文章编号：2095-2163(2022)02-0073-06 中图分类号：TP751.1 文献标志码：A

一种改进的 Mask R-CNN 卫星影像船舶尾迹检测方法

吴荣峰，唐希源
（南京理工大学 电子工程与光电技术学院，南京 210094）

摘 要：为提高运用深度学习算法进行船舶尾迹检测的准确度，本文提出了一种改进的 Mask R-CNN 网络结构，在传统的 Mask R-CNN 深度学习算法结构基础上，引入平衡特征金字塔串联结构，增强特征的融合和可辨识性，并引入 GCNet 提高特征提取能力，改善船舶尾迹的检测效果。以 landsat8 卫星遥感图像为数据集，通过在不同背景中的船舶航行图像下，比较改进结构与一般 Mask R-CNN 的检测效果，说明在相同条件下，改进结构较传统的 Mask R-CNN 算法能够得到更好的检测效果。
关键词：Mask R-CNN；平衡特征金字塔；GCNet；卫星遥感图像；船舶尾迹检测

An improved satellite image ship wake detection method based on Mask R-CNN

WU Rongfeng, TANG Xiyuan
(School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing 210094, China)

【Abstract】 In order to improve the accuracy of ship wake detection using deep learning algorithms, this paper proposes an improved Mask R-CNN network structure. Based on the structure of the traditional Mask R-CNN, a series structure of balanced feature pyramids is introduced to enhance the fusion and recognizability of features, and GCNet is incorporated to improve the feature extraction ability. Comparison of detection effects on landset8 dataset show that our proposed improved method achieves better results than the traditional Mask R-CNN.
【Key words】 Mask R-CNN; balanced feature pyramids; GCNet; satellite remote sensing image; ship wake detection

0 引 言

中国海域面积辽阔，使用卫星遥感技术实时监测海面船舶对国防事业、海运贸易等都具有十分重要的意义。为了尽可能多地获取海面船舶信息，往往会选择超广角的卫星，然而在这类卫星的遥感图像上，船舶往往表现为很小的白色点状，难于识别，而海面复杂的环境状况又会进一步加大识别的难度，基于这种情况，转向识别船舶的尾迹。船舶尾迹的目标范围远大于船舶，且尾迹在遥感图像上的灰度变化和周边的海域有着明显的区别，大大降低了目标检测的难度。此外，尾迹还能提供船只的航速以及航向方向等信息[1]。

传统的船舶尾迹检测算法往往依赖于人为的特征提取，耗时费力，且这类方法的鲁棒性和泛化能力较差，不利于系统自动地识别目标。近年来，深度学习技术不断地发展和完善，逐渐被引入到遥感图像目标检测与识别领域，并且取得了很好的效果[2]。基于此，本文提出了一种基于改进的 Mask R-CNN 算法的船舶尾迹检测技术。

Mask R-CNN 是由 Faster R-CNN 改进而来，用于实例分割的目标检测算法，可以在一个网络中同时做目标检测和实例分割，其在原来 Faster R-CNN 的基础上把 ROI Pooling 层改为 ROI Align，使得区域划分更加精准，此外还额外引入了一个 Mask 层用于实例的分割。

由于遥感卫星图像往往图像不清晰，噪声很大。为了更好地实现检测，本文在原有的 Mask R-CNN 算法的基础上做出了两点改进：

(1)在原有的特征金字塔网络(FPN)结构上引入平衡特征金字塔(BFP)串联结构，以增强图像特征信息的融合，降低原图的噪声，增强目标的可辨识性；

(2)使用 ResNet50 作为主干网络，在主干网络上引入 GCNet，增加特征的提取能力。

实验结果表明，经过改进之后的 Mask R-CNN 对于船舶尾迹的目标检测能力明显提升。

1 Mask R-CNN 简述

Mask R-CNN 是一种实例分割的深度学习神经网络，在目标检测领域有着十分优秀的表现，很适合遥感图像的检测。主干网络与特征金字塔网络层

作者简介：吴荣峰(1999-)，男，本科生，主要研究方向：信号与信息处理；唐希源(1999-)，男，本科生，主要研究方向：微电子科学与工程。
收稿日期：2021-10-16

