文章编号:1674-2974(2022)10-0069-08 DOI:10.16339/j.cnki.hdxbzkb.2022360

MPANet-YOLOv5:多路径聚合网络复杂海域目标检测

王文亮,李延祥†,张一帆,韩鹏,刘识灏
[中船(浙江)海洋科技有限公司,浙江 舟山 316000]

摘 要:船舶智能化的发展对船舶视觉感知系统实时目标检测能力提出了更高要求,YOLOv5作为YOLO(You Only Look Once)系列算法的最新成果,以良好的速度和精度被广泛应用于海上目标检测.但在实际海上航行中往往伴随着多变的自然条件和复杂的活动场景,这使其在复杂海域中小目标检测能力和多目标分类效果并不理想.因此,为提升YOLOv5在复杂海域中目标检测能力,本文提出多路径聚合网络结构(MPANet).在自底向上特征传递过程中融合多层次特征信息以增强多尺度定位能力,同时结合SimAM注意力模块和Transformer结构增强高阶特征语义信息.在自定义数据集中实验结果表明:MPANet-YOLOv5相较于YOLOv5模型AP提升了5.4%,召回率提升了3.3%,$AP_{0.5}$提升了3.3%,$AP_{0.5:0.95}$提升了2.2%,不同海域测试结果显示MPANet-YOLOv5海面小目标检测能力明显优于YOLOv5.

关键词:目标检测;注意力机制;Transformer;船舶检测;多路径聚合网络
中图分类号:TP391.41 **文献标志码**:A

MPANet-YOLOv5: Multi-path Aggregation Network for Complex Sea Object Detection

WANG Wenliang, LI Yanxiang†, ZHANG Yifan, HAN Peng, LIU Shihao
[CSSC (Zhejiang) Ocean Technology Co., Ltd., Zhoushan 316000, China]

Abstract: The development of ship intelligence puts forward higher demands on the real-time object detection capability of ship vision perception systems. YOLOv5, the latest achievement of the YOLO (You Only Look Once) series of algorithms, is widely used for object detection at sea with good speed and accuracy. However, in actual sea navigation, it is often accompanied by variable natural conditions and complex activity scenarios, which makes its ability to detect small objects and multi-target classification in complex waters unsatisfactory. Therefore, to improve the target detection capability of YOLOv5 in complex seas, this paper proposes a Multi-Path Aggregation Network (MPANet) structure. MPANet enhances multi-scale localization capability by fusing multi-level feature information in the bottom-up feature transfer process, and enhances higher-order feature semantic information by combining the SimAM attention module and Transformer structure. The experimental results of the custom dataset show that

* **收稿日期**:2021-11-09
基金项目:智慧海洋舟山大数据应用服务平台建设项目(ZCHYGC201901)
作者简介:王文亮(1987—),男,北京人,中船(浙江)海洋科技有限公司高级工程师,硕士
† 通信联系人,E-mail: 1437042520@qq.com

MPANet-YOLOv5 improves AP by 5.4%, recall by 3.3%, $AP_{0.5}$ by 3.3%, and $AP_{0.5:0.95}$ by 2.2%, compared with the YOLOv5 model. The results of different sea area tests show that MPANet-YOLOv5 has significantly better detection capability for small objects on the sea surface than YOLOv5.

Key words: object detection; attention mechanism; transformer; ship detection; multi path aggregation network

经济全球化的不断发展使海上运输需求迅猛上升,随着船只数量日益增长,海上航行的安全性备受人们关注.船舶智能化的快速发展进一步提高了海上航行的安全性,其视觉感知系统能够实时获取周围船舶及障碍信息.近年来,深度学习技术被广泛用于目标检测领域,这进一步提升了船舶视觉感知系统的信息获取能力.

深度学习目标检测算法包括以 Faster R-CNN[1,2]为代表的两阶段算法和以 YOLO[3-6]系列为代表的单阶段目标检测算法.由于 YOLO 系列算法在保证较高精度的同时具有明显的实时性优势,因此被广泛部署于实时目标检测项目中.YOLOv5 作为 YOLO 系列最新的研究成果,其检测速度和精度都达到了 SOTA 水准.然而,在海上目标实时检测中往往伴随其他问题,如:小目标集群、船只密集且类型复杂、海面大雾影响造成目标模糊等,这些问题使得检测海域具有较高的复杂性,同时对 YOLOv5 的小目标检测能力及分类能力有了更高的要求.在实际部署中发现 YOLOv5 算法在复杂环境海域海面实时目标检测应用中仍需进一步改进.

