【引用格式】蔡自清，王力，梁镜，等. 基于连续帧与注意力机制的水下小目标自主检测算法[J]. 数字海洋与水下攻防，2024，7（5）：529-535.

基于连续帧与注意力机制的水下小目标自主检测算法

蔡自清[1, 2]，王　力[1, 2]，梁　镜[1, 2]，李孟霏[1, 2]，徐凯凯[1, 2]
（1. 中国船舶集团有限公司第七一〇研究所，湖北 宜昌 443003；
2. 清江创新中心，湖北 武汉 430200）

摘　要　针对声呐对水下小目标探测时目标特征少，常规目标检测算法性能不佳的问题，提出了一种以 YOLOv5 目标识别算法为基础的连续帧识别改进算法。该算法通过连续帧数据提取模块以及轻量的通道空间注意力模块，提取声呐连续帧信息，提升了 YOLOv5 算法的识别能力。湖上前视声呐时序数据集算法验证试验表明，在几乎不增加推理时间的前提下，改进算法的平均检测精度比 YOLOv5 算法提升了 13.7%。该改进算法预期可在水下小目标自主检测任务中应用。

关键词　YOLOv5；小目标自主检测；连续帧信息；注意力机制
中图分类号　TP242.6　　**文献标识码**　A　　**文章编号**　2096-5753(2024)05-0529-07
DOI　10.19838/j.issn.2096-5753.2024.05.009

Underwater Small Target Detection Based on Continuous Frame and Attention Mechanism

CAI Ziqing[1, 2]，WANG Li[1, 2]，LIANG Jing[1, 2]，LI Mengfei[1, 2]，XU Kaikai[1, 2]
（1. No. 710 R&D Institute，CSSC，Yichang 443003，China；
2. Qingjiang Innovation Center，Wuhan 430200，China）

Abstract　In order to solve the problems of few target features of underwater small targets collected by sonar，and poor performance of conventional target detection algorithm，an improved continuous frame recognition algorithm based on YOLOv5 is proposed. The algorithm uses a continuous frame data extraction module and a lightweight channel spatial attention module to extract sonar continuous frame information，which improves the recognition ability of YOLOv5 algorithm. The lake experimental results based on the forward-looking sonar time series dataset show that the accuracy of the algorithm is improved by 13.7%，and the reasoning time is basically unchanged. The improved algorithm is expected to be applied to the autonomous detection of underwater small targets.

Key words　YOLOv5；underwater small target detection；continuous frame information；attention mechanism

0　引言

近年来，水下无人机器人技术不断发展，随着水下机器人的无人化、智能化程度的提高，基于水下探测器的智能感知技术的重要性也随之提高。

水下小目标自主检测技术能有效降低相关任务上的人工成本，在民用及军用领域均具备良好的发展与应用前景。

收稿日期：2024-08-08
作者简介：蔡自清（1997-），硕士，助理工程师，主要从事声呐目标检测算法研究。

水下声探测作为水下目标探测的主要手段，存在图像分辨率低，图像细节不清晰等问题。传统的水下小目标检测方法主要通过人工特征提取与分类器进行检测，人工特征主要有尺度不变特性变换[1]、方向梯度直方图[2]、局部二值模式[3]等方式，分类器主要有 Adaboost 迭代[4]、支持向量机[5]等算法。而随着人工神经网络的兴起，目标检测算法也由传统的人工特征提取检测转变为基于深度卷积神经网络[6]的智能特征提取检测。RCNN[7]、Fast-RCNN[8]、Faster-RCNN[9]、SDD[10]、YOLOv3[11]、YOLOv5[12]等算法被相继提出，并在光学图像目标检测任务上取得了优异的成效。近年来，相关算法也逐渐被应用于水下小目标检测上，并取得了一定的成效。

针对水下小目标自主检测任务中存在的目标尺度小与图像特征少的问题，设计了一种改进 YOLOv5 算法。该算法基于声探测的数据连续采集的特性进行设计，使用数据处理模块对声呐数据连续帧信息进行整合，替代单张图片作为算法的输入；设计了一种基于注意力机制的特征提取模块进行连续帧特征提取。相较于 YOLOv5 算法，本文算法极大提高了水下小目标自主检测的准确性。