(Backbone + FPN)、区域建议网络层(RPN)、RoI Align 层、卷积层(CONV)、边框回归支路(class)、边框分类支路(box),以及一条并行的 Mask 支路[3],如图 1 所示。

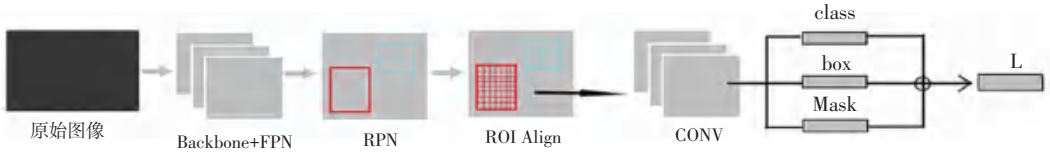

图 1 Mask R-CNN 结构示意图

Fig. 1 Structure diagram of Mask R-CNN

ROI Align 是 Mask R-CNN 的第一个重大改进,明显改进了量化误差的影响。在 Faster R-CNN 当中,RoI Pooling 引入了两次量化误差,一次是在原始图像映射到特征图的过程中,图像尺寸的浮点数取整;另一次是对特征图进行最邻近插值[4]。整个过程的两次取整操作,会给坐标引入很大的误差。为了解决该问题,文献[4]提出了 RoI Align 使用双线性插值方法,利用原图中虚拟点四周的 4 个像素点的值,来共同决定目标图中的一个像素值,这样就可以将虚拟点对应的像素值估计出来。

Mask R-CNN 的另一个改进是在原有的损失函数中引入了 Mask 预测损失函数 L_{MASK},损失函数如式(1):

$$L = L_{CLS} + L_{BOX} + L_{MASK} \tag{1}$$

其中,L_{CLS}、L_{BOX} 分别为类别、位置预测的损失函数。

对于 *Mask* 支路,每个 *ROI* 的输出维度是 $m \times m \times k$,$m \times m$ 表示 *Mask* 的大小,k 代表类别数。得到预测 *Mask* 后,对 *Mask* 的每一个像素点求 *Sigmoid* 函数值,并把结果作为 L_{MASK} 的输入。虽然会有 k 个 *Mask*,但在计算时只有对应类别的 *Mask* 才有效,其他的 *Mask* 不会对 L_{MASK} 造成影响。

2 平衡特征金字塔(Balanced Feature Pyramid,BFP)

在遥感图像中,船舶尾迹目标的长短大小往往很不一致,并且由于分辨率低,目标的辨识度很低,图像噪声也很大,即使是依靠人眼也很难快速确定目标,因此需要加工处理,加强特征,提高辨识度,而平衡特征金字塔结构可以很好地满足这一要求。

BFP 结构旨在解决特征层信息的不平衡,以更加高效地利用不同尺度各自的特征[5]。传统的 FPN 是一种致力于解决特征融合问题的结构,使用自下而上后再自上而下的结构,低层的特征图包含了更多的位置细节信息,有利于小物体的目标检测,而高层次的特征图则是包含了更多的语义信息,更加适合做大尺度物体的识别,通过两者的组合来进行不同尺度物体的识别[6]。但这种结构更多地关注于相邻层的关系,忽略非相邻层间的依赖关系,而非相邻层的依赖关系在目标识别当中往往起着重要的作用。

平衡特征金字塔结构很好地解决了这一问题,同时获取并聚合了来自不同层级的特征,使得高层语义特征和底层位置细节等信息同时汇聚到一起,并通过使用嵌入式高斯 Non-Local 注意力模块进一步精炼了特征,提高了目标的可辨识度。