针对 YOLO 算法的改进方案层出不穷,在船舶目标检测中改进方法主要包括数据增强[7]、多尺度[8]、特征融合[9]与添加注意力机制[10]等.受海上船舶检测数据集的限制,现有改进方案仅在较小数据集上进行训练并不能进行实际部署.因此,本文通过收集东海、南海部分航线图像构建了大型船舶检测数据集,提出了多路径聚合网络结构通过融合多层特征以增强模型多尺度定位能力,同时结合 SimAM[10]注意力模块和 Transformer[11]结构增强高阶特征语义信息.在自定义船舶数据集上实验结果表明:MPANet-YOLOv5 模型在增加少量参数的情况下明显提升了海上目标检测能力,取得了良好的效果.

1 YOLOv5 概述

YOLOv5 模型共包含四个版本 YOLOv5s、YOLOv5m、YOLOv5l 和 YOLOv5x,模型参数和性能依次提升.YOLOv5 依旧延续 Input、Backbone、Neck 和 Head 输出的网络结构,其结构如图 1 所示.

YOLOv5s 是 YOLOv5 系列中最小的模型结构,模型宽度和深度分别为 0.33 m 和 0.5 m,YOLOv5m、YOLOv5l 以及 YOLOv5x 在 YOLOv5s 模型基础上不断加深加宽.在数据输入端 YOLOv5 模型延续了 YOLOv4 中的 mosaic 数据增强方法,该方法将四张图片

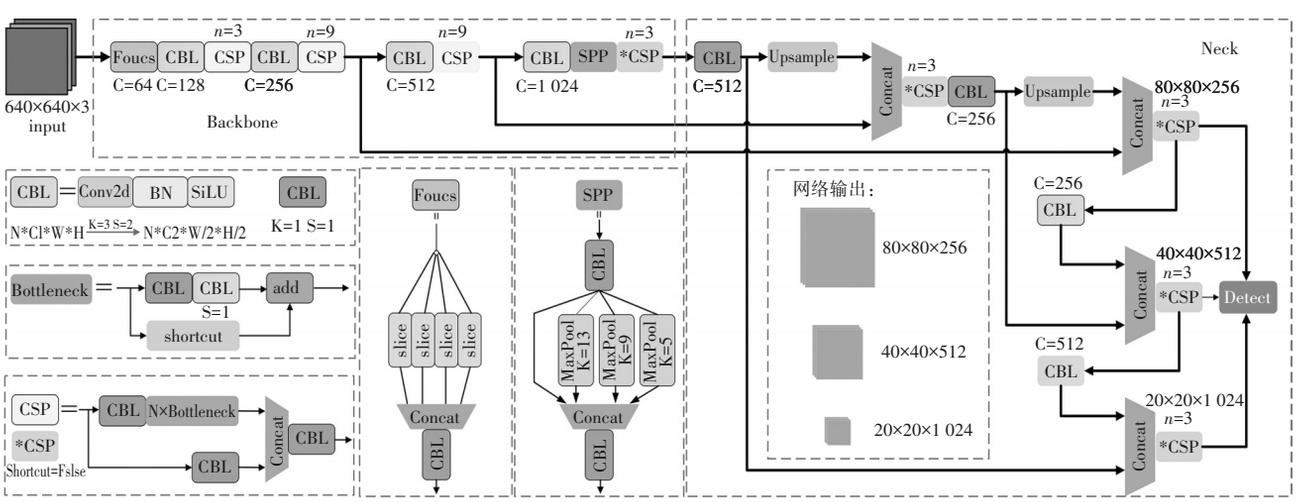


图 1 YOLOv5 网络结构

Fig.1 YOLOv5 Network Structure

进行随机裁剪拼接至一张图像中作为训练数据,使输入端同时获得四张图片的信息,一方面丰富了图像背景信息,另一方面也减少了模型对 Batch Size 的依赖. YOLOv5 的另一个特点是放弃了设定固定长宽初始锚框的思想,提出了自适应锚框,根据训练集的差异自适应计算最佳锚框数值. 在 Backbone 中,YOLOv5 主要用到了 Focus 结构、SPP 和 CSP 结构.

Focus 结构主要是将输入图像进行切片操作,增加特征通道数并减小特征尺寸,减少了 FLOPs(Floating Point Operations Per second),提升了运算速度,并且减少了模型层数. CSP 网络结构由 CBL 模块、Bottleneck 模块和 Concat 结构组成. 其中 CBL 模块由 Conv、Batch Normalization[11]、SiLu[12]激活函数构成. Bottleneck 模块包含 CBL 模块和残差连接[13],每个 CSP 结构中包含 n 个 Bottleneck 模块,并且 CSP 模块可通过设置是否使用残差连接生成两种不同的 CSP 结构. CSP 将输入特征分为不同分支,分别进行卷积使得特征通道数减半,其中一个分支进行 Bottleneck 操作,经过 Concat 将两个分支特征合并,增加特征图信息. SPP[14](空间金字塔池化)使用多个最大池化操作,对于不同的输入特征 SPP 结构都产生固定大小的输出. 在 YOLOv5 中 SPP 采用 [5,9,13] 三个尺度特征与输入特征进行融合,SPP 处理结果进一步提升了不同尺度和长宽比输入图像的尺度不变性.

