(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 113469073 B
(45)授权公告日 2024.02.20

(21)申请号 202110765081.4
(22)申请日 2021.07.06
(65)同一申请的已公布的文献号
申请公布号 CN 113469073 A
(43)申请公布日 2021.10.01
(73)专利权人 西安电子科技大学
地址 710071 陕西省西安市雁塔区太白南路2号
(72)发明人 陈潇钰 侯彪 焦李成 张丹 马文萍 马晶晶 王爽
(74)专利代理机构 西安通大专利代理有限责任公司 61200
专利代理师 高博
(51)Int.Cl.
G06V 20/10(2022.01)
G06V 10/25(2022.01)
G06V 10/774(2022.01)
G06V 10/82(2022.01)
G06N 3/0464(2023.01)

(56)对比文件
CN 110929593 A,2020.03.27
CN 111259740 A,2020.06.09
CN 112819771 A,2021.05.18
CN 112308019 A,2021.02.02
CN 112464846 A,2021.03.09
CN 112686180 A,2021.04.20
CN 110909667 A,2020.03.24
梁文楷.基于深度数据特征与统计特征学习的高分辨 率SAR图像分类.中国优秀博士学位论文全文数据库.2023,全文.
肖恩.基于深度学习的SAR 车辆目标分类与识别.中国优秀硕士学位论文全文数据库.2021,全文.
(续)

审查员 杨晓华

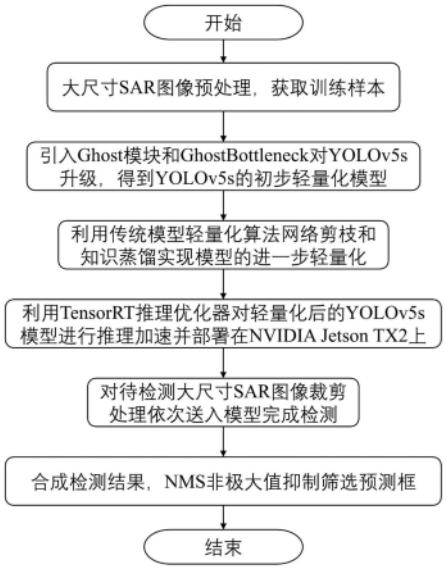
权利要求书3页 说明书13页 附图6页

(54)发明名称

一种基于轻量级深度学习的SAR图像舰船检测方法及系统

(57)摘要

本发明公开了一种基于轻量级深度学习的SAR图像舰船检测方法及系统,对大尺寸SAR图像预处理,选取训练样本;引入Ghost模块和GhostBottleneck对YOLOv5s升级,得到YOLOv5s的初步轻量化模型;在初步轻量化模型的基础上利用传统模型轻量化算法网络剪枝和知识蒸馏实现模型的进一步轻量化;利用TensorRT推理优化器对轻量化后的YOLOv5s模型进行推理加速并部署在NVIDIA Jetson TX2上;将待检测大尺寸SAR图像裁剪处理依次送入模型完成检测;合成检测结果,并在最终的大尺寸SAR图像上使用NMS非极大值抑制筛选预测框。在满足可接受精度损失的前提下,压缩模型的参数量和浮点运算量,提升检测速度。



CN 113469073 B

[转续页]

[接上页]

(56)对比文件

Rongfang Wanga,Fan Dinga.A Light-Weighted Convolutional Neural Network for Bitemporal SAR Image Change Detection.IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.2020,全文.

Xi Yang , Member, IEEE, Jianan Zhang, Chengzeng Chen, and Dong Yang.An Efficient and Lightweight CNN Model With Soft Quantification for Ship Detection in SAR Images.IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.2022,全文.

Xiaowo Xu, Xiaoling Zhang * and Tianwen Zhang.Lite-YOLOv5: A Lightweight Deep Learning Detector for On-Board Ship Detection in Large-Scene Sentinel-1 SAR Images.Remote Sens.2022,全文.

Xuemeng Zhao,Yinglei Song,Sanxia Shi, Shunxin Li.Improving YOLOv5n for lightweight ship target detection.IEEE 3rd International Conference on Computer Systems.2023,全文.

Hang Yu,Suiping Zhou.VS-LSDet: A Multiscale Ship Detector for Spaceborne SAR Images Based on Visual Saliency and Lightweight CNN.IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING.2023,全文.

1.一种基于轻量级深度学习的SAR图像舰船检测方法,其特征在于,包括以下步骤:

S1、对大尺寸SAR图像进行预处理,选取包含目标信息的子图作为训练样本;

S2、引入Ghost模块和GhostBottleneck对YOLOv5s模型进行升级,得到初步轻量化的YOLOv5s模型,使用步骤S1选取的训练样本对YOLOv5s模型进行训练,具体为:

S201、使用Ghost模块和GhostBottleneck替换YOLOv5s模型主干网络中的卷积模块和瓶颈模块,利用Ghost模块和GhostBottleneck对YOLOv5s模型进行升级;

S202、将宽度乘数调整为0.15,深度乘数调整为0.35,网络层数减少至212层,得到初步轻量化的YOLOv5s模型;

S3、对步骤S2训练后得到的YOLOv5s模型进行蒸馏处理,然后进行稀疏化训练和剪枝处理,对剪枝处理后的YOLOv5s模型进行微调训练,具体为:

S301、以YOLOv5m作为教师模型,用L2 loss作为蒸馏基础函数,损失中的蒸馏dist平衡系数选择为1,蒸馏训练100个epoch;

S302、在正常训练得到过度参数化模型后,设定稀疏参数为6e-4,稀疏化训练对BN层gamma参数进行L1正则化,产生稀疏权值矩阵作为评价神经元贡献大小的标准,根据30%稀疏率确定阈值,剪去小于阈值的层以及对应层的依赖层,若对应层中所有通道都需要移除,保留最大通道;

S303、在步骤S302剪枝处理完成后,对步骤S302得到的模型继续训练50个epoch,通过微调训练学习稀疏连接的最终权重;

S4、利用TensorRT推理优化器对步骤S3微调训练后的YOLOv5s模型进行推理加速,并部署在NVIDIA Jetson TX2上;

S5、对待检测的SAR图像进行裁剪处理后依次送入步骤S4部署在NVIDIA Jetson TX2上的YOLOv5s模型进行检测,得到对应的子图检测结果,具体为:

S501、将待检测图片的子图送入训练好的轻量化YOLOv5s模型进行检测前,若待检测图片的子图不满足模型对于图片大小的要求,则进行自适应图片缩放,将子图送入特征提取网络,得到S×S大小的特征图,将输入图像分为S×S的小格;

S502、每个网格使用逻辑回归预测B个边界框,若预测边界框中心在网格单元内,则网格单元的B个边界框负责对目标进行分类和边框预测,得到每个网格对B个边界框的预测结果,输出边界框的位置信息、表明网格是否包含目标的置信度和C个类别的概率信息,每个边界框预测得到t_x、t_y、t_w、t_h、t_o,t_x、t_y是边界框中心坐标相对于当前网格单元的偏移值;使用逻辑激活对t_x和t_y进行归一化处理,将值限制在0～1内,t_w、t_h是边界框宽和高的尺度缩放,t_o是置信度confidence;

S503、采用特征金字塔网络下采样自顶向下传达强语义特征和路径聚合网络上采样自底向上传达强定位特征融合三个尺度分别的检测结果;针对图片输入大小为960×960,输出特征图分别为120×120,60×60,30×30,分别为8倍、16倍和32倍下采样的结果;

S6、对步骤S5得到的子图检测结果进行拼接,并在最终的大尺寸SAR图像上使用NMS非极大值抑制筛选预测框,根据筛选后的预测框数值在原始大尺寸图像上绘制预测框并标记类别,实现SAR图像舰船检测。

2.根据权利要求1所述的方法,其特征在于,步骤S1具体为:

S101、对5张单通道TIF图像Img10K和31张单通道TIFF图像AIR-SARShip-1.0以50%的

重合率进行切块,得到大尺寸遥感图像的子图;

S102、将1000张8位JPG图像SAR-train-int进行放大;

S103、将步骤S101得到的Img-10K、AIR-SARShip-1.0和步骤S102得到的SAR-train-int图像格式统一为8位单通道TIF图像,得到包括2551张图片的数据集,划分2351张作为训练样本,200张作为验证样本;

S104、利用Mosaic数据增强算法对步骤S103的训练样本进行随机操作,对训练样本中每四张图片以随机缩放、随机裁剪、随机排布的方式进行拼接。

3.根据权利要求1所述的方法,其特征在于,步骤S4中,TensorRT推理优化器进行部署包括Build阶段和Deploymeng阶段,具体为:

S401、Build阶段优化用Pytorch训练模型得到*.pt文件,并将其转化为onnx模型,然后在TensorRT中加载onnx模型,并转换成TensorRT模型;再将TensorRT模型序列化存储到磁盘或内存中,称为plan文件;

S402、Deployment阶段进行轻量化YOLOv5模型的部署,先对步骤S401获得的plan文件反序列化,创建runtime引擎,完成前向推理过程。

4.根据权利要求1所述的方法,其特征在于,步骤S502中,根据每个边界框预测得到的5个值得到预测的边界框在整个特征图中的中心点坐标b_x、b_y和长宽b_w、b_h如下:

$b_x=\sigma(t_x)+c_x$

$b_y=\sigma(t_y)+c_y$

$$b_w = p_w e^{t_w}$$

$$b_h = p_h e^{t_h}$$

其中,σ函数为逻辑激活,c_x和c_y分别为当前网格单元相对于特征图左上角的距离,p_w和p_h为先验框的长和宽。

5.根据权利要求4所述的方法,其特征在于,限制坐标偏移和置信度在0~1内,当真实框落在网格单元内时,Pr(object)为1,否则,Pr(object)为0,网格单元在包含目标的条件下属于某个类别的概率$Pr(class_i|object)$表示为

$$\Pr(class_i|object)*\Pr(object)*IOU_{pred}^{truth}=\Pr(class_i)*IOU_{pred}^{truth}$$