1 基于连续帧信息与注意力机制的改进目标检测算法

1.1 基线网络

本文使用目标检测算法为 YOLOv5 网络的小参数量版本 YOLOv5s 进行基线网络的搭建。

YOLOv5 算法是一种高性能、高效率的目标检测算法，基于深度学习和神经网络技术。该算法通过创新的设计和优化，在处理大规模图像数据时展现出卓越的能力。其主要由主干网络、融合层和目标检测头 3 个部分组成，其模型结构如图 1 所示。

YOLOv5 主干网络采用了创新的 C3 模块和 SPP[13]模块。

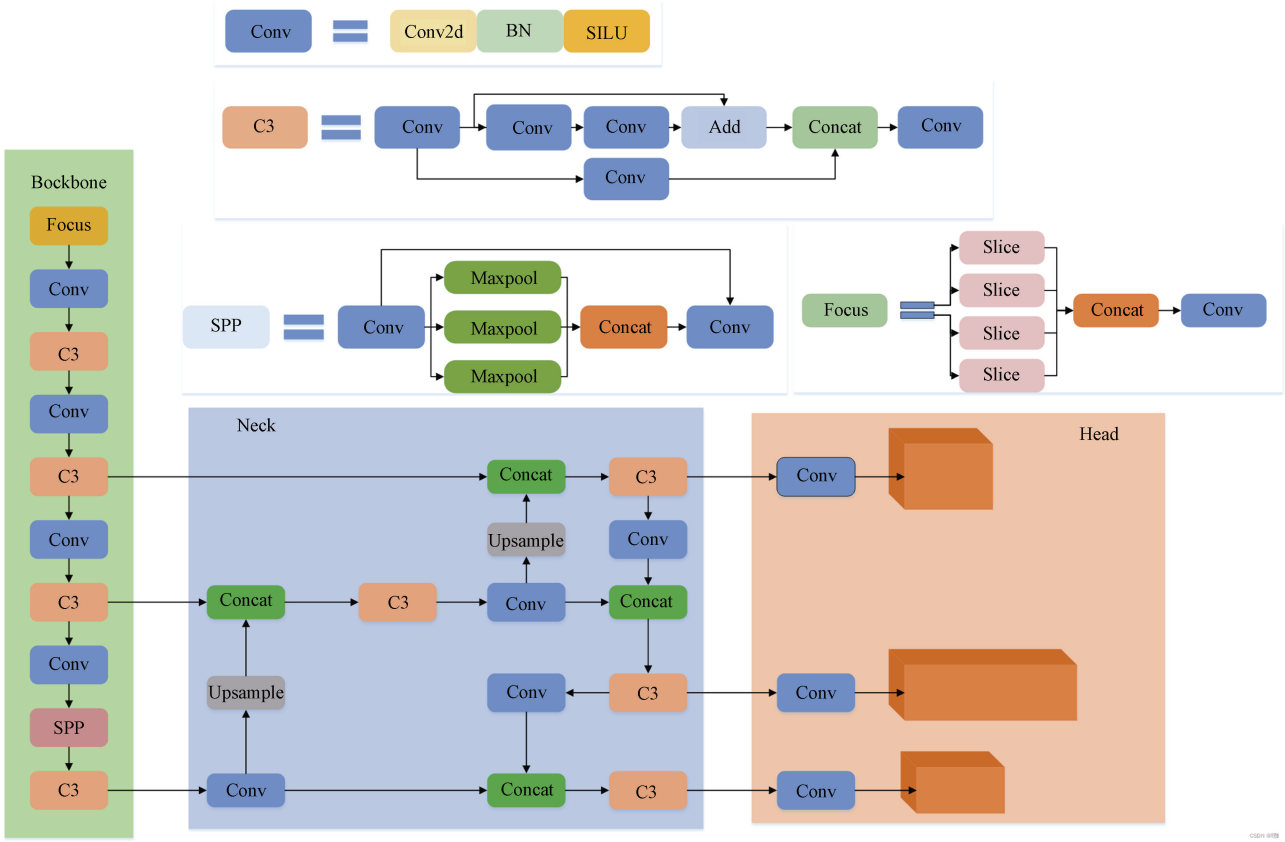

图 1 YOLOv5 算法网络结构

Fig. 1 Network structure of YOLOv5 algorithm

C3 模块是 YOLOv5 主干网络的核心组成部分，其相较于传统的卷积堆叠方法，通过跨阶段部分连接实现了特征图的融合，从而充分利用不同层级的特征信息，提高了目标检测的准确性和感受野。