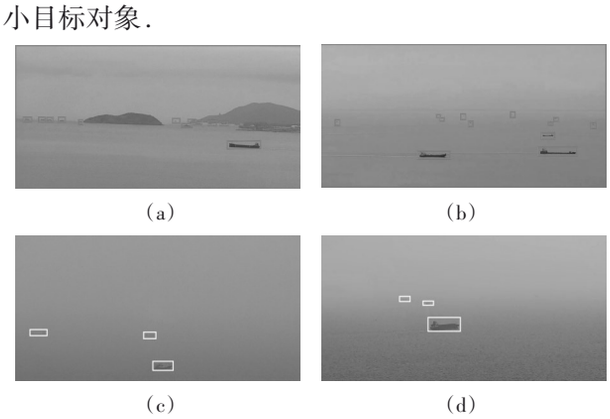
2　注意力机制

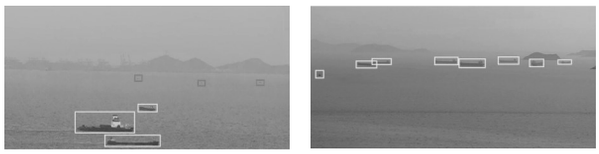
众所周知,卷积神经网络一直是计算机视觉任务中的主要方法,通过设计不同的网络结构和局部连接以丰富特征信息进而提升图像识别的性能,如全连接网络-FCN[15]、区域生成网络-RPN[1]、残差网络-Resnet[14]、级联网络-Cascade R-CNN[16]、特征金字塔网络-FPN[17]、空间金字塔池化-SPP[14]等在大规模图像识别项目中都获得了很好的效果. 但卷积神经网络中往往通过增加网络深度提高卷积神经网络的表示能力,在模型中大量的模块堆叠使得网络结构十分庞大. 与网络堆叠不同,注意力机制基于视觉感知过程,聚焦全局和局部特征,在减少网络参数的同时增强了图像特征信息.

作为注意力机制的代表著作,SENet[18]从全局获取上下文信息,显式地构建特征通道之间的相互依赖关系,输出结果逆向完成在通道维度上对原始特征的重新标定. ECANet[19]通过对 SENet 模块进行改进,提出了一种不降维的局部跨信道交互策略和自适应选择一维卷积核大小的方法实现了性能提升. SENet 在设计中并未引入空间维度进行特征融合,CBAM[20]注意力机制在此基础上沿通道、空间两个维度推断注意力,通过将注意力图与输入特征图融合以进行自适应特征优化. 同时,还有一些其他的注意力机制,如多谱通道压缩注意力方法-频域注意力网络 FcaNet[21]、位置像素注意力模块-Non-local[22]、CCNet[23]、双重注意力机制-DANet[24]等. 但以上注意力方法只能生成 1-D 或 2-D 注意力权重,这在一定程度上限制了注意力权重在通道和空间中的灵活性. Yang[25]等提出了一种即插即用的三维权重注意力模块 SimAM,能够直接评估特征的三维注意力权重,基于神经科学理论在已有的空间抑制理论的基础上,为每个神经元设计了能量函数评估其重要性.

值得一提的是,Vaswani[26]等创造性提出的 Transformer 机制在自然语言处理领域取得的巨大成功,吸引人们开始尝试将其应用于计算机视觉任务中. 主要的应用方向包括:1)将自注意力机制与 CNN 架构结合[22, 27],如 Zhu 等[28]提出 CABM 注意力机制与 Transformer 检测头结合的 YOLOv5 模型改进方法,在无人机捕获场景的目标检测中表现出良好的性能;Dai[29]等提出动态检测框架,采用多种注意力机制相结合的方法提升了目标检测头的表示能力. 2)使用自注意力机制完全替代卷积结构[30-31],如多头注意力机制[32-33]、稀疏 Transformer[34]和分层 Transformer[35].

本文根据第一种思想将 SimAM 注意力模块和 Transformer 融入 YOLOv5 网络,并提出多路径聚合网络以提升 YOLOv5 复杂海域目标检测和多目标分类能力. 复杂海域目标检测及目标分类面临的主要问题如图 2 所示. 图中(a)、(b)包含大量的小目标对象,并且呈现集群式分布;(c)、(d)图像为受海雾影响的海面目标;(e)、(f)包含了多种船只类型和部分小目标对象.