BFP 的结构示意图,如图 2 所示,包括调整大小、融合、精炼和增强 4 个步骤。

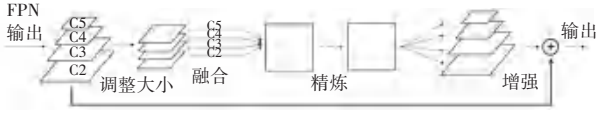

图 2 BFP 结构示意图

Fig. 2 Structure diagram of BFP

(1)调整大小。在 FPN 结构输出的特征图中,不同层次的特征图大小不一,为了便于后续的整合,需要调整为同一尺寸。比如,以 C4 层作为目标,对于更大的 C3 和 C2,使用最大池化(Max Pooling)方法进行缩小,对于更小的 C5 层,则可以使用双线性插值的方法放大到 C4 的尺寸。

(2)融合。把几张同尺寸特征图相互叠加,并求平均值即可。

(3)精炼。使用嵌入式高斯 Non-Local 注意力模块进行特征精炼,通过建立图像上两个有一定距离的像素之间的联系来增强识别的效果,同时基于传统数字图像处理中的非局部均值理论,该方法还可以明显降低图像中的噪声[7]。该方法有比卷积更好的稳定性,其关键公式如式(2)所示。

$$y_i = \frac{1}{C(x)} \sum_{\forall j} f(x_i, x_j)\, g(x_j) \tag{2}$$

输入信号 x_i 代表目标图像,x_j 是所有特征可能与 x_i 相似的图像,两者大小相等。通过 f 函数计算得到两者的关联系数,g 函数代表位置 j 处的输入

信号，之后以 f 函数为权重进行加权求和，$C(x)$ 代表归一化系数。相关的函数表达式如式(3)～式(7)所示。

$$f(x_i,x_j)=e^{\theta(x_i)^{\mathrm{T}}\varphi(x_j)} \tag{3}$$

$$g(x_j)=W_g\,x_j \tag{4}$$

$$\phi(x_j)=W_\phi\,x_j \tag{5}$$

$$\theta(x_i)=W_\theta\,x_i \tag{6}$$

$$C(x)=\sum_{\forall j}f(x_i,x_j) \tag{7}$$

最后，需要把该结构插入到原有的网络中，并且不能破坏初始信息，这里需要增加一个残差链接，其表达式如式(8)所示。

$$z_i=W_z\,y_i+x_i \tag{8}$$

Non-Local 模块的结构示意图，如图 3 所示。

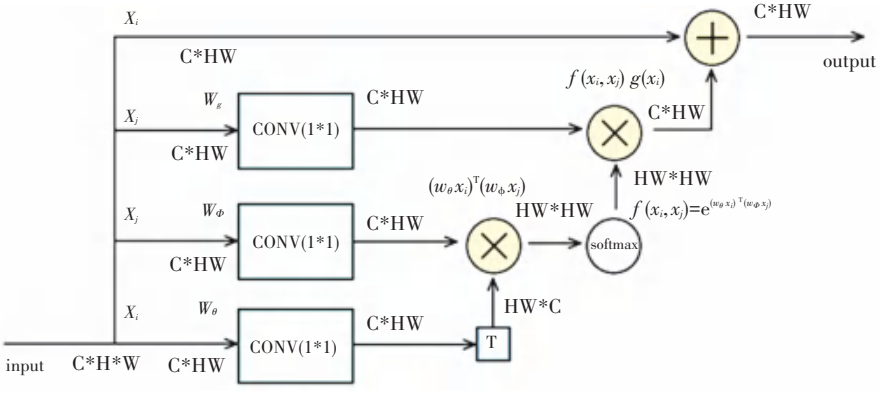

图 3　Non-Local 模块的结构示意图

Fig. 3　Structure diagram of non-local block

(4)增强。把已经融合的特征图恢复到原有大小。对精炼后的图像使用双线性插值操作可以恢复到 C2，C3 大小，使用最大池化操作则恢复到 C5 的大小，对于 C4 大小的输出则不需要操作。恢复完成后，再把其和原始的 C2、C3、C4、C5 相互叠加后输出。