其中,IOU_{pred}^{truth}为真实框与预测框的交并比,$Pr(class_i)$为某个单元格内目标对应类别的概率。

6.根据权利要求1所述的方法,其特征在于,步骤S6具体为:

S601、通过目标在子图上的位置信息和子图在大图上的相对位置推算目标在大图上的位置信息;

S602、针对某一类别,设置NMS阈值为0.65,选择置信度最高的边界框,根据边界框与其他边界框的DIOU值滤除超过NMS阈值的所有边界框,筛选完预测框后根据保留的预测框进行画框,完成大尺寸SAR图像的舰船检测。

7.一种基于轻量级深度学习的SAR图像舰船检测系统,其特征在于,基于权利要求1所述的方法,包括:

数据模块,对大尺寸SAR图像进行预处理,选取包含目标信息的子图作为训练样本;

处理模块,引入Ghost模块和GhostBottleneck对YOLOv5s模型进行升级,得到初步轻量

化的YOLOv5s模型,使用数据模块选取的训练样本对YOLOv5s模型进行训练;

微调模块,对处理模块训练后得到的YOLOv5s模型进行蒸馏,然后进行稀疏化训练和剪枝,对剪枝后的YOLOv5s模型进行微调训练;

推理模块,利用TensorRT推理优化器对微调模块微调训练后的YOLOv5s模型进行推理加速,并部署在NVIDIA Jetson TX2上;

检测模块,对待检测的SAR图像进行裁剪处理后依次送入推理模块部署在NVIDIA Jetson TX2上的YOLOv5s模型进行检测,得到对应的子图检测结果;

去除模块,对检测模块得到的子图检测结果进行拼接,并在最终的大尺寸SAR图像上使用NMS非极大值抑制筛选预测框,根据筛选后的预测框数值在原始大尺寸图像上绘制预测框并标记类别,实现大尺寸SAR图像舰船检测。

一种基于轻量级深度学习的SAR图像舰船检测方法及系统

技术领域

[0001] 本发明属于计算机视觉技术领域,具体涉及一种基于轻量级深度学习的SAR图像舰船检测方法及系统。

背景技术

[0002] 2012年Alexnet横空出世,在计算机领域掀起一股深度卷积神经网络的应用热潮。更深的模型,往往意味着模型具有更好的非线性表达能力,可以完成更加复杂的变换,从而可以拟合更加复杂的特征。基于这样一个假设,深度卷积神经网络朝着越来越深和越来越宽的方向发展,虽然在各类任务中表现出愈发出色的性能,但随之而来的是网络模型的体量变得愈发庞大,这与当前移动端各类嵌入式设备硬件条件相悖,深度神经网络研究的各项成果只能束之高阁,无法落地。与深度神经网络发展速度相当的是各类移动端设备,这些设备通常并没有图形处理单元(Graphic Processing Unit,GPU)的高性能计算集群,只有中央处理单元(Central Processing Unit,CPU)完成计算任务,并不能为现阶段用于提取表达能力更强深度特征的大型卷积神经网络提供与之匹配的存储空间和算力条件,这严重阻碍了深度卷积升级网络在便携式设备上的发展与应用。为大力推动人工智能产业落地,学术界和工业界大批学者投身网络模型轻量化算法的研究,以提高便携式设备的在图像处理方面的性能和效率。

[0003] 现有使网络轻量化的方法主要可以分为两大类:模型压缩和紧凑模型设计。模型压缩是指针对模型结构以及参数压缩神经网络模型,从而减少模型对于存储设备和计算资源的需求,满足移动端便携式的内存和算力限制要求。模型压缩面向的是网络结构和网络权重的冗余部分,在一定程度上牺牲准确率以换得冗余更少,速度更快、更为精简的模型。目前已提出的算法有网络剪枝(NetWork Pruning)、模型量化(Model Quantization)、二值化方法(Binarization Method)、低秩分解(Low-rank Decomposition)和知识蒸馏(Knowledge Distillation)等。由于深度神经网络各层的冗余程度不同,不能一概而论,传统的模型压缩算法往往对具体模型过拟合,而针对每个模型若都人工探索适应其各层冗余程度的模型压缩算法,费时费力,这便推动了自动机器学习算法(AutoML)的发展,它自动学习探索到局部最优的网络超参数和架构,避免了人工的干扰,同时可以推广到各个模型。基于AutoML,西安交通大学和谷歌研究团队提出自动模型压缩算法(AMC),将强化学习引入模型压缩算法中,相较于传统基于规则的压缩策略,在保持网络模型性能的情况下压缩比更高。近几年也提出一系列紧凑模型像Xception、MobileNetV1、MobileNetV2、MobileNetV3、ShuffleNet、ShuffleNetv2等。这些网络模型通常从减少卷积核的冗余、压缩通道数、用高效卷积模块替代传统卷积的角度出发。通过在卷积层中多使用小卷积核来减少卷积核冗余,有效减少网络参数。SqueezeNet中提出的Fire模块,由一个squeeze层和一个expand层组成,利用减少squeeze层中1×1卷积核的数目进而减少3×3卷积核的输入通道数。MobilenetV1中利用深度可分离卷积(Depthwise Separable Convolution)将普通卷积分解为深度卷积(depthwise convolution)和点卷积(pointwise convolution);ShuffleNet

进一步提出了置乱(Shuffle)操作和分组逐点卷积(group pointwise convolution),对特征进行重新排列,使得特征信息在各个通道分组内互相流通;MobilenetV2提出反残差结构(Inverted residual block)、MobilenetV3使用神经网络架构搜索技术(NAS),引入SE(squeeze and excitation)模块,选用H-swish激活函数对网络结构进一步压缩模型。这些优秀的轻量级网络模型在损失少量精度的情况下在模型压缩和加速上已经获得不错的成果。

[0004] 目标检测又称目标类别检测或目标分类检测,返回图像中感兴趣目标的类别信息和位置信息。在近二十年内都是计算机视觉和数字图像处理领域的研究热点内容。2012年Alexnet提出之前都是基于传统手工特征的目标检测方法,如为人熟知的:V-J检测、HOG检测、结合Bounding box回归的DPM检测。2012年之后随着卷积神经网络兴起和GPU性能指数级增长,深度学习迎来爆发式发展,目标检测也走入了深度学习时期,根据算法是否需要生成预选框,基于深度学习的目标检测算法又可分为单阶段(One-stage)检测算法和两阶段(Two-stage)检测算法。单阶段检测算法中的代表网络有YOLO系列、SSD、RetinaNet。其主要特点是检测精度低,检测速度快。两阶段检测算法的典型网络有R-CNN、SPP-Net、Fast R-CNN和Faster R-CNN。与单阶段检测算法不同,两阶段检测检测精度高但时间成本高。截止目前,性能最卓越的目标检测算法还是难以与人眼检测媲美。当前的目标检测仍旧面临着诸多难题。针对高准确率的要求,同类物体纹理、颜色、材质导致的多样性;目标实例姿态、形变的多样性;采样过程环境的差异性以及图像噪声的影响都影响算法对类内形变鲁棒性。至于类与类的可区分性,这通常由类间的相似性和类的多样性决定。针对在时间和内存占用高效性的要求,自然界类别的丰富性、目标检测任务包含定位及分类的双重性以及图像数据体积愈发庞大,这些都给当前的目标检测算法提出了更高的要求,也正是各大研究学者跻身的领域。

[0005] 基于大数据的高分影像目标检测一直以来都是遥感图像处理领域炙手可热的研究方向,传统的目标检测识别方法针对遥感图像的海量数据无法自适应调整,并且需要人为设计大量图像特征,在带来极大时间成本的同时,给研究者在专业知识和对数据特征的理解上都提出了极高的要求,而且搜索高效分类器以充分理解数据犹如大海捞针。而深度学习强大的高级(更具抽象和语义意义)特征表示和学习的能力可以为图像中的目标提取提供有效的框架。相关研究包括车辆检测、船舶检测、农作物检测、建筑物等地物检测。

发明内容

[0006] 本发明所要解决的技术问题在于针对上述现有技术中的不足,提供一种基于轻量级深度学习的SAR图像舰船检测方法,将轻量化目标检测网络在嵌入式设备NVIDIA Jetson TX2部署以实现大尺寸SAR图像舰船检测。以目标检测网络YOLOv5为基线网络,结合传统模型压缩算法和Ghost轻量化模块对基线网络实现轻量化。

[0007] 本发明采用以下技术方案:

[0008] 一种基于轻量级深度学习的SAR图像舰船检测方法,包括以下步骤:

[0009] S1、对大尺寸SAR图像进行预处理,选取包含目标信息的子图作为训练样本;

[0010] S2、引入Ghost模块和GhostBottleneck对YOLOv5s模型进行升级,得到初步轻量化的YOLOv5s模型,使用步骤S1选取的训练样本对YOLOv5s模型进行训练;

[0011] S3、对步骤S2训练后得到的YOLOv5s模型进行蒸馏处理,然后进行稀疏化训练和剪枝处理,对剪枝处理后的YOLOv5s模型进行微调训练;

[0012] S4、利用TensorRT推理优化器对步骤S3微调训练后的YOLOv5s模型进行推理加速,并部署在NVIDIA Jetson TX2上;

[0013] S5、对待检测的SAR图像进行裁剪处理后依次送入步骤S4部署在NVIDIA Jetson TX2上的YOLOv5s模型进行检测,得到对应的子图检测结果;

[0014] S6、对步骤S5得到的子图检测结果进行拼接,并在最终的大尺寸SAR图像上使用NMS非极大值抑制筛选预测框,根据筛选后的预测框数值在原始大尺寸图像上绘制预测框并标记类别,实现SAR图像舰船检测。

[0015] 具体的,步骤S1具体为:

[0016] S101、对5张单通道TIF图像Img10K和31张单通道TIFF图像AIR-SARShip-1.0以50%的重合率进行切块,得到大尺寸遥感图像的子图;

[0017] S102、将1000张8位JPG图像SAR-train-int进行放大;