(e) (f)

图2 复杂海域目标信息

Fig.2 Complex sea object information

3 多路径聚合YOLOv5网络

3.1 多路径聚合网络(MPANet)

神经网络中特征信息的流动方式对结果影响明显,因为低层特征拥有精确的定位信息,高层特征拥有较强的语义信息.低层特征在向顶层特征传播过程中将会越来越难以获取准确的定位信息,而小目标信息随着特征传播将会逐渐丢失.特征金字塔网络(FPN)[36]通过自顶向下的横向连接方式丰富了每一层的特征信息,并在自顶向下传播过程中给出每一级的预测.FPN结构解决了目标检测中多尺度变化的问题,在小目标识别能力上有明显提升.PANet[37]在FPN网络结构基础上增加了自底向上的网络结构,保留横向连接的同时添加了路径增强和聚合,缩短了低层特征向顶层特征的传播路径,增强了低层特征传递能力,保留了精确的位置信息.

研究结果表明,特征融合的方法能够明显提升特征的分类能力[38].基于此,我们在PANet基础上对YOLOv5结构进行改进,我们继承了YOLOv5主干网络的同时增加了多路径聚合.每一个顶层特征都通过融合三个不同路径特征产生,本文称之为多路径聚合网络(MPANet),如图3所示.图中(a)和(b)部分是自顶向下的网络结构,通过横向连接进行特征融合.(c)部分是自底向上的网络结构,保持横向连接同时聚合低层特征,特征上传过程中给出每层特征的预测结果(d).MPANet保持自顶向下过程中的横向连接,使用多路径聚合方式将低层特征、中间特征与顶层特征相互融合,进一步丰富特征上下文信息并提升多尺度定位能力,保证顶层特征具有丰富语义信息的同时拥有精确的定位信息.从图4中可以看出,MPANet在顶层特征中聚合了三个不同路径的特征信息,在研究中,我们巧妙地设计了网络结构使这些特征拥有相同的尺寸和不同的通道数,融合特征通过1×1卷积后输入检测网络.MPANet中仍然保留三个检测头部,检测头部特征属性与YOLOv5保持一致,检测尺寸与通道数分别为[256,512,1 024]和[80,40,20].

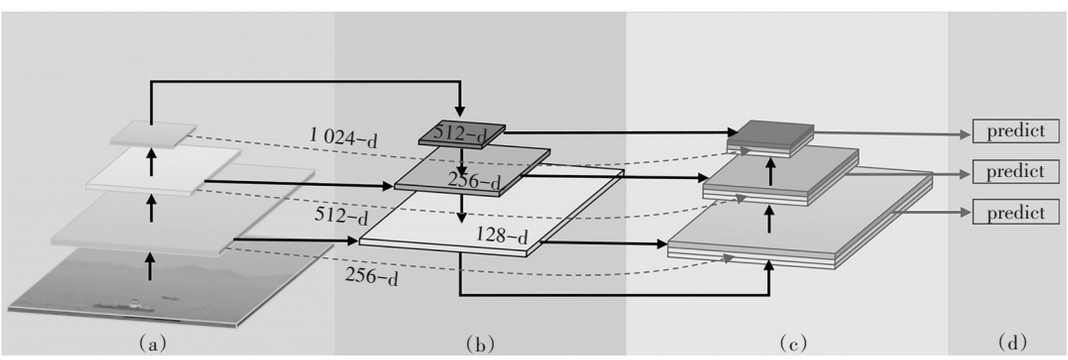


图3 MPANet

Fig.3 MPANet

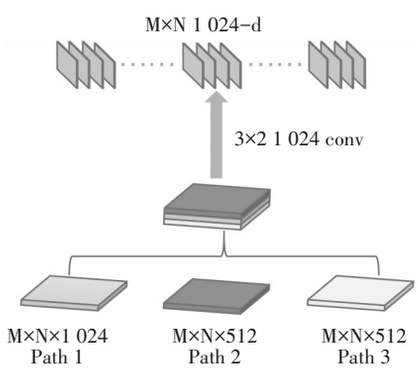


图4 顶层特征多路径聚合

Fig.4 Top-level feature multi-path aggregation

在MPANet-YOLOv5中,我们使用了在目标检测中效果更好的Mish[39]激活函数,如图5所示.Mish激活函数是一个非单调、有下界、无上界、正则化的平滑激活函数,它允许一部分负梯度流入保证信息流动,非单调平滑的特性保证了梯度下降效果较好.

3.2 MPANet-YOLOv5

我们将多路径聚合网络融入YOLOv5结构中,并在网络不同阶段添加了SimAM和Transformer注意力模块.SimAM是一个三维权重注意力模块,能够直接评估三维注意力权重,如图6所示. SimAM基于神