SPP 模块通过多层级的空间金字塔池化操作，实现了多尺度特征池化，能够处理不同尺度的目标物体。这样可以在保持输入特征图尺寸不变的情况下，获取到更全局和更细粒度的特征信息，提高了目标检测的准确性。

YOLOv5 的融合层主要负责提取多尺度特征并进行特征融合，以便更好地检测不同大小的目标。其采用了一种轻量级的特征金字塔[14]的结构，由一个下采样层和一个上采样层组成。在融合层中，每个下采样层和上采样层之间都有一个特征融合模块，用于融合不同尺度的特征图。这种多尺度的特征融合使得 YOLOv5 可以有效地检测不同大小的目标，并提高检测性能。

目标检测头部分用于对融合层进行多尺度目标检测。

本文的改进目标检测算法沿用 YOLOv5s 的基本架构，将 YOLOv5s 网络作为基线网络，通过模块设计以及优化，提升算法对于水下小目标的检测能力。

1.2　图像连续帧信息提取模块

传统目标检测算法通常基于光学图片进行设计，针对单张图片进行解析，对图片中的目标特征进行提取，用于检测与定位目标。光学图片具有细节丰富，目标尺度相对较大等优点，使用传统目标检测算法一般能够取得较为优异的检测效果。

而声探测图像存在的细节信息少、目标尺度极小的问题，使用传统的单图检测策略进行目标检测时，由于信息不足的问题，往往难以取得较好的检测效果。本文基于实际探测过程中获取数据具有序列连续性的特点，设计了一个高效的图像连续帧信息提取模块。

声呐数据在解析生成图片的过程中一般由信号强度信息通过特定的仿射变换形成伪彩图，存在一定的信息冗余性。

该模块同时接收当前帧以及当前帧前 2 帧的图像信息，根据声呐图片解析的特性，使用固定编码的方式进行反解码，将其通道数由三通道压缩至单通道，获取 3 帧精炼的图像信息，固定编码压缩过程如图 2 所示。该压缩过程具有可逆性，不会产生信息损失。

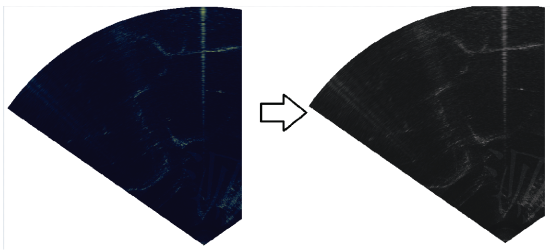

图 2　固定编码数据压缩精炼效果示意图

Fig. 2　Schematic diagram of compression effect for fixed coding data

随后将压缩后获取的精炼图像数据进行通道上的叠加，获取如图 3 所示的包含连续帧信息的多维输入。将多维输入送入网络后，网络通过自主学习不同通道间的差异性，可以自主对声探测图像中容易误识别的偶发干扰进行滤除；并且多帧信息中的目标特征，会通过叠加得到增强，从而得到更丰富的目标特征用于检测与识别。

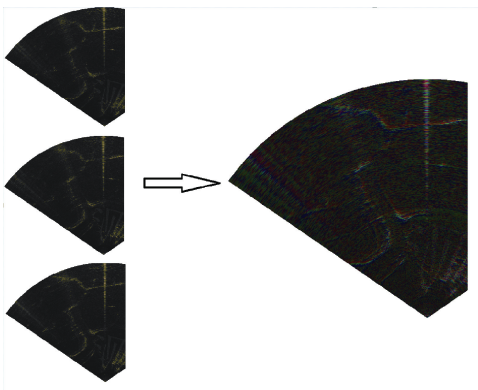

图 3　多维连续帧信息整合

Fig. 3　Integration of multi-dimensional continuous frame information

1.3　通道空间注意力模块

本文针对多维连续帧信息的特性对目标检测算法的特征提取模块进行了改进设计，设计了一种高效的通道空间注意力模块，用以对多维连续帧信息进行提取。

其中，使用通道注意力部分高效整合不同通道之间的差异性以及依赖性，达到针对性的解析获取不同帧信息之间的特异性以及相关性；使用空间注意力部分获取图片不同位置的重要性信息，两者通过级联进行整合。