[0018] S103、将步骤S101得到的Img-10K、AIR-SARShip-1.0和步骤S102得到的SAR-train-int图像格式统一为8位单通道TIF图像,得到包括2551张图片的数据集,划分2351张作为训练样本,200张作为验证样本;

[0019] S104、利用Mosaic数据增强算法对步骤S103的训练样本进行随机操作,对训练样本中每四张图片以随机缩放、随机裁剪、随机排布的方式进行拼接。

[0020] 具体的,步骤S2具体为:

[0021] S201、使用Ghost模块和GhostBottleneck替换YOLOv5s模型主干网络中的卷积模块和瓶颈模块,利用Ghost模块和GhostBottleneck对YOLOv5s模型进行升级;

[0022] S202、将宽度乘数调整为0.15,深度乘数调整为0.35,网络层数减少至212层,得到初步轻量化的YOLOv5s模型。

[0023] 具体的,步骤S3具体为:

[0024] S301、以YOLOv5m作为教师模型,用L2 loss作为蒸馏基础函数,损失中的蒸馏dist平衡系数选择为1,蒸馏训练100个epoch;

[0025] S302、在正常训练得到过度参数化模型后,设定稀疏参数为6e-4,稀疏化训练对BN层gamma参数进行L1正则化,产生稀疏权值矩阵作为评价神经元贡献大小的标准,根据30%稀疏率确定阈值,剪去小于阈值的层以及对应层的依赖层,若对应层中所有通道都需要移除,保留最大通道;

[0026] S303、在步骤S302剪枝处理完成后,对步骤S302得到的模型继续训练50个epoch,通过微调训练学习稀疏连接的最终权重。

[0027] 具体的,步骤S4中,TensorRT推理优化器进行部署包括Build阶段和Deploymeng阶段,具体为:

[0028] S401、Build阶段优化用Pytorch训练模型得到*.pt文件,并将其转化为onnx模型,然后在TensorRT中加载onnx模型,并转换成TensorRT模型;再将TensorRT模型序列化存储到磁盘或内存中,称为plan文件;

[0029] S402、Deployment阶段进行轻量化YOLOv5模型的部署,先对步骤S401获得的plan文件反序列化,创建runtime引擎,完成前向推理过程,。

[0030] 具体的,步骤S5具体为:

[0031] S501、将待检测图片的子图送入训练好的轻量化YOLOv5s模型进行检测前,若待检测图片的子图不满足模型对于图片大小的要求,则进行自适应图片缩放,将子图送入特征提取网络,得到S×S大小的特征图,将输入图像分为S×S的小格;

[0032] S502、每个网格使用逻辑回归预测B个边界框,若预测边界框中心在网格单元内,则网格单元的B个边界框负责对目标进行分类和边框预测,得到每个网格对B个边界框的预测结果,输出边界框的位置信息、表明网格是否包含目标的置信度和C个类别的概率信息,每个边界框预测得到t_x、t_y、t_w、t_h、t_o,t_x、t_y是边界框中心坐标相对于当前网格单元的偏移值;使用逻辑激活对t_x和t_y进行归一化处理,将值限制在0~1内,t_w、t_h是边界框宽和高的尺度缩放,t_o是置信度confidence;

[0033] S503、采用特征金字塔网络下采样自顶向下传达强语义特征和路径聚合网络上采样自底向上传达强定位特征融合三个尺度分别的检测结果;针对图片输入大小为960×960,输出特征图分别为120×120,60×60,30×30,分别为8倍、16倍和32倍下采样的结果。

[0034] 进一步的,步骤S502中,根据每个边界框预测得到的5个值得到预测的边界框在整个特征图中的中心点坐标b_x、b_y和长宽b_w、b_h如下:

[0035] $b_x=\sigma(t_x)+c_x$

[0036] $b_y=\sigma(t_y)+c_y$

[0037] $$b_w = p_w e^{t_w}$$

[0038] $$b_h = p_h e^{t_h}$$

[0039] 其中,σ函数为逻辑激活,c_x和c_y分别为当前网格单元相对于特征图左上角的距离,p_w和p_h为先验框的长和宽。

[0040] 更进一步的,限制坐标偏移和置信度在0~1内,当真实框落在网格单元内时,Pr(object)为1,否则,Pr(object)为0,网格单元在包含目标的条件下属于某个类别的概率Pr($class_i$|object)表示为

[0041] $$\Pr(class_i|object) * \Pr(object) * IOU_{pred}^{truth} = \Pr(class_i) * IOU_{pred}^{truth}$$

[0042] 其中,IOU_{pred}^{truth}为真实框与预测框的交并比,Pr($class_i$)为某个单元格内目标对应类别的概率。

[0043] 具体的,步骤S6具体为:

[0044] S601、通过目标在子图上的位置信息和子图在大图上的相对位置推算目标在大图上的位置信息;

[0045] S602、针对某一类别,设置NMS阈值为0.65,选择置信度最高的边界框,根据边界框与其他边界框的DIOU值滤除超过NMS阈值的所有边界框,筛选完预测框后根据保留的预测框进行画框,完成大尺寸SAR图像的舰船检测。

[0046] 本发明的另一技术方案是,一种基于轻量级深度学习的SAR图像舰船检测系统,包括:

[0047] 数据模块,对大尺寸SAR图像进行预处理,选取包含目标信息的子图作为训练样本;

[0048] 处理模块,引入Ghost模块和GhostBottleneck对YOLOv5s模型进行升级,得到初步轻量化的YOLOv5s模型,使用数据模块选取的训练样本对YOLOv5s模型进行训练;

[0049] 微调模块,对处理模块训练后得到的YOLOv5s模型进行蒸馏,然后进行稀疏化训练和剪枝,对剪枝后的YOLOv5s模型进行微调训练;

[0050] 推理模块,利用TensorRT推理优化器对微调模块微调训练后的YOLOv5s模型进行推理加速,并部署在NVIDIA Jetson TX2上;

[0051] 检测模块,对待检测的SAR图像进行裁剪处理后依次送入推理模块部署在NVIDIA Jetson TX2上的YOLOv5s模型进行检测,得到对应的子图检测结果;

[0052] 去除模块,对检测模块得到的子图检测结果进行拼接,并在最终的大尺寸SAR图像上使用NMS非极大值抑制筛选预测框,根据筛选后的预测框数值在原始大尺寸图像上绘制预测框并标记类别,实现大尺寸SAR图像舰船检测。

[0053] 与现有技术相比,本发明至少具有以下有益效果:

[0054] 本发明一种基于轻量级深度学习的SAR图像舰船检测方法,采用的技术手段就是网络剪枝、知识蒸馏和Ghost算法,针对遥感影像尺寸大的特点,直接缩放输入网络中会导致过多信息的丢失,采用对图片以一定重合度切块的方式既能避免网络信息的丢失,也能保证图片大小与网络输入相匹配;将传统模型压缩算法网络剪枝和知识蒸馏与人工设计的轻量级模型Ghost相结合,升级目标检测网络YOLOv5;在很大程度上减少模型的参数量和浮点运算量,提升推理速度。

[0055] 进一步的,对于尺寸格式各异的数据集针对性地以一定的重合度切块,统一图片尺寸和格式,保证输入模型的训练样本尺寸和格式一致。

[0056] 进一步的,用轻量化模型Ghost优化升级YOLOv5s模型中的卷积模块和瓶颈模块,并将宽度乘数调整为0.15,将深度乘数调整为0.35,将网络层数减少至212层,从而减少了模型的参数了和浮点运算量。

[0057] 进一步的,为了对模型进行进一步的压缩,引入传统模型压缩算法,网络剪枝和知识蒸馏,知识蒸馏将大模型YOLOv5m优越性能教授给轻量化后的YOLOv5s,在一定程度上提升了模型性能,网络剪枝通过衡量神经元的重要性剪去相对不重要的神经元,更进一步的减少了模型参数和浮点运算量。

[0058] 进一步的,轻量化模型的意义是为了实现深度学习模型在嵌入式设备上的部署,步骤S4利用TensorRT推理优化器将轻量化后的YOLOv5s部署在NVIDIA Jetson TX2上。

[0059] 进一步的,通过步骤S5说明了轻量化YOLOv5s对大尺寸SAR图像的检测过程,最终得到标有预测框和类别信息的结果图。

[0060] 进一步的,通过步骤S502中对预测边界框在整个特征图中的中心点坐标和长宽的计算公示进行了说明。

[0061] 进一步的,网格单元在包含目标的条件下属于某个类别的概率$Pr(class_i|object)$是网络输出的结果

[0062] 进一步的,步骤S6将待检测图片的子图结果还原到原尺寸图片上,并通过NMS滤除重复度较高的预测框,得到最终的检测结果。

[0063] 综上所述,本发明提出一个完整的模型轻量化流程,最终得到一个轻量化后的YOLOv5s模型,并将其部署在嵌入式设备NVIDIA Jetson TX2上,完成大尺寸SAR图像的舰船

舰船任务。

[0064] 下面通过附图和实施例,对本发明的技术方案做进一步的详细描述。

附图说明

[0065] 图1为本发明的流程示意图;

[0066] 图2为Ghost模型图解;

[0067] 图3为GhostBottleneck图解;

[0068] 图4为复杂模型对复杂场景检测所得结果隐去置信度后的关键部分示意图;

[0069] 图5为复杂模型对复杂场景检测所得结果未隐去置信度的关键部分示意图;

[0070] 图6为复杂模型对复杂场景检测所得结果未隐去置信度的关键部分示意图;

[0071] 图7为简单模型对简单场景检测所得结果隐去置信度后的关键部分示意图;

[0072] 图8为简单模型对简单场景检测所得结果未隐去置信度的关键部分示意图。

具体实施方式

[0073] 下面将结合本发明实施例中的附图,对本发明实施例中的技术方案进行清楚、完整地描述,显然,所描述的实施例是本发明一部分实施例,而不是全部的实施例。基于本发明中的实施例,本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例,都属于本发明保护的范围。

[0074] 在本发明的描述中,需要理解的是,术语“包括”和“包含”指示所描述特征、整体、步骤、操作、元素和/或组件的存在,但并不排除一个或多个其它特征、整体、步骤、操作、元素、组件和/或其集合的存在或添加。