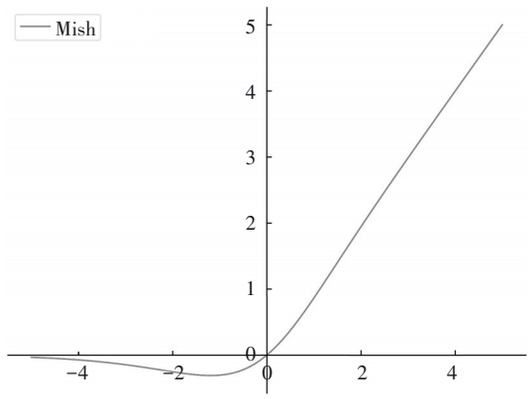


图 5　Mish 激活函数

Fig.5　Mish activation function

经科学理论在已有的空间抑制理论的基础上,为每个神经元设计了能量函数评估其重要性.能量函数定义如下:

$$e_t(\omega_t, b_t, y, x_i) = (y_t - \hat{t})^2 + \frac{1}{M-1}\sum_{i=1}^{M-1}(y_0 - \widehat{x_i})^2 \tag{1}$$

式中:$\hat{t} = \omega_t t + b_t$, $\widehat{x_i} = \omega_t x_i + b_t$ 分别是 t 和 x_i 的线性变换,可得到目标神经元 t 与其他神经元 x_i 在单一通道的线性可分离性,i 是空间维度索引,$M = H \times W$ 是通道中神经元数量,ω_t 和 b_t 是权重和偏置.整个模块的细化结果为:

$$\tilde{X} = \text{sigmoid}(\frac{1}{E}) \odot X \tag{2}$$

式中:E 将所有 e_t^* 跨通道和空间维度分组,Sigmoid 用来限制 E 中过大的值.

在 MPA-YOLOv5 模型中我们设计 SimAM 模块在自底向上过程中计算三维注意力权重,这有利于在多路径聚合时提供更丰富的特征信息.

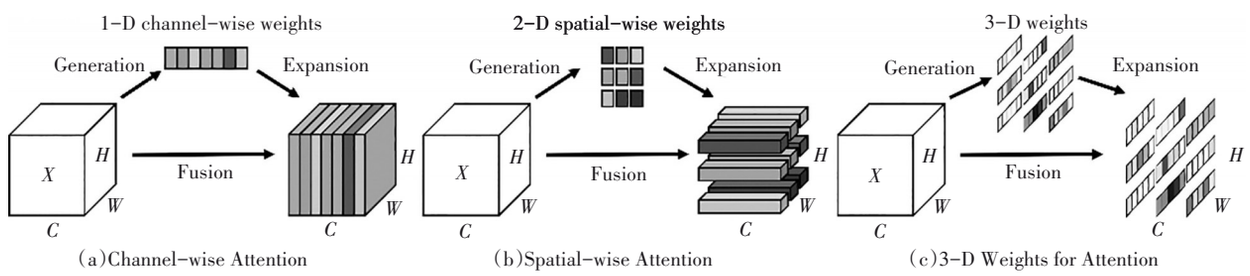


图 6　不同注意力机制比较

Fig.6　Comparison of different attention mechanisms

Transformer Encoder Block 为自注意力机制.Transformer 在目标检测领域已经有很多的应用,在本次实验中我们引入 Transformer encoder 在多路径聚合之前对特征进一步优化.Transformer encoder(图 7)能够捕获全局信息和丰富的上下文信息.作为多路径之一,Transformer encoder 的输出特征对多路径特征聚合结果具有正向增强作用.在本文中所使用的 Transformer encoder block 包含一个多头注意力层和一个 MLP 层.多头注意力机制能够使模型在不同的表示子空间学习到相关信息[26],而 MLP 能够阻止输出退化,增强自注意力机制的表达能力[40].

MPA-YOLOv5 在主干网络中依然继承 YOLOv5 的网络设计,在主干网络之后使用 SPP 固定输出尺寸,并将特征提取网络的高层特征输入 Transformer 中进一步增强特征信息.在 Neck 部分我们只使用了 1×1 的卷积结构用于降低图像通道数,并在每个上采样之后使用 CSP 进一步提取特征信息.上采样的特征经过 SimAM 注意力模块生成三维注意力权重并传递至 Transformer 模块.SimAM 注意力模块主要添加在至自底向上的特征传递过程中并保持横向连接,同时聚合低层特征信息进行特征融合.

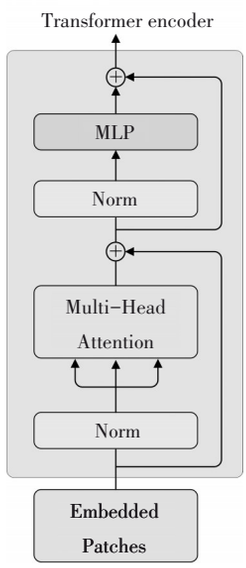


图 7　Transformer 编码器结构

Fig.7　Transformer encoder structure

多路径聚合过程中我们使用低层、中层和顶层三个层级的特征信息进行融合,三个路径特征尺寸相同,通道数不同,通过多路径聚合可得到预测所需特征.多路径聚合一方面保证了融合后特征信息具有定位信息和语义信息,另一方面缩短了低层次特征向顶层特征传递的距离.