该模块可以使算法自主可以对偶发干扰与任务目标进行多维度的解析，并针对性地对多维信息中的丰富目标特征信息进行关注，从而达到更优的检测效果。

1.3.1 通道注意力部分

通道注意力模块于 2017 年，由论文压缩激活网络（Squeeze-and-Excitation Networks，SENet）[15]提出，并被命名为 SE 模块。

SE 模块的结构如图 4 所示。

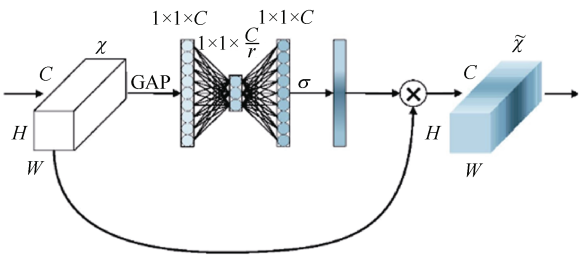

图 4 SE 模块结构图

Fig. 4 Schematic diagram of SE module

SE 模块通过全局信息嵌入以及通道自适应矫正可以获取高维特征的不同间的相关性，可以有效提高模型对特征的提取整合能力，但是其也存在一些问题。SE 模块仅使用全局平均池化进行全局信息嵌入导致很多典型特征例如峰值特征被忽视；使用了全连接层进行通道自适应矫正，造成信息损失的同时还会增加较大的运算量。

针对 SE 模块中存在的问题，本文通道注意力部分进行了优化，结构如图 5 所示。

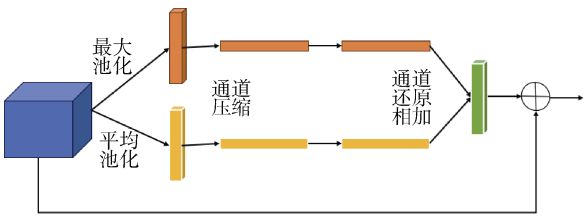

图 5 改进通道注意力模块结构图

Fig. 5 Schematic diagram of improved channel attention module

通道注意力部分对输入特征使用全局平均池化、全局最大池化的方式进行全局信息混合嵌入，以获取不同通道的不同关注度的全局表征。

使用一维卷积代替全连接层，减少信息丢失的同时，使用更低的运算量进行通道自适应矫正，最终使用聚合模块以及归一化模块进行整合，获取混合通道注意力。

在进行一维卷积设计时，设计卷积核大小与输入特征通道进行匹配，使两者满足公式（1）比例要求。

$$X = \text{int}\left|\text{Log}_2\frac{C}{2}+\frac{1}{2}\right| \qquad (1)$$

式中：X 为卷积核大小；C 为输入特征通道。

改进通道注意力模块，通过一维卷积代替全连接层进行跨通道的信息的捕获，减少了参数量的同时还减少了信息在全连接压缩过程种的损失；使用一种映射规则，将一维卷积的感受野与特征的通道数进行了合理的匹配；引入了全局最大池化作为全局信息嵌入的补充，让改进通道注意力能够获取连续帧信息输入中的更加丰富的特征信息。

1.3.2 空间注意力部分

空间注意力的思想最早于 2015 年，由论文空间转换网络（Spatial Transformer Networks，STN）[16]提出。2018 年，注意力块网络（Bottlenet attention Module，BAM）[17]与卷积注意力模块网络（Convolutional Block Attention Module，CBAM）[18]基于该思想，设计了轻量可靠的空间注意力模块。其中 CBAM 网络的空间注意力模块的结构如图 6 所示。

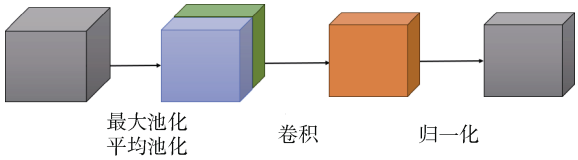

图 6 CBAM 空间注意力模块结构图

Fig. 6 Schematic diagram of CBAM spatial attention module

CBAM 空间注意力模块通过最大池化与均值池化的方式获取每个像素点的空间重要程度信息，可以以极低的运算量获取较为明显的性能提升，但是其也有很多不足。CBAM 空间注意力模块信息