综上所述，通过 BFP 的操作实现了不同特征层的信息融合，并加强了目标的特征，增加了可辨识度，对低分辨率的遥感图像识别十分关键。

3　GCNet 模块

传统的卷积神经网络通过图像的一部分作为卷积核，在图像上以滑窗的形式不断进行卷积，直至整幅图像均以该卷积核进行过卷积操作，后对特征图进行池化。然而这样的操作产生了一个问题，当另外有相似或关系密切的目标距离卷积核所在位置较远，那么该卷积核只能观察到其卷积范围内的部分图像，无法提高长距离依赖的检测能力。引入 GCNet 的目的正是提高长距离依赖特征提取能力。

GCNet 由 Non-local 与 SE 两大模块组成。

Non-local 操作是为提高长距离依赖，某一输入信号处的响应是其他所有与其大小相等的位置特征权重和，将每一个信号与其他所有的信号相关联，实现 Non-local 的思想。2019 年 Yue Cao 等人[8]指出，所选取的注意力 x_i 对最终的识别效果只能产生很小的影响，对于每个 x_i 均计算其注意力分布是很浪费计算资源的行为，因此，在 GCNet 当中，Non-local 模块被进一步简化。

由于不再对 x_i 进行操作，因此传统的 non-local 模块中的 W_θ 路被移除，不再加入该卷积模块，以节约计算资源。将 W_g 移至 y_i 的乘法运算之后，单独生成一个模块称为 Transform，虽然会牺牲一定的准确度，但是会大大节省计算的成本，提高运算的速度[8]。

简化的 Non-local 模块结构如图 4 所示，可以将整个简化 Non-local 模块划分为上下文建模(Context Modeling)、变换(Transform)以及融合(Fusion)3 个部分。

其数学模型如式(9)所示。

$$z_i=x_i+W_g\sum_{\forall j}\frac{e^{W_\Phi x_j}}{\sum\limits_{\forall m}e^{W_\Phi x_m}}x_j \tag{9}$$

其中，x_i、x_j 表示输入信号，W_g，W_Φ 表示卷积因子。

在简化的 non-local 模块的操作中，将 W_g 移至乘法运算之后，在显著减少运算量的同时，会降低准确度，为了弥补这个问题从而引入了第二个模块 SE 模块[9]，其结构示意图如图 5 所示。