[0075] 还应当理解,在本发明说明书中所使用的术语仅仅是出于描述特定实施例的目的而并不意在限制本发明。如在本发明说明书和所附权利要求书中所使用的那样,除非上下文清楚地指明其它情况,否则单数形式的“一”、“一个”及“该”意在包括复数形式。

[0076] 还应当进一步理解,在本发明说明书和所附权利要求书中使用的术语“和/或”是指相关联列出的项中的一个或多个的任何组合以及所有可能组合,并且包括这些组合。

[0077] 在附图中示出了根据本发明公开实施例的各种结构示意图。这些图并非是按比例绘制的,其中为了清楚表达的目的,放大了某些细节,并且可能省略了某些细节。图中所示出的各种区域、层的形状及它们之间的相对大小、位置关系仅是示例性的,实际中可能由于制造公差或技术限制而有所偏差,并且本领域技术人员根据实际所需可以另外设计具有不同形状、大小、相对位置的区域/层。

[0078] 本发明提供了一种基于轻量级深度学习的SAR图像舰船检测方法,面向嵌入式设备NVIDIA Jetson TX2,涉及模型压缩方法,利用传统模型压缩算法和人工设计的轻量级模型压缩优化目标检测网络,可应用于大尺寸合成孔径雷达图像中某些特定类目标的检测;在满足可接受精度损失的前提下,压缩模型的参数量和浮点运算量,提升检测速度。

[0079] 请参阅图1,本发明一种基于轻量级深度学习的SAR图像舰船检测方法,包括以下步骤:

[0080] S1、对大尺寸SAR图像进行预处理,选取包含目标信息的子图作为训练样本,在训练样本上对步骤S2得到轻量化的YOLOv5s模型500个epoch;

[0081] S101、对5张10000×10000像素、16位单通道TIF图像Img1OK和31张3000×3000像素、16位单通道TIFF图像AIR-SARShip-1.0以50％的overlap(重合率)进行切块,切块得到大尺寸遥感图像的子图,将作为训练样本输入网络进行训练;

[0082] S102、将1000张800×800像素、8位JPG图像SAR-train-int放大至1000×1000的尺寸;

[0083] S103、统一Img-1OK、AIR-SARShip-1.0、AIR-SARShip-2.0和SAR-train-int图像格式为8位单通道TIF图像,最终建立的数据集共包括2551张图片,划分训练样本2351张,验证样本200张;

[0084] S104、利用Mosaic数据增强算法,对四张图片以随机缩放、随机裁剪、随机排布的方式进行拼接,增加小目标样本数量使得训练数据分布趋向均匀。

[0085] S2、引入Ghost模块和GhostBottleneck对YOLOv5s模型进行升级,完成YOLOv5s模型的初步轻量化,使用步骤S1选取的训练样本对YOLOv5s模型进行500个epoch的训练;

[0086] Ghost模块替代标准卷积的思想是用少量本征特征图进行廉价线性变换后的“重影”作为输出特征图。它利用了冗余特征图对之间的相似性,基于少量本征特征图通过简单线性变换可得到大量相似的冗余特征图的假设,实现对卷积参数量和运算量压缩的目的。Ghost模块将一个标准卷积分解为两个部分,第一部分用少量标准卷积生成少量的本征特征图,第二部分对本征特征图进行简单线性运算以极低的成本生成大量“重影”特征图,即冗余特征图。

[0087] S201、利用Ghost模块和GhostBottleneck对YOLOv5s升级的具体操作是将YOLOv5s模型主干网络中的卷积模块和瓶颈模块分别用Ghost模块和GhostBottleneck替换,如图3所示。

[0088] S202、由于Ghost模块会给网络深度带来大幅增加,考虑通过更改深度乘数来减少由Ghost模块带来的网络深度的增加。将宽度乘数调整为0.15,深度乘数调整为0.35,网络层数减少至212层,得到初步轻量化的YOLOv5s模型。

[0089] 调整了两个乘数:宽度乘数和深度乘数,通过调整这两个乘数让网络层数减少,这个过程叫初步的轻量化。

[0090] S3、在步骤S2得到的YOLOv5s模型基础上,利用传统模型轻量化算法网络剪枝和知识蒸馏实现YOLOv5s初步轻量化模型的进一步轻量化,得到YOLOv5s模型;

[0091] 对步骤S2初步轻量化的YOLOv5s模型先进行蒸馏,蒸馏后进行稀疏化训练,再进行剪枝,然后对剪枝后的模型进行微调训练以恢复精度。

[0092] S301、以YOLOv5m作为教师模型(T-model),用L2 loss作为蒸馏基础函数,损失中的蒸馏dist平衡系数选择为1,蒸馏训练100个epoch;

[0093] S302、在正常训练得到过度参数化模型后,设定稀疏参数6e-4,稀疏化训练对BN层gamma参数进行L1正则化,产生稀疏权值矩阵。以此作为评价神经元贡献大小的标准,并根据30％稀疏率确定阈值。剪去小于阈值的层以及该层的依赖层,若该层中所有通道都需要被移除,为保证网络结构,保留最大通道;

[0094] 步骤S301是对这个初步轻量化的模型进行蒸馏优化,步骤S302是对蒸馏后的模型进一步剪枝,得到的剪枝后的模型。

[0095] S303、在完成剪枝处理后,为了保证模型精度不会出现大幅下降,对步骤S302得到

的剪枝后的模型继续训练50个epoch,通过微调训练学习稀疏连接的最终权重。

[0096] S4、利用TensorRT推理优化器对步骤S3得到的YOLOv5s模型进行推理加速,并将其部署在NVIDIAJetson TX2上,TensorRT推理优化器进行部署包括Build阶段和Deploymeng阶段;

[0097] S401、Build阶段优化用Pytorch训练模型得到*.pt文件,并将其转换成onnx模型,然后在TensorRT中加载onnx模型,并转换成TensorRT模型,将TensorRT模型序列化存储到磁盘或内存中,称为plan文件;

[0098] S402、Deployment阶段是进行轻量化YOLOv5模型的部署,完成前向推理过程。先对Build过程中获得的plan文件反序列化,创建runtime引擎,进行推理。

[0099] S5、对待检测大尺寸SAR图像进行裁剪处理后,依次送入步骤S4部署在NVIDIA Jetson TX2上的YOLOv5s模型完成检测;

[0100] 与生成训练样本类似,对大尺寸SAR图像以50%的overlap(重合率)进行切块为1000×1000的子图,依次送入模型进行检测。

[0101] S501、将待检测图片的子图送入训练好的轻量化YOLOv5s模型进行检测前,若子图不满足模型对于图片大小的要求,则进行自适应图片缩放,将子图送入特征提取网络,得到S×S大小的特征图,将输入图像分为S×S的小格;

[0102] S502、每个网格使用逻辑回归预测B个边界框,若预测边界框中心在网格单元内,则该网格单元的B个边界框负责对该目标进行分类和边框预测,得到每个网格对B个边界框的预测结果;

[0103] 输出边界框的位置信息、表明该网格是否包含目标的置信度和C个类别的概率信息。每个边界框预测得到5个值:t_x、t_y、t_w、t_h、t_o。t_x、t_y是边界框中心坐标相对于当前网格单元的偏移值。同时为了保证将边界框的中心约束在当前网格单元内,使用逻辑激活(Logistic)对t_x和t_y进行归一化处理,将t_x和t_y的值限制在0~1内,使得模型训练更为稳定;t_w、t_h是边界框宽和高的尺度缩放,t_o是置信度confidence,RCNN中提到$t_w = \ln(\frac{b_w}{p_w})$,$t_h$的计算也是如此。

[0104] 根据每个边界框预测得到的5个值进而可以根据下面的公式计算得到预测的边界框在整个特征图中的中心点坐标b_x、b_y和长宽b_w、b_h。

[0105] $b_x=\sigma(t_x)+c_x$ (1)

[0106] $b_y=\sigma(t_y)+c_y$ (2)

[0107] $$b_w = p_w e^{t_w} \quad (3)$$

[0108] $$b_h = p_h e^{t_h} \quad (4)$$

[0109] $$\Pr(object) * IOU_{pred}^{truth} = \sigma(t_o) \quad (5)$$

[0110] 其中,c_x和c_y是当前网格单元相对于特征图左上角的距离,p_w和p_h是为先验框长和宽。σ函数为逻辑激活,限制坐标偏移和置信度在0~1内。当真实框落在网格单元内时,真实框落在网格单元的概率Pr(object)为1,否则,Pr(object)为0。

[0111] 某网格单元在包含目标的条件下属于某个类别的概率$\Pr(class_i|object)$表示

为:

[0112] $$\Pr(class_i|object)*\Pr(object)*IOU_{pred}^{truth}=\Pr(class_i)*IOU_{pred}^{truth} \quad (6)$$

[0113] 其中,IOU_{pred}^{truth}为真实框与预测框的交并比,Pr(class$_i$)为某个单元格内目标对应类别的概率。

[0114] S503、特征金字塔网络FPN下采样自顶向下传达强语义特征和路径聚合网络PAN上采样自底向上传达强定位特征融合三个尺度分别的检测结果。

[0115] 针对图片输入大小为960×960,输出特征图分别为120×120,60×60,30×30,分别为8倍、16倍和32倍下采样的结果。

[0116] S6、将步骤S5中得到的子图检测结果进行拼接,并在最终的大尺寸SAR图像上使用NMS非极大值抑制筛选预测框,根据筛选后的预测框数值在原始大尺寸图像上绘制预测框,在原图上绘制保留的预测框并标记类别,完成大尺寸SAR图像目标检测。

[0117] S601、拼接过程是切块过程的逆过程,通过目标在子图上的位置信息和子图在大图上的相对位置推算目标在大图上的位置信息;

[0118] S602、针对某一类别,设置NMS阈值为0.65,选择置信度最高的边界框,根据该边界框与其他边界框的DIOU值滤除超过NMS阈值的所有边界框,去除了重复率较大的边界框,NMS已经筛选了预测框,筛选完预测框后根据保留的预测框进行画框,此时完成大尺寸SAR图像的舰船检测。