在检测头部分依然保持[20,40,80]三个尺寸的特征输出并与自适应锚点匹配.添加SimAM注意力模块与Transformer的MPANet-YOLOv5网络结构如图8所示.

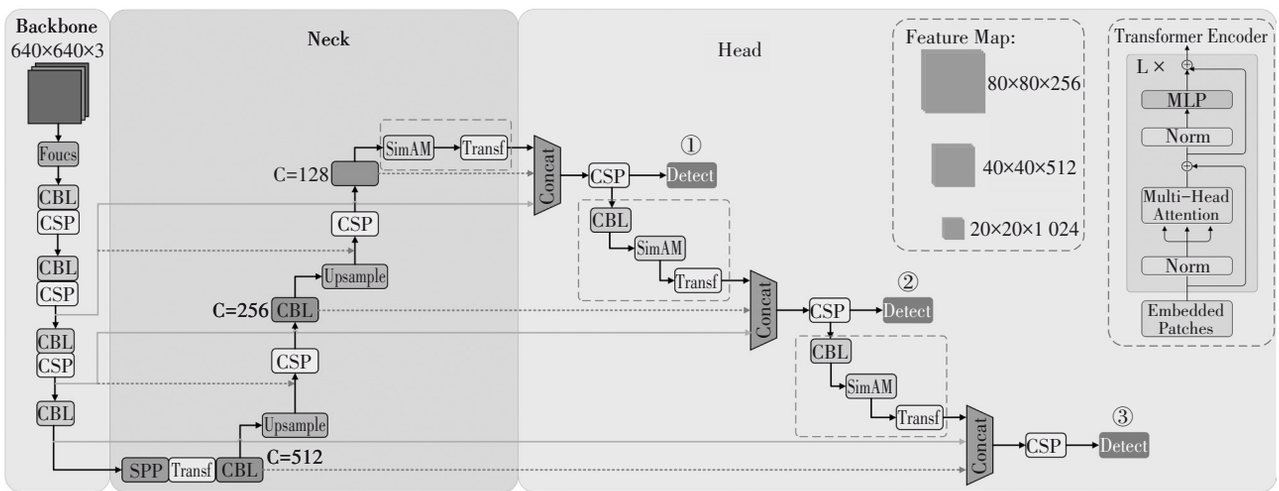


图8 MPANet-YOLOv5网络结构

Fig.8 MPANet-YOLOv5 network structure

4 实 验

本次研究采集了复杂场景下海面图像,以船舶作为主要的海上目标建立了自定义船舶检测数据集,并使用该数据集进行模型训练和测试.数据集包含五种目标类型(货船、集装箱船、渔船、游轮、岛屿)共计5 380张图片,训练集和测试集分别占80%和20%,在训练过程中使用640×640尺寸作为输入.

本文所有模型都在2张TITAN RTX 2080TI 24G GPU上训练和测试,MPANet-YOLOv5基于Pytorch 1.9深度学习框架.在对比实验中将MPANet-YOLOv5与YOLOv3、YOLOv4、YOLOv5、YOLOv5-SPPF和YOLOv5-Transformer算法进行对比,所有模型均设置300次迭代.通过模型测试结果可以发现:YOLOv5-Transformer仅在Backbone中添加Transformer后模型计算量有所下降,模型性能与基础YOLOv5模型相比平均精度上升了1.4%,召回率降低了1%,$AP_{0.5}$与$AP_{0.5:0.95}$有微弱下降,分别降低了0.2%和0.5%.YOLOv5 -SPPF使用卷积替换Focus结构,结合SPPF使模型计算量下降了0.6GFLOPs,精度提升了1%.MPANet-YOLOv5结果优于所有的对比模型,与YOLOv3相比,我们的模型精度提升了8.2%,召回率提升了1.2%,$AP_{0.5}$与$AP_{0.5:0.95}$分别提升了4.9%和3.8%.与YOLOv4模型相比,MPANet-YOLOv5精度提升了3.6%,召回率提升了3.8%,$AP_{0.5}$与$AP_{0.5:0.95}$分别提升了3%和1.9%.MPANet-YOLOv5与YOLOv5模型测试性能相比,AP提升了5.4%,召回率提升了3.3%,$AP_{0.5}$与$AP_{0.5:0.95}$分别提升了3.3%和2.2%.可以发现,MPANet-YOLOv5在性能表现上大幅领先,特别是与YOLOv5模型相比有明显提升.各模型具体测试结果如表1所示.