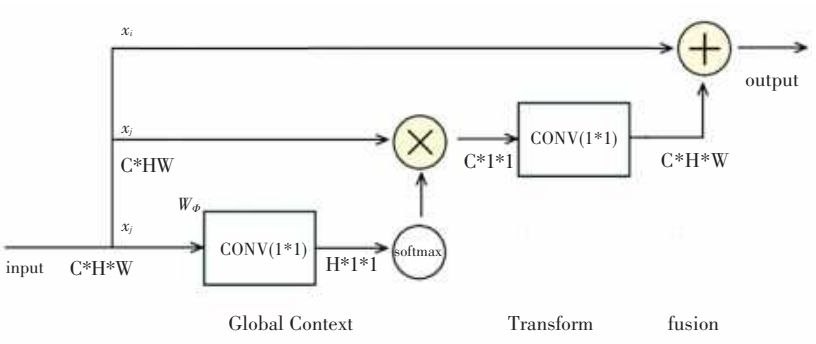

图 4 简化的 non-local 模块结构示意图

Fig. 4 Structure diagram of simplified non-local block

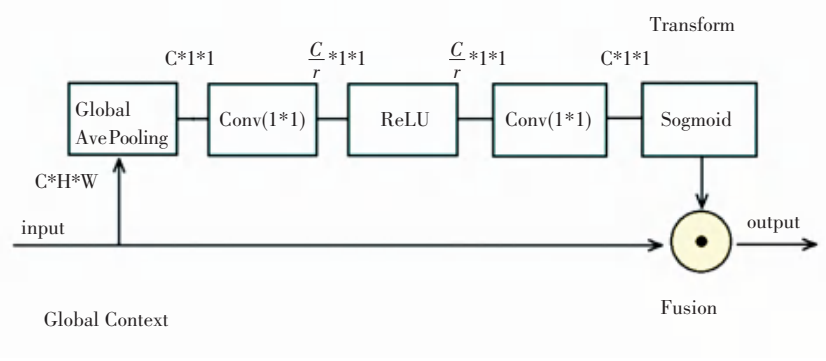

图 5 SE 模块的结构示意图

Fig. 5 Structure diagram of SE block

SE 模块的上支路会先将输入的图像做一次全局平均池化(Global Average Pooling),后接 bottleneck 结构,即先使用卷积降低维度,之后做一次 ReLU 非线性激活,再做一次卷积恢复维度,最后通过 *sigmoid* 产生归一化权重。上支路最后和恒等映射进行乘积操作,形成 SE 模块的输出。SE 模块的显著特点便是通过 bottleneck 结构减小参数量,这是 GCNet 引入 SE 的重要原因。

融合简化后的 Non-local 模块以及 SE 模块,最终的 GCNet 模块结构如图 6 所示。层标准化(Layer Normalization, LayerNorm)的作用是改善 bottleneck 结构难以优化的问题,提高模型泛化能力,同时可以弥补传统神经网络不断以相同函数堆叠导致提取的特征缺少多样性的问题。

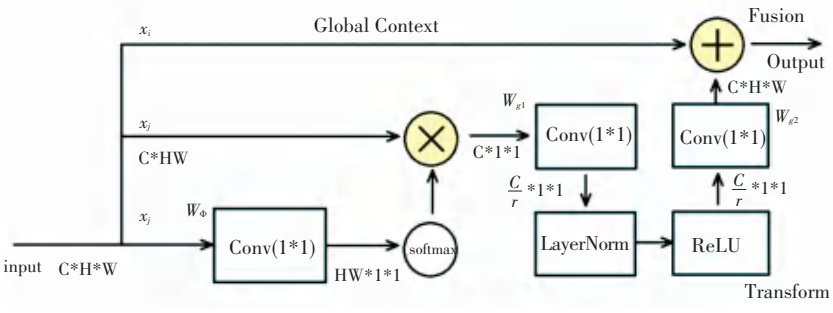

图 6 Gcnet 模块结构示意图

Fig. 6 Structure diagram of GCnet block

GCNet 的数学表达如式(10)所示:

$$z_i = x_i + W_{g2}ReLU(LN(W_{g1}\sum_{\forall j}\frac{e^{W_\Phi x_j}}{\sum_{\forall m}e^{W_\Phi x_m}}x_j)) \quad (10)$$

其中，*ReLU* 即 ReLU 非线性激活函数，*LN* 即层标准化。

在原来的简化的 non-local 模块的变换部分，融合了 SE 模块中 bottleneck 结构，并使用层标准化运算解决优化问题，而上下文建模部分保留了简化的 Non-local 模块的结构，这样即能够得到 Non-local 适应特征之间长距离的依赖的性能，又能像 SE 模块一般减少计算量，解决提取特征多样性的丢失问题，提高了检测的准确率。

4　融合 BFP+GCNet 的 Mask R-CNN 网络整体结构

融合 BFP+GCNet 的 Mask R-CNN 网络整体结构如图 7 所示。在主干网络（Backbone）部分选用 Resnet50，并在其中引入了 GCNet 结构，以加强特征的提取能力；在 FPN 和 RPN 之间增加了串联的 BFP 结构，用于提高特征的融合，增加目标的可辨识性。