[0119] 本发明再一个实施例中,提供一种基于轻量级深度学习的SAR图像舰船检测系统,该系统能够用于实现上述基于轻量级深度学习的SAR图像舰船检测方法,具体的,该基于轻量级深度学习的SAR图像舰船检测系统包括数据模块、处理模块、微调模块、推理模块、检测模块以及去除模块。

[0120] 其中,数据模块,对大尺寸SAR图像进行预处理,选取包含目标信息的子图作为训练样本;

[0121] 处理模块,引入Ghost模块和GhostBottleneck对YOLOv5s模型进行升级,得到初步轻量化的YOLOv5s模型,使用数据模块选取的训练样本对YOLOv5s模型进行训练;

[0122] 微调模块,对处理模块训练后得到的YOLOv5s模型进行蒸馏,然后进行稀疏化训练和剪枝,对剪枝后的YOLOv5s模型进行微调训练;

[0123] 推理模块,利用TensorRT推理优化器对微调模块微调训练后的YOLOv5s模型进行推理加速,并部署在NVIDIA Jetson TX2上;

[0124] 检测模块,对待检测的SAR图像进行裁剪处理后依次送入推理模块部署在NVIDIA Jetson TX2上的YOLOv5s模型进行检测,得到对应的子图检测结果;

[0125] 去除模块,对检测模块得到的子图检测结果进行拼接,并在最终的大尺寸SAR图像上使用NMS非极大值抑制筛选预测框,根据筛选后的预测框数值在原始大尺寸图像上绘制预测框,实现SAR图像舰船检测。

[0126] 本发明再一个实施例中,提供了一种终端设备,该终端设备包括处理器以及存储器,所述存储器用于存储计算机程序,所述计算机程序包括程序指令,所述处理器用于执行所述计算机存储介质存储的程序指令。处理器可能是中央处理单元(Central Processing Unit,CPU),还可以是其他通用处理器、数字信号处理器(Digital Signal Processor、

DSP)、专用集成电路(Application Specific Integrated Circuit,ASIC)、现成可编程门阵列(Field-Programmable GateArray,FPGA)或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等,其是终端的计算核心以及控制核心,其适于实现一条或一条以上指令,具体适于加载并执行一条或一条以上指令从而实现相应方法流程或相应功能;本发明实施例所述的处理器可以用于基于轻量级深度学习的SAR图像舰船检测方法的操作,包括:

[0127] 对大尺寸SAR图像进行预处理,选取包含目标信息的子图作为训练样本;引入Ghost模块和GhostBottleneck对YOLOv5s模型进行升级,得到初步轻量化的YOLOv5s模型,使用训练样本对YOLOv5s模型进行训练;对训练后得到的YOLOv5s模型进行蒸馏,然后进行稀疏化训练和剪枝处理,对剪枝处理后的YOLOv5s模型进行微调训练;利用TensorRT推理优化器对微调训练后的YOLOv5s模型进行推理加速,并部署在NVIDIA Jetson TX2上;对待检测的SAR图像进行裁剪处理后依次送入部署在NVIDIA Jetson TX2上的YOLOv5s模型进行检测,得到对应的子图检测结果;对得到的子图检测结果进行拼接,并在最终的大尺寸SAR图像上使用NMS非极大值抑制筛选预测框,根据筛选后的预测框数值在原始大尺寸图像上绘制预测框,实现SAR图像舰船检测。

[0128] 本发明再一个实施例中,本发明还提供了一种存储介质,具体为计算机可读存储介质(Memory),所述计算机可读存储介质是终端设备中的记忆设备,用于存放程序和数据。可以理解的是,此处的计算机可读存储介质既可以包括终端设备中的内置存储介质,当然也可以包括终端设备所支持的扩展存储介质。计算机可读存储介质提供存储空间,该存储空间存储了终端的操作系统。并且,在该存储空间中还存放了适于被处理器加载并执行的一条或一条以上的指令,这些指令可以是一个或一个以上的计算机程序(包括程序代码)。需要说明的是,此处的计算机可读存储介质可以是高速RAM存储器,也可以是非不稳定的存储器(non-volatile memory),例如至少一个磁盘存储器。

[0129] 可由处理器加载并执行计算机可读存储介质中存放的一条或一条以上指令,以实现上述实施例中有关基于轻量级深度学习的SAR图像舰船检测方法的相应步骤;计算机可读存储介质中的一条或一条以上指令由处理器加载并执行如下步骤:

[0130] 对大尺寸SAR图像进行预处理,选取包含目标信息的子图作为训练样本;引入Ghost模块和GhostBottleneck对YOLOv5s模型进行升级,得到初步轻量化的YOLOv5s模型,使用训练样本对YOLOv5s模型进行训练;对训练后得到的YOLOv5s模型进行蒸馏,然后进行稀疏化训练和剪枝处理,对剪枝处理后的YOLOv5s模型进行微调训练;利用TensorRT推理优化器对微调训练后的YOLOv5s模型进行推理加速,并部署在NVIDIA Jetson TX2上;对待检测的SAR图像进行裁剪处理后依次送入部署在NVIDIA Jetson TX2上的YOLOv5s模型进行检测,得到对应的子图检测结果;对得到的子图检测结果进行拼接,并在最终的大尺寸SAR图像上使用NMS非极大值抑制筛选预测框,根据筛选后的预测框数值在原始大尺寸图像上绘制预测框,实现SAR图像舰船检测。

[0131] 为使本发明实施例的目的、技术方案和优点更加清楚,下面将结合本发明实施例中的附图,对本发明实施例中的技术方案进行清楚、完整地描述,显然,所描述的实施例是本发明一部分实施例,而不是全部的实施例。通常在此处附图中的描述和所示的本发明实施例的组件可以通过各种不同的配置来布置和设计。因此,以下对在附图中提供的本发明

的实施例的详细描述并非旨在限制要求保护的本发明的范围,而是仅仅表示本发明的选定实施例。基于本发明中的实施例,本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例,都属于本发明保护的范围。

[0132] 本发明的效果可以通过以下实验进一步说明:

[0133] 1.实验环境

[0134] 训练主机仿真环境为:Ubuntu 18.04,Intel(R)Xeon(R)Gold 5118CPU,GPU为GeForce RTX 2080ti,python 3.8.5,CUDA 10.0.130,CuDNN 7.0。

[0135] Jetson TX2推理环境:Ubuntu 18.04,CPU为HMP Dual Denver 2/2MB L2+Quad ARM® A57/2MB L2,GPU为NVIDIA PascalTM,256CUDA cores,python3.8.5,CUDA 10.2.89,CuDNN 8.0.0.180,TensorRT版本7.1.3.0,Jetpack版本4.4.1。

[0136] 2.实验内容

[0137] (1)以YOLOv5s为基线模型,分别在训练主机的GPU和CPU上验证Ghost模块对于模型轻量化的有效性,记录模型参数量、浮点运算量、平均精度AP50、AP50:95、精确率P、召回率R、处理一张1000×1000图片的推理时间和总处理时间。证明Ghost模型对于参数量和浮点运算量压缩性能,如图2所示。

[0138] (2)对初步轻量化模型进行蒸馏剪枝,在NVIDIA Jetson TX2上测试模型,记录处理一张1000×1000图片的推理时间和总处理时间。验证模型在推理时间加速上的性能。

[0139] 3.仿真实验结果

[0140] 实验结果表明,Ghost模块对基线模型YOLOv5s在参数量和浮点运算量上均有较大压缩。使用的蒸馏和剪枝策略可以使模型进一步压缩,并且大大提升推理速度。

[0141] 表1 Mobile和Ghost模块在YOLOv5中的性能对比

[0142]

Model	All	L	Weights(M)	FLOPs(G)	AP_{50}	$AP_{50:95}$	P	R	Infer(ms)	Total(ms)
YOLOv5s	\	224	6.72	16.3	61.0%	33.5%	84.5%	56.6%	8	54
YOLOv5m	\	308	20.06	50.3	64.3%	33.5%	77.9%	60.1%	13	61
GhostYOLOv5		326	4.44	9.7	62.8%	33.7%	78.5%	59%	11	60
GhostYOLOv5	√	362	2.69	6.5	60.5%	32.9%	80.3%	58.4%	12	60.9

[0143] 各个模型的性能对比如表1所示。表格中的实验均在实验室服务器上完成。图像尺寸为适应网络输入,均为960×960。其中all代表是否替换网络中全部卷积模块和瓶颈块,L代表网络层数。可见Ghost对于模型的参数量和浮点运算量都有显著压缩,Ghost将YOLOv5s的参数量由6.72M减少至4.44M,浮点运算量由16.3G减少为9.7G,同时能保持一定精度GhostYOLOv5s在AP_{50}和$AP_{50:95}$上均略高于YOLOv5s,但在GPU上的推理速度却不如参数量和浮点运算量均大于它们的YOLOv5s。考虑GPU算力瓶颈在于访存带宽,为减少网络层数,提高推理速度,仅替换主干网络。在CPU上测试模型推理速度。结果如表2所示。

[0144] 表2模型CPU推理时间对比

[0145]

Model	Weights(M)	FLOPs(G)	AP_{50}	$AP_{50:95}$	P	R	Infer(ms)	Total(ms)
YOLOv5s	6.72	16.3	61.0%	33.5%	84.5%	56.6%	510	546
GhostYOLOv5	4.44	9.7	62.8%	33.7%	78.5%	59%	440	499

[0146] GhostYOLOv5s在CPU上的推理速度明显快于YOLOv5s,总计的处理时间也快于YOLOv5s,这说明Ghost对网络模型的轻量化是有效的。足以证明Ghost模型在网络压缩上的卓越性能。

[0147] 表3 depth乘数和width乘数对推理时间影响

[0148]

Model	Depth	Width	Weights(M)	FLOPs(G)	AP_{50}	$AP_{50:95}$	Infer(ms)	Total(ms)
GhostYOLOv5	0.33	0.50	4.44	9.7	62.8%	33.7%	11	60
GhostYOLOv5	0.15	0.35	2.22	5.1	63.0%	34.8%	12	60.9