丢失较大，且进行空间重要程度评估时，未对像素点周边小范围其他像素点进行考虑。

针对 CBAM 空间注意力中的不足，本文空间注意力部分进行了优化调整，其结构如图 7 所示。

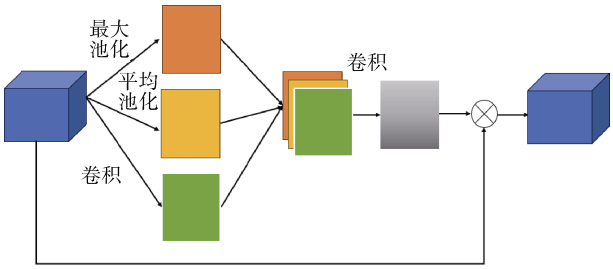

图 7　改进空间注意力模块结构图

Fig. 7　Schematic diagram of improved spatial attention module

空间注意力部分对特征图中的各个位置分别使用通道最大池化、通道平均池化与卷积自适应池化进行不同关注方向的空间重要程度信息提取，获取了多类关注方向的特征图。

将 3 个特征图进行通道上的叠加后，使用卷积以及归一化模块，将多类空间重要程度信息进行进行整合，获取混合空间重要程度信息。

将混合空间重要程度信息与原始输入进行加权残差求和，获取最终的输出结果。

改进空间注意力模块同时使用特征图各个位置上的通道平均值表征、通道最大值表征以及卷积自适应学习到的通道表征，从不同的关注方向获取空间重要程序信息，使新型空间注意力模块关注连续帧信息输入中的重要位置的信息，针对性的进行信息解析。

2　试验与分析

2.1　试验数据集

本文基于真实的声呐数据进行了前视声呐时序数据集的搭建。

数据通过湖上测试采集获得，采用蓝衡前视声呐，数据采集共计耗时 6 d。数据采集完成后，由人工对试验数据进行整理，选取存在目标的直线航次中的连续序列图片进行标注，并在命名时通过序列号与帧号用以表示单张图片在某一序列中的时序信息。