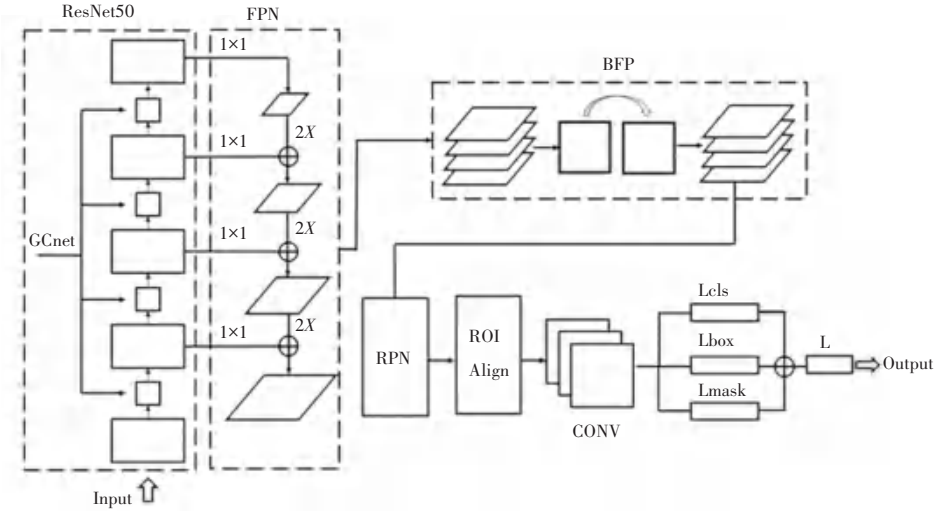

图 7　融合 BFP+GCNet 的 Mask R-CNN 网络整体结构

Fig. 7　The overall structure of Mask R-CNN network integrated with BFP+GCNet

5　实验方法与实验结果

5.1　实验环境

硬件环境：配有两块 NVIDIA RTX 2080 Ti 显卡的计算机；

软件环境：Ubuntu 18 操作系统，Python 语言编程实现算法网络，使用 PyTorch 学习框架，mmdetection 框架；

训练集：64 张图片进行 mosaic 混合，大图裁剪拼接，以提高背景与场景特征多样性，提升数据质量与数据集泛化性，每轮训练取所有图片的 80%，重复十次，共计十二轮训练；

测试集：64 张图片，大小均为 1 400×1 000。

5.2　评价指标

识别对象分别为船只和尾迹，根据测试程序返回的指标，选取各检测对象“框选”和“分割”的平均准确度进行评价，评价的对照组为传统 Mask R-CNN，实验组为仅融合 BFP 的 Mask R-CNN、仅融合 GCNet 的 Mask R-CNN、融合 BFP+GCNet 的 Mask R-CNN，测试结果见表 1。

表 1　测试集输出的模型准确度测试结果

Tab. 1　Accuracy on test dataset

测试模型	船只框选准确度	尾迹框选准确度	船只分割准确度	尾迹分割准确度
融合 BFP+GCNet 的 Mask R-CNN	0.491	0.913	0.338	0.355
仅融合 GCNet 的 Mask R-CNN	0.449	0.895	0.332	0.199
仅融合 BFP 的 Mask R-CNN	0.358	0.786	0.264	0.192
传统 Mask R-CNN	0.262	0.697	0.199	0.116

由表 1 可以看出，相较于传统的 Mask R-CNN，不论是仅采取一个改进措施或是将两项改进结合，本文所述的改进措施具有显著效果。同时，对尾迹的标定准确度比船只都高，说明针对尾迹对船只的位置进行勘测是可行的。

5.3　检测效果

本文采用的数据集来自于 landsat8 遥感影像，实际检测效果如图 8 所示。由于目标物较为模糊，且图像的噪声大，对需要检测的目标存在较大干扰，

需要通过对已有的卫星影像进行裁剪,放大小目标的精度,并进行 mosaic 融合,以提升检测数据的质量,并扩充数据集。图片经过预处理后,进入神经网络中的数据质量得到提升,从而使得识别结果较为清晰,基本能够正确地标注出船只与尾迹所在的位置。