表1 模型测试结果对比

Tab.1 Comparison of model test results

模型	激活函数	backbone	Transformer	AP/%	R/%	$AP_{0.5}$/%	$AP_{0.5:0.95}$/%	FLOPs/G
YOLOv3	Leaky ReLu	Darknet-53		71.4	78.2	70.9	34.4	155
YOLOv4	Mish	CSPDarknet-53		76.0	75.6	72.8	36.3	120.6
YOLOv5	Leaky ReLu	Focus+CSP		74.2	76.1	72.5	36.0	16.4
YOLOv5-Transformer	Leaky ReLu	Focus+CSP+Transformer	√	75.6	75.1	72.3	35.5	16.2
YOLOv5 -SPPF	Leaky ReLu	Focus+CSP		75.2	75.3	73.6	36.1	15.8
MPANet-YOLOv5(ours)	Mish	Focus+CSP+Transformer+MPANet	√	79.6	79.4	75.8	38.2	18.2

在实际检测中,MPANet-YOLOv5小目标检测效果突出,在发生遮挡时也能准确地检测出目标类型.两种模型检测结果如图9所示.

(a)YOLOv5　　(b)MPANet-YOLOv5

图9　模型小目标检测结果对比

Fig.9　Comparison of model small target detection results

在研究中,我们分别在东海和南海通过两种不同方法对MPANet-YOLOv5复杂海域目标检测效果进行测试.1)在固定海域搭建检测站,对多条航线目标实时检测;2)通过在固定航线航行,获取航线上图像进行目标检测.两种测试结果如图10、图11所示.

图10　东海固定检测站点目标检测结果

Fig.10　East China Sea fixed detection site object detection results

图11　南海固定航线目标检测结果

Fig.11　South China Sea fixed route object detection results

5　结　论

本文提出了一种基于注意力机制的多路径聚合网络结构MPANet-YOLOv5,以提升YOLOv5在复杂海域的目标检测与分类能力.实验结果表明:使用多路径聚合网络将带有定位信息的低层特征与高层语义特征融合增强了多尺度定位能力,丰富了顶层特征语义信息和上下文信息,有效提升了YOLOv5复杂海域目标检测精度,特别是在小目标检测方面有明显的提升.事实证明:MPANet-YOLOv5在复杂海域场景中目标检测性能优于YOLOv3、YOLOv4、YOLOv5及其变体,是一种可靠的海上目标检测算法.

参考文献

[1] REN S 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, 2017, 39(6): 1137-1149.

[2] 张学军,黄爽,靳伟,等. 基于改进Faster R-CNN的农田残膜识别方法[J]. 湖南大学学报(自然科学版),2021,48(8):161-168.
ZHANG X J, HUANG S, JIN W, *et al*. Identification method of agricultural film residue based on improved faster R-CNN[J]. Journal of Hunan University (Natural Sciences), 2021, 48(8): 161-168.(In Chinese)

[3] REDMON J, FARHADI A. YOLO9000: better, faster, stronger[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA: IEEE, 2017: 6517-6525.

[4] REDMON J, FARHADI A. YOLOv3: an incremental improvement [EB/OL]. 2018: arXiv: 1804.02767 [cs. CV]. https://arxiv. org/abs/1804.02767

[5] BOCHKOVSKIY A, WANG C, LIAO H M. Yolov4: Optimal speed and accuracy of object detection [J]. arXiv: 2004.10934 [cs. CV]. https://doi.org/10.48550/arXiv.2004.10934

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

[7] 聂鑫,刘文,吴巍.复杂场景下基于增强YOLOv3的船舶目标检测[J].计算机应用,2020,40(9):2561-2570.
NIE X, LIU W, WU W. Ship detection based on enhanced YOLOv3 under complex environments[J]. Journal of Computer Applications, 2020, 40(9): 2561-2570.(In Chinese)

[8] 徐海祥,龙泽升,冯辉.面向船舶智能航行的多尺度目标检测算法[J].华中科技大学学报(自然科学版),2021,49(5):50-55.
XU H X, LONG Z S, FENG H. Multi-scale object detection algorithm for ship intelligent navigation[J]. Journal of Huazhong University of Science and Technology (Natural Science Edition), 2021, 49(5): 50-55.(In Chinese)

[9] 盛明伟,李俊,秦洪德,等.基于改进YOLOv3的船舶目标检测算法[J].导航与控制,2021,20(2):95-109.
SHENG M W, LI J, QIN H D, *et al*. Ship target detection algorithm based on the improved YOLOv3 [J]. Navigation and Control, 2021, 20(2): 95-109.(In Chinese)

[10] 梁月翔,冯辉,徐海祥. 面向智能船舶的水面小目标检测算法

[J]. 大连理工大学学报,2021,61(3):255-264.
LIANG Y X,FENG H,XU H X. Detection algorithm for small objects on water surface for intelligent ships[J]. Journal of Dalian University of Technology,2021,61(3):255-264.(In Chinese)

[11] IOFFE S, SZEGEDY C. Batch normalization: accelerating deep network training by reducing internal covariate shift [EB/OL]. 2015: arXiv: 1502.03167 [cs. LG]. https://arxiv. org/abs/1502.03167

[12] ELFWING S, UCHIBE E, DOYA K. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning[J]. Neural Networks,2018,107:3-11.