[0149] YOLOv5通过调整宽度乘数(width multiple)和深度乘数(depth multiple)来实现四个不同大小的模型。由于Ghost模块会给网络深度带来大幅增加,考虑通过更改深度乘数来减少由Ghost模块带来的网络深度的增加。将宽度乘数调整为0.15,深度乘数调整为0.35。网络层数减少至212层。在服务器GPU上测试推理速度和总处理速度均有一定提高,但模型的精度并未出现下降。说明更改宽度乘数和深度乘数来控制模型宽度和深度的措施确实有效。

[0150] 表4 TensorRT推理性能对比

[0151]

Model	剪枝	蒸馏	Weights(M)	FLOPs(G)	AP_{50}	$AP_{50:95}$	Infer(ms)	Total(ms)
YOLOv5s			6.72	16.3	61.1%	33.5%	70.38	121.82
GhostYOLOv5			2.22	5.1	63.0%	34.8%	61	109.77
GhostYOLOv5	√		1.62	3.0	61.6%	32.2%	40.5	90.7
GhostYOLOv5		√	2.22	5.1	63.0%	32.3%	59.4	108.66
GhostYOLOv5	√	√	0.89	1.8	57.3%	27.7%	30.2	84.5

[0152] 经Ghost轻量化后的YOLOv5s再经过蒸馏、剪枝和微调训练后,模型的参数量、浮点运算量大大减少。不可避免造成了一定精度的损失,但这个损失在可接受的范围内。在TX2上一张1000×1000的图片推理时间仅为30.2ms,包括读取图片和后处理的总时长也仅为84.5ms。足以证明GhostYOLOv5的优越性。分析表4,发现蒸馏对于模型精度的提高效果甚微,但经过蒸馏再以同样剪枝策略剪枝的模型可以实现更进一步的轻量化。考虑蒸馏可以使模型权重分布更为密集,使得重要的权重更加重要,相对不重要的权重也更为不重要。可以使在稀疏训练中获得更为稀疏的稀疏矩阵。

[0153] 在10000×10000的测试图像上,图4展示了检测结果的关键部分。为不对小目标产生遮挡,隐去置信度信息。实验最终得到两个模型分别是只经过剪枝的复杂模型和经过蒸馏剪枝的简单模型。两个模型在参数量和浮点运算量上相差悬殊,针对的图像复杂度也有区别。复杂模型在GhostYOLOv5s在TX2上单张图片的推理时间为40ms,若包括加载模型,读取图片和后处理阶段,单张图片的总处理时间约为92ms。简单模型在GhostYOLOv5s在TX2上单张图片的推理时间为30.2ms,若包括加载模型,读取图片和后处理阶段,单张图片的总处理时间约为84.5ms。

[0154] 表5不同模型在复杂场景中检测效果对比

[0155]

Model	Weights(M)	FLOPs(G)	AP_{50}	$AP_{50:95}$	Miss	Fake	F_1	Infer(s)	Total(s)
Complex	1.62	3.0	69%	42.3%	35.6%	13.5%	0.738	14.44	32.2
Simple	0.89	1.8	56.1%	27.3%	49.5%	22.9%	0.61	10.9	30.5

[0156] 表6不同模型在简单场景中检测效果对比

[0157]

Model	Weights(M)	FLOPs(G)	AP_{50}	$AP_{50:95}$	Miss	Fake	F_1	Infer(s)	Total(s)
Complex	1.62	3.0	81.6%	50%	27.9%	13%	0.789	14.44	32.2
Simple	0.89	1.8	55.1%	24.6%	46.3%	36.7%	0.581	10.9	30.5

[0158] 图5和图6展示了复杂模型未隐去置信度图片。该图均为河道,靠岸检测复杂度较大,检测的置信度均能保持较高水平。

[0159] 复杂主要针对复杂场景,简单模型主要针对简单场景。图7和图8展示了简单模型对简单场景的检测效果。在范围海域内简单模型表现出优越的性能。因此针对不同复杂度的图像选择恰当的模型可以实现推理速度的最优化。

[0160] 综上所述,本发明一种基于轻量级深度学习的SAR图像舰船检测方法,将得到的轻量化后的YOLOv5s模型部署在嵌入式设备NVIDIA Jetson TX2上,完成大尺寸SAR图像的舰船舰船任务,在SAR图像的简单场景和复杂场景均能有效检测舰船。

[0161] 本领域内的技术人员应明白,本申请的实施例可提供为方法、系统、或计算机程序产品。因此,本申请可采用完全硬件实施例、完全软件实施例、或结合软件和硬件方面的实施例的形式。而且,本申请可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质(包括但不限于磁盘存储器、CD-ROM、光学存储器等)上实施的计算机程序产品的形式。

[0162] 本申请是参照根据本申请实施例的方法、设备(系统)、和计算机程序产品的流程图和/或方框图来描述的。应理解可由计算机程序指令实现流程图和/或方框图中的每一流程和/或方框、以及流程图和/或方框图中的流程和/或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理设备的处理器以产生一个机器,使得通过计算机或其他可编程数据处理设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的装置。

[0163] 这些计算机程序指令也可存储在能引导计算机或其他可编程数据处理设备以特定方式工作的计算机可读存储器中,使得存储在该计算机可读存储器中的指令产生包括指令装置的制造品,该指令装置实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能。

[0164] 这些计算机程序指令也可装载到计算机或其他可编程数据处理设备上,使得在计算机或其他可编程设备上执行一系列操作步骤以产生计算机实现的处理,从而在计算机或其他可编程设备上执行的指令提供用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的步骤。

[0165] 以上内容仅为说明本发明的技术思想,不能以此限定本发明的保护范围,凡是按照本发明提出的技术思想,在技术方案基础上所做的任何改动,均落入本发明权利要求书的保护范围之内。

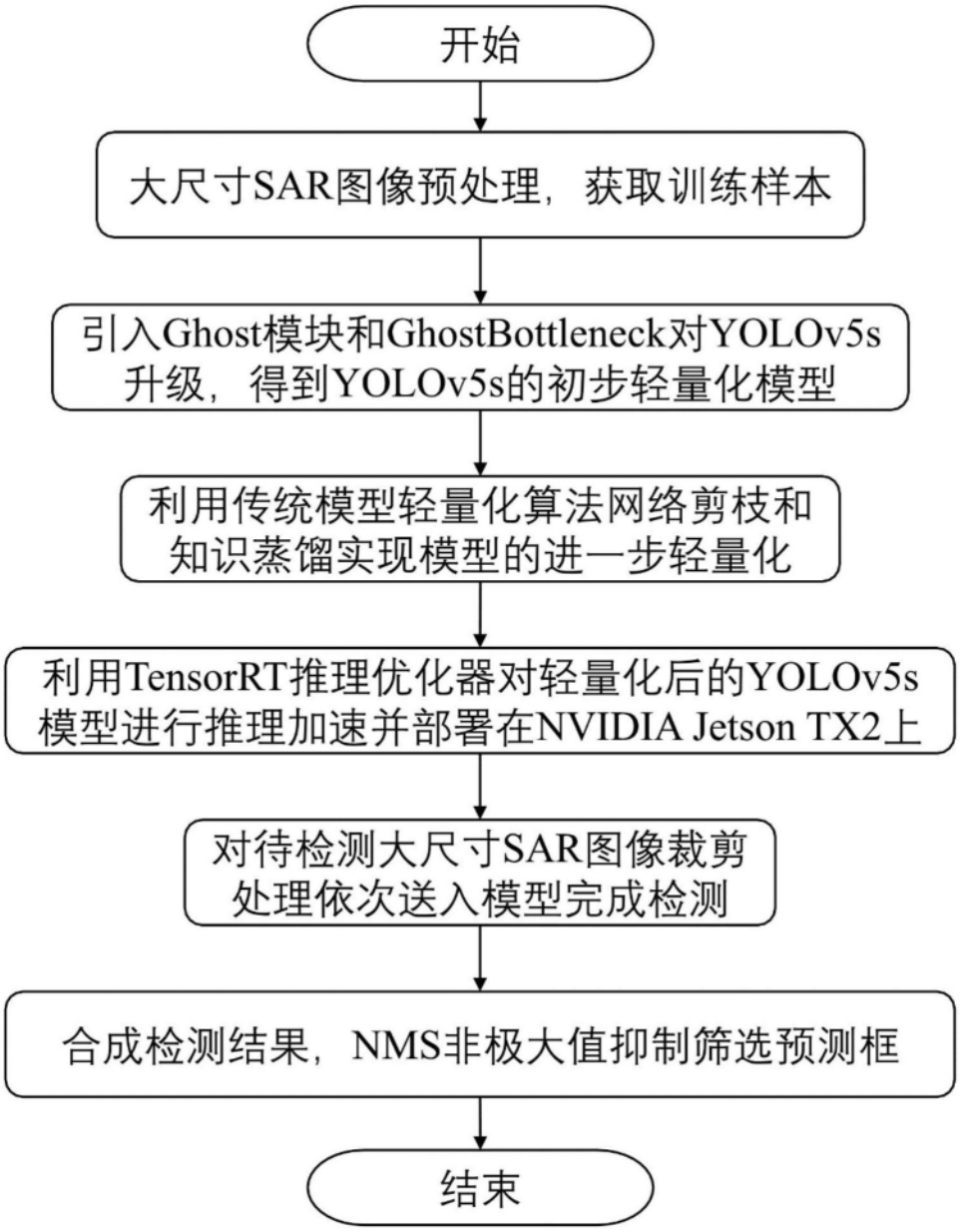

开始
大尺寸SAR图像预处理，获取训练样本
引入Ghost模块和GhostBottleneck对YOLOv5s升级，得到YOLOv5s的初步轻量化模型
利用传统模型轻量化算法网络剪枝和知识蒸馏实现模型的进一步轻量化
利用TensorRT推理优化器对轻量化后的YOLOv5s模型进行推理加速并部署在NVIDIA Jetson TX2上
对待检测大尺寸SAR图像裁剪处理依次送入模型完成检测
合成检测结果，NMS非极大值抑制筛选预测框
结束

