfeng, Jia Shijie. Drone-DETR: Efficient small object detection for remote sensing image using enhanced RT-DETR model. Sensors, 2024, 24(17): 5496

[37] Xiao Yunze, Di Nan. SOD-YOLO: A lightweight small object detection framework. Scientific Reports, 2024, 14(1): 25624

ZHANG Yun-Zuo, Ph. D., professor, Ph. D. supervisor. His current research interests include computer vision, artificial intelligence, and big data.

KANG Yao-Xing, Ph. D. candidate, engineer. His current research interests include computer vision and artificial intelligence.

LIU Ting, M. S. candidate. His current research interests include computer vision and image processing.

CHENG Yu, Ph. D., professor. His current research interests include computer application and artificial intelligence.

REN Ya-Heng, M. S., senior engineer. His current research interests include computer vision and image processing.

Background

Small object detection is currently a key and difficult problem in the field of computer vision. Industrial defect images, UAV images, satellite remote sensing images, and other types of images all contain a large number of small objects. When identifying and locating objectts in the image, the phenomenon of false detection and omission of small ob-

jects is very serious. This directly affects the application of computer vision technology in related industries.

The main difficulty of small object detection lies in the fact that small objects have few pixels, unclear features, and are easily affected by background information interference. In response to these difficulties in small object detection, various optimization methods have been proposed in related research to improve the performance of small object detection. The main methods include data augmentation, multi-scale fusion, context learning, generative adversarial learning, etc. These methods have all improved the performance of small object detection to a certain extent, providing important references for the research of small object detection. In recent years, with the introduction of Transfomer architecture into the field of vision, object detection models based on Transfomer have been widely studied. For small object detection, the global context modeling capability of self attention mechanism has significant advantages. Transfomer provides another paradigm for small object detection.

Hower existing small objects detection algorithms still fall short of meeting the rigorous demands of practical scenarios, with their limitations mainly manifested in two key aspects: insufficient feature retention capability, which leads to the easy loss of fine-grained details of small objects during feature extraction; and inadequate utilization of global contextual information, making it hard to distinguish small objects from complex backgrounds effectively.

To address these issues, this paper proposes SOD-DETR, a small object detection algorithm based on RT-DETR. First, a small target feature retention network is presented, which adopts a dual-branch structure to extract spatial and frequency features separately. Fusing these two features enhances representation and improves small target feature retention. For frequency feature extraction, combining frequency band separation and attention mechanism enables adaptive learning of key frequency features, boosting processing efficiency. Then, a subpixel-excitation module is constructed, integrating subpixel processing, Focus strategy and squeeze-and-excitation mechanism. It preserves subpixel information in the channel dimension and guides focus on key channels, enhancing small object detection performance. Finally, improving the hybrid encoding strategy further strengthens small target extraction capability. Experimental results on the VisDrone, TT100K and UAVDT-2024-DET datasets demonstrate that SOD-DETR significantly outperforms state-of-the-art methods, with notable accuracy improvements in detecting objects of different scales. Specifically, the key metrics mAP50 and mAP50-95 are increased by up to 6% and 5.1%, respectively.

This work is jointly supported by the National Natural Science Foundation of China (No. 61702347), the Central Guidance on Local Science and Technology Development Fund (No. 226Z0501G), the Major Science and Technology Projects of Universities in Hebei Province (No. 2512602307A), the Key Program of Hebei Natural Science Foundation (No. F2024210008), the Key Sciences Promotion Project of Hebei Academy of Sciences (No. 25A03).