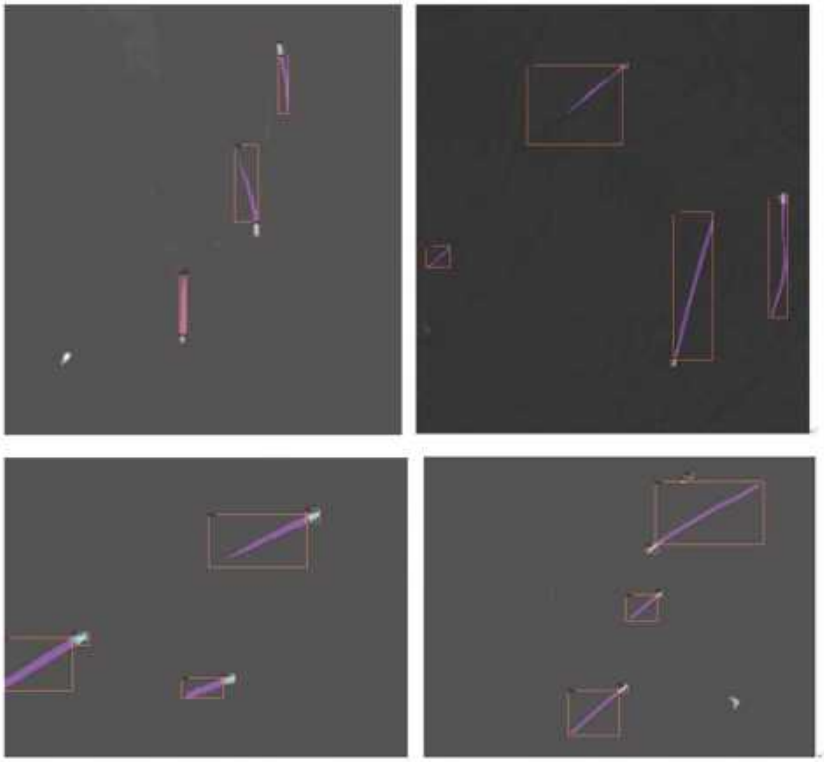

图 8 检测效果

Fig. 8 Detection effects

6 结束语

本文讨论了一种改进的 Mask R-CNN 的结构,该结构做出了两个改进:一在骨干网络 Resnet50 中加入 GCNet 全局注意力模块;二在 FPN 特征提取网络中引入 BFP 串联结构。首先,从理论上证明这样的改进结构能够使 Mask R-CNN 的检测准确率得以提升;利用实验分别测试融合了 BFP/GCNet/BFP+GCnet 改进的 Mask R-CNN 以及对照组(传统 Mask R-CNN)的检测准确率,最终证明 BFP+GCNet 的改进结构明显比其他模型的检测能力更好,对于尾迹的检测比对于船只的检测准确率更高,说明了融合 BFP+GCNet 的 Mask R-CNN 能够更好地适应船舶尾迹的检测任务。

参考文献

[1] 巩彪, 黄韦艮. SAR 图像船只尾迹检测研究综述[J]. 遥感技术与应用, 2012, 27(6):829-836.

[2] 陈琳. 基于深度学习的 SAR 图像目标识别与分类[D]. 山东:山东大学,2021.

[3] HE K, GKIOXARI G, P DOLLÁR, et al. Mask R-CNN[C]// Proceedings of the IEEE international conference on computer vision. 2017: 2961-2969.

[4] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6):1137-1149.

[5] PANG J, CHEN K, SHI J, et al. Libra R-CNN: Towards Balanced Learning for Object Detection[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020:821-830.

[6] LIN T Y, DOLLAR P, GIRSHICK R, et al. Feature Pyramid Networks for Object Detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017: 936-944.

[7] WANG X, GIRSHICK R, GUPTA A, et al. Non-local neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).2018: 7794-7803.

[8] CAO Y, XU J, LIN S, et al. GCNet: Non-Local Networks Meet Squeeze-Excitation Networks and Beyond[C]//2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), 2020. 1971-1980.

[9] JIE H, LI S, GANG S, et al. Squeeze-and-Excitation Networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018:7132-7141.