图1

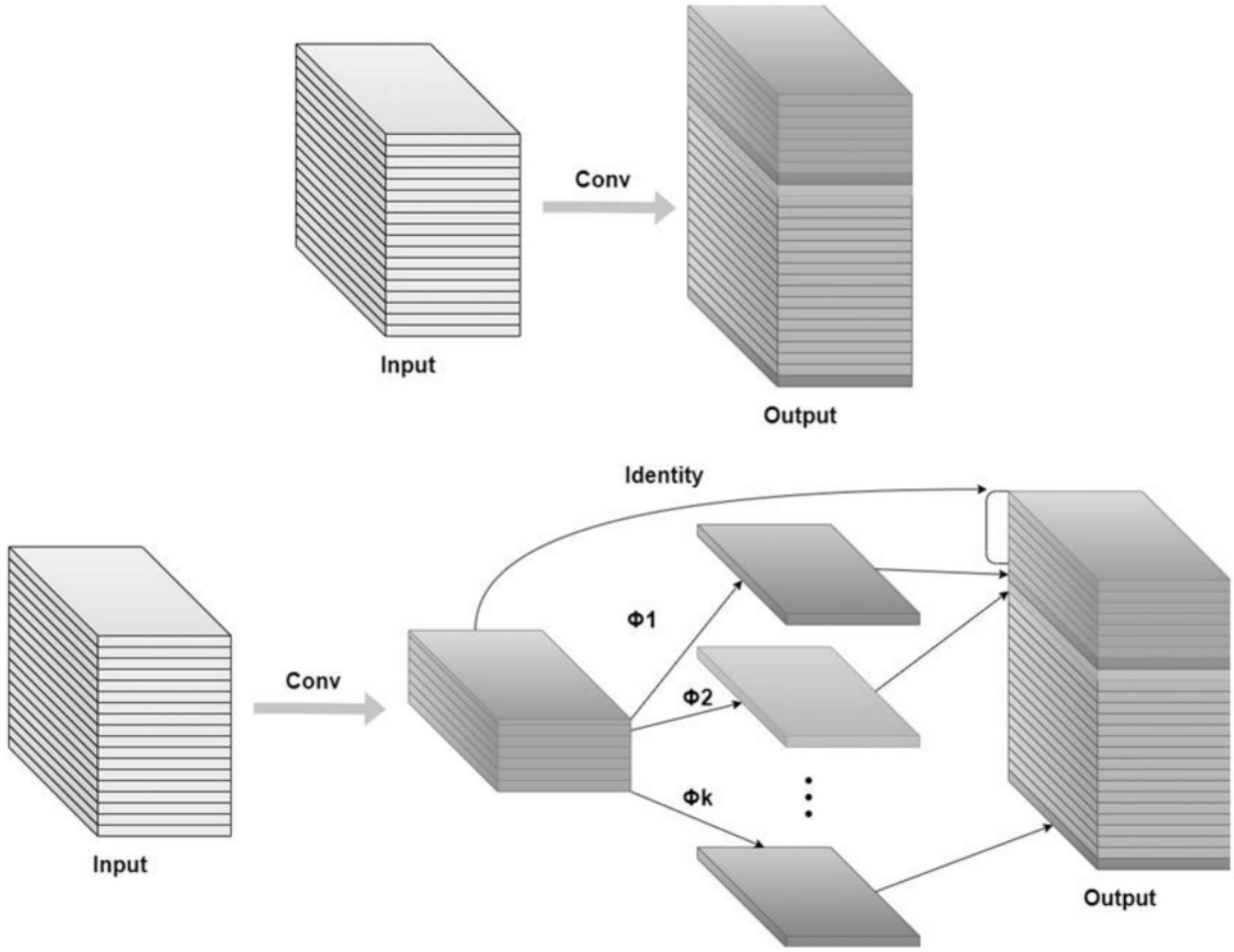


图2

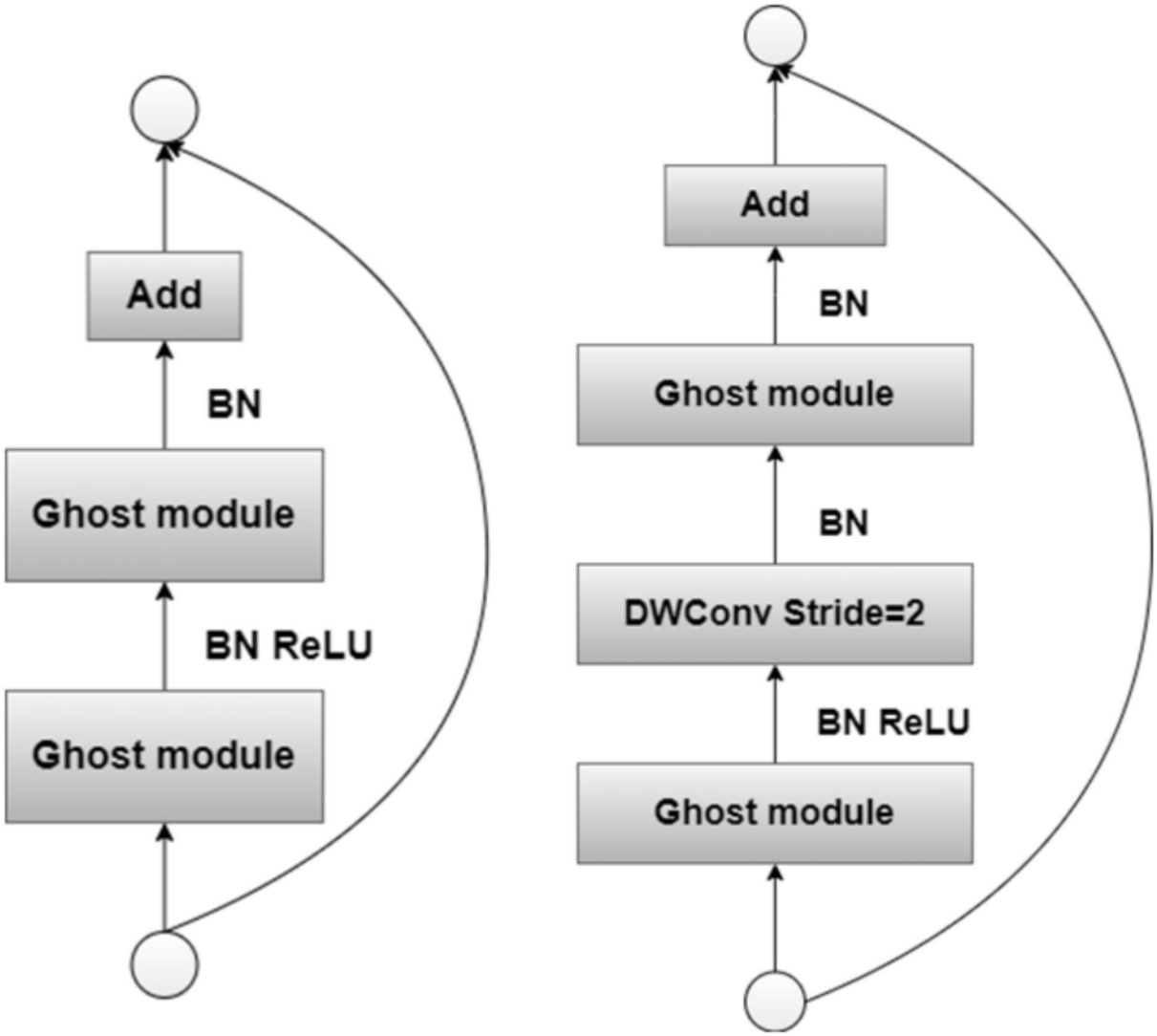

Add
BN
Ghost module
BN ReLU
Ghost module
Add
BN
Ghost module
BN
DWConv Stride=2
BN ReLU
Ghost module