[13] HE K M,ZHANG X Y,REN S Q,*et al*. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas,NV,USA:IEEE,2016:770-778.

[14] HE K M,ZHANG X Y,REN S Q,*et al*. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2015,37(9):1904-1916.

[15] SHELHAMER E,LONG J,DARRELL T. Fully convolutional networks for semantic segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2017,39(4):640-651.

[16] CAI Z W, VASCONCELOS N. Cascade R-CNN: delving into high quality object detection[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA:IEEE,2018:6154-6162.

[17] LIN T Y, DOLLÁR P, GIRSHICK R, *et al*. Feature pyramid networks for object detection[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA: IEEE, 2017:936-944.

[18] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City,UT,USA: IEEE,2018.

[19] WANG Q L,WU B G,ZHU P F,*et al*. ECA-net: efficient channel attention for deep convolutional neural networks[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle,WA,USA:IEEE,2020:11531-11539.

[20] WOO S,PARK J,LEE J Y,*et al*. CBAM: convolutional block attention module[C]//Computer Vision-ECCV 2018.

[21] QIN Z Q,ZHANG P Y,WU F,*et al*. FcaNet: frequency channel attention networks[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal,QC,Canada:IEEE,2021:763-772.

[22] WANG X L,GIRSHICK R,GUPTA A,*et al*. Non-local neural networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City,UT,USA:IEEE,2018.

[23] HUANG Z, WANG X, WEI Y, *et al*. CCNet: Criss-Cross Attention for Semantic Segmentation[J]. IEEE Trans Pattern Anal Mach Intell, 2020. doi: 10.1109/TPAMI.2020.3007032.

[24] FU J, LIU J, TIAN H J, *et al*. Dual attention network for scene segmentation[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE,2019:3141-3149.

[25] YANG L,ZHANG R Y,LI L,*et al*. Simam: A simple, parameter-free attention module for convolutional neural networks[C]//International Conference on Machine Learning. PMLR, 2021.

[26] VASWANI A, SHAZEER N, PARMAR N, *et al*. Attention is all you need[C]//Advances in Neural Information Processing Systems. 2017:1-15.

[27] CARION N, MASSA F, SYNNAEVE G, *et al*. End-to-end object detection with transformers[C]//Computer Vision - ECCV 2020.

[28] ZHU X K, LYU S C, WANG X, *et al*. TPH-YOLOv5: improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios[C]//2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). Montreal, BC,Canada:IEEE,2021:2778-2788.

[29] DAI X Y, CHEN Y P, XIAO B, *et al*. Dynamic head: unifying object detection heads with attentions[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville,TN,USA:IEEE,2021:7369-7378.

[30] WANG H Y, ZHU Y K, GREEN B, *et al*. Axial-DeepLab: stand-alone axial-attention for panoptic segmentation [C]//Computer Vision-ECCV 2020.

[31] RAMACHANDRAN P, PARMAR N, VASWANI A, *et al*. Stand-alone self-attention in vision models[C]// Neural Information Processing Systems. 2019.

[32] ZHAO H S,JIA J Y,KOLTUN V. Exploring self-attention for image recognition[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). WA, USA: IEEE, 2020: 10073-10082.

[33] HU H,ZHANG Z,XIE Z D,*et al*. Local relation networks for image recognition[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE, 2019: 3463-3472.

[34] CHILD R, GRAY S, RADFORD A, *et al*. Generating long sequences with sparse transformers [EB/OL]. 2019: arXiv: 1904.10509[cs. LG]. https://arxiv. org/abs/1904.10509

[35] LIU Z,LIN Y T,CAO Y,*et al*. Swin transformer: hierarchical vision transformer using shifted windows[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal,QC, Canada:IEEE,2021:9992-10002.

[36] LIN T Y, DOLLÁR P, GIRSHICK R, *et al*. Feature pyramid networks for object detection[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA: IEEE, 2017:936-944.

[37] LIU S,QI L,QIN H F,*et al*. Path aggregation network for instance segmentation[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018:8759-8768.

[38] 王军,夏利民. 基于深度学习特征的异常行为检测[J]. 湖南大学学报(自然科学版),2017,44(10):130-138.
WANG J,XIA L M. Abnormal behavior detection based on deep-learned features[J]. Journal of Hunan University (Natural Sciences),2017,44(10):130-138.(In Chinese)

[39] MISRA D . Mish: A self regularized non-monotonic activation function[C]//British Machine Vision Conference. 2020.

[40] DONG Y H,CORDONNIER J B,LOUKAS A. Attention is not all You need: pure attention loses rank doubly exponentially with depth[EB/OL]. 2021: arXiv: 2103.03404[cs. LG]. https://arxiv. org/abs/2103.03404