图3

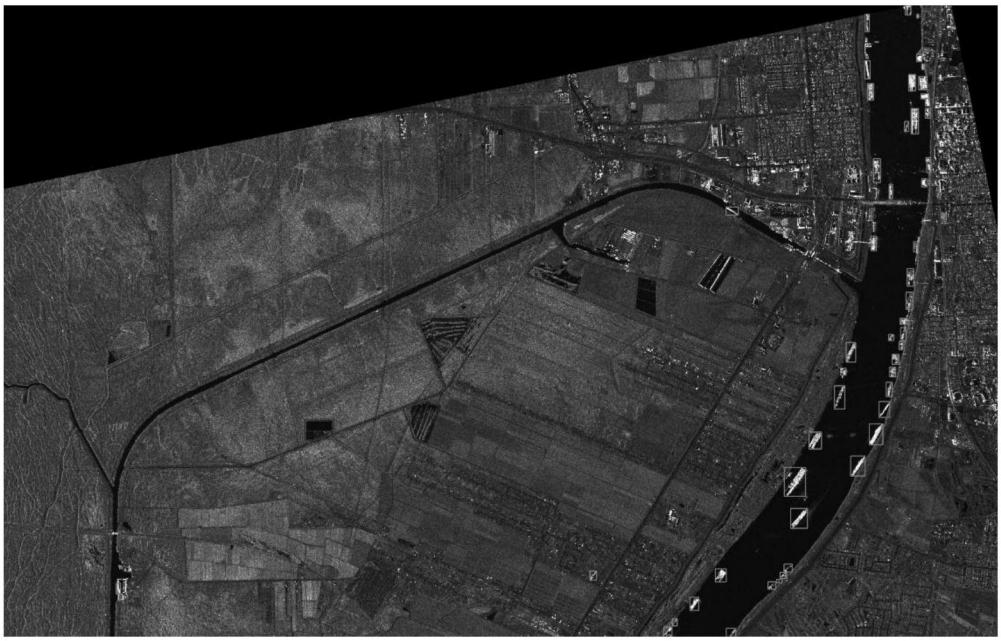


图4

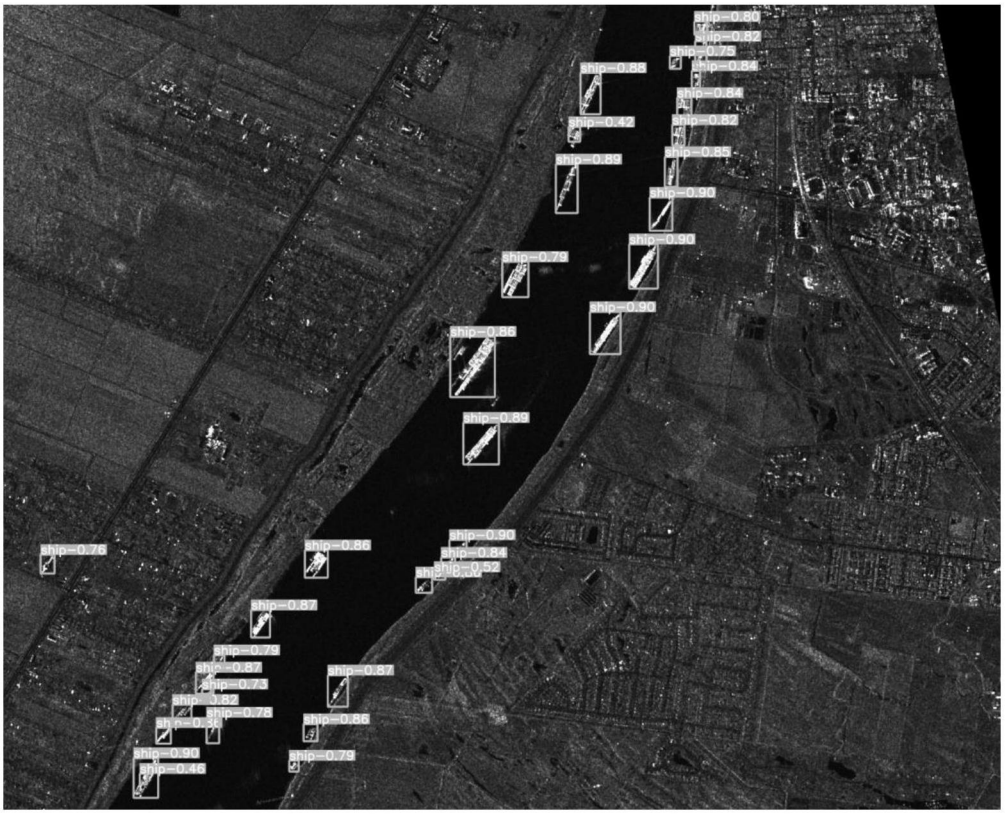

ship=0.88
ship=0.42
ship=0.89
ship=0.90
ship=0.90
ship=0.76
ship=0.90
ship=0.86
ship=0.89
ship=0.76
ship=0.86
ship=0.90
ship=0.84
ship=0.87
ship=0.79
ship=0.87
ship=0.87
ship=0.86
ship=0.90
ship=0.79

图5

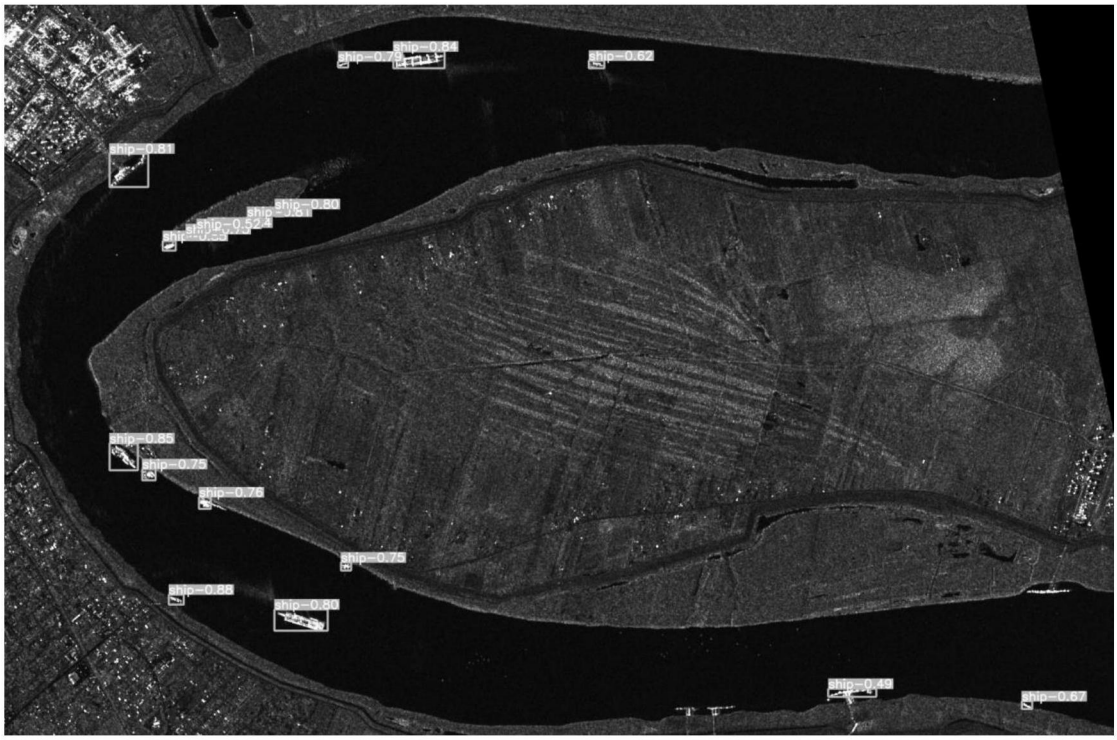

ship=0.84
ship=0.62
ship=0.81
ship=0.80
ship=0.85
ship=0.75
ship=0.76
ship=0.75
ship=0.88
ship=0.80
ship=0.46
ship=0.67

图6

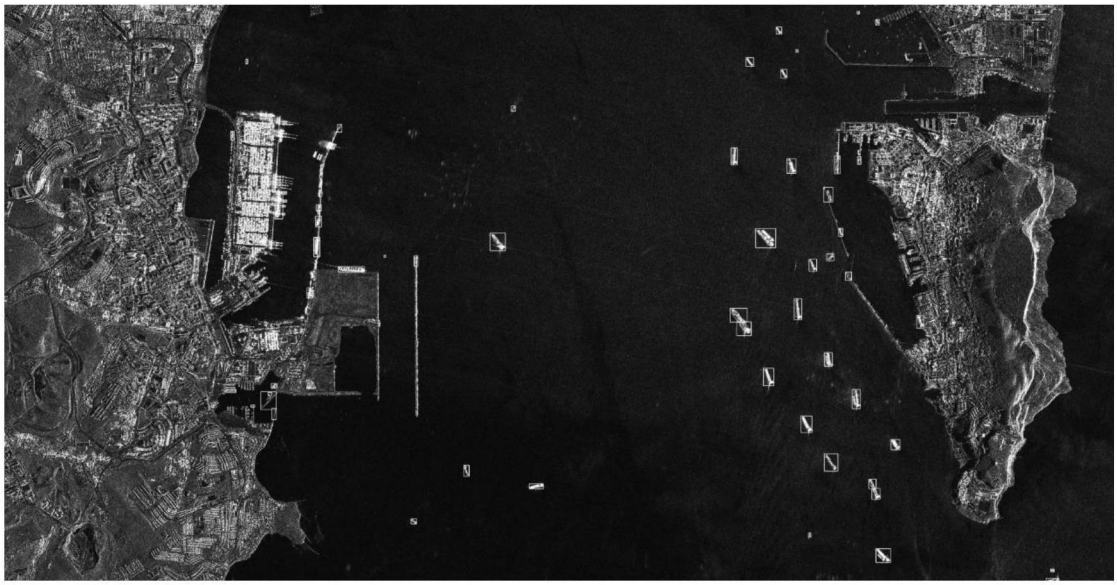


图7

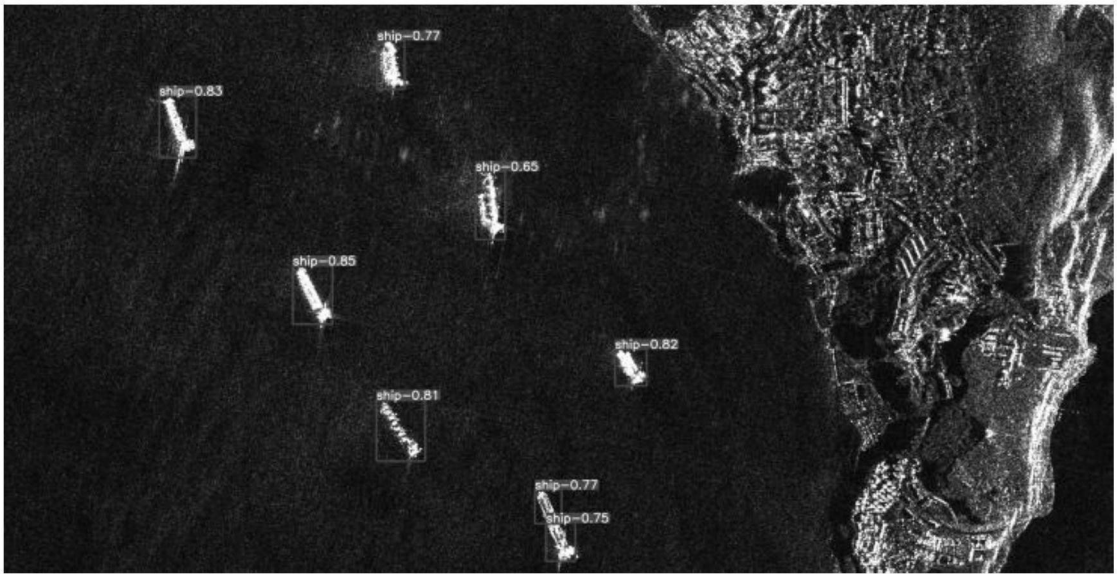

ship-0.77
ship-0.83
ship-0.65
ship-0.85
ship-0.82
ship-0.81
ship-0.77
ship-0.75

图8