数据集共计包含 1 778 张分辨率为 1 200×600 的声呐图像，数据集对人工目标进行了标注。

数据集由多个序列的声呐图片组成，不同序列的长度分布如图 8 所示，图中横轴代表序列的长度，纵轴代表该长度的序列的个数。

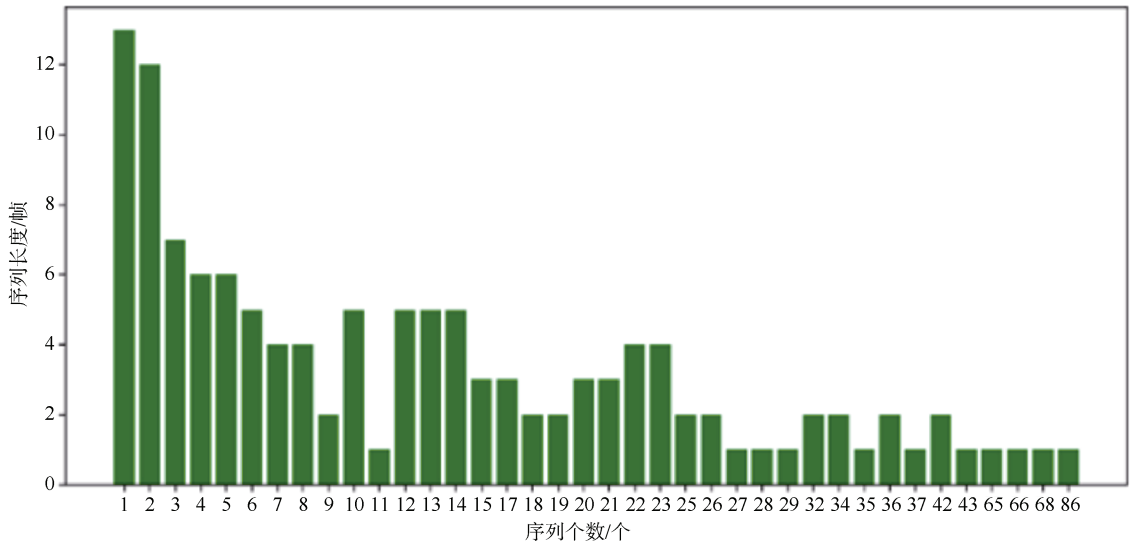

图 8　数据集序列长度分布图

Fig. 8　Dataset sequence length distribution map

以单个序列为最小单位，将数据集尽可能以 2∶1 的比例划分为训练集与验证集，最终共获取训练集 1 142 张带标注图像，验证集 636 张带标注图像。

数据集典型图片示例如图 9 所示。

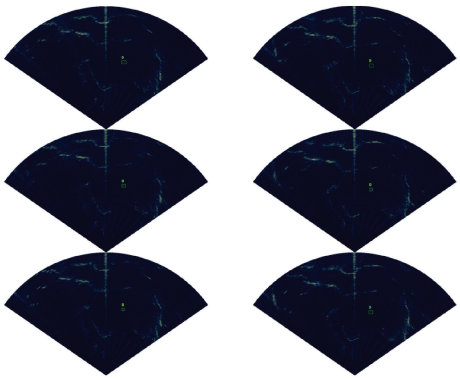

图 9 数据集典型图片示例
Fig. 9 Typical images of dataset

2.2 试验数据处理以及训练参数设置

2.2.1 试验数据处理

本文对前视声呐时序数据集原始分辨率为 1 200×600 的声呐图片进行裁剪与缩放处理，转换为 2 张分辨率为 640×640 的图片并对标签进行对应转换。

本文使用扩充后 2 284 张声呐图片作为训练集，1 272 张声呐图片作为验证集。

2.2.2 训练参数设置

试验相关硬件设备平台与软件版本参数如表 1 所示。

表 1 试验硬件平台与软件版本
Table 1 Test hardware platform and software versions

内容	明细
操作系统	Ubuntu 20.04
GPU 型号	Nvidia GeForce RTX 3090
GPU 显存	24G
CUDA 版本	11.3
编程语言	Python
深度学习框架	PyTorch

进行算法训练时，设置输入图片大小为 640×640，批处理大小为 16，训练轮次为 300，初始学习率为 0.01。并开启图像旋转、平移、缩放、扭曲等用于训练过程中的数据增强。

2.3 图像连续帧信息提取模块性能验证试验

本小节使用 YOLOv5s 算法作为图像连续帧信息提取性能验证试验的基线，使用前视声呐时序数据集对引入图像连续帧信息提取模块前后的算法进行训练。

使用验证集数据对算法性能进行验证，试验结果如表 2 所示。

表 2 图像连续帧信息提取模块性能验证试验结果
Table 2 Image continuous frame test results

算法	mAP50	准确率	召回率	运算速度/ms
YOLOv5s	0.368	0.434	0.463	3.6
YOLOv5s+连续帧	0.451	0.534	0.491	3.6

试验结果表明，使用图像连续帧信息提取模块后，在不增加运算量的同时引入了更多的初始信息，使算法在保证相同运算速度的同时，获得了较大的性能提升，充分说明了该模块的有效性。

2.4 通道空间注意力模块性能验证试验

本小节使用 2.3 小节试验的优势算法作为通道空间注意力模块性能验证试验的基线，用于验证该模块对图像连续帧信息提取的增强作用。

使用验证集数据对算法性能进行验证，试验结果如表 3 所示。

表 3 通道空间注意力模块性能验证试验结果
Table 3 Experimental results of channel spatial attention module

算法	mAP50	准确率	召回率	运算速度/ms
YOLOv5s+连续帧	0.451	0.534	0.491	3.6
YOLOv5s+连续帧+空间通道注意力	0.505	0.551	0.535	3.7

试验结果表明，使用空间通道注意力模块后，算法可以更加高效且准确的利用连续帧信息中的额外特征，进一步发挥连续帧信息的优势性，获得更好的识别效果，进一步佐证了注意力机制与连续帧信息的小目标检测网络的优势性。

3 结束语

本文针对常规目标检测算法 YOLOv5 在水下小目标自主检测任务上表现较差的问题，对 YOLOv5 算法进行了改进优化，提出了一种基于连续帧信息和注意力机制的水下小目标检测改进算法。改进算法对声呐连续帧数据进行高效整合处理后作为算法输入，根据连续帧输入的特性设计了通道空间注

意力模块用于强化算法对相关特征的提取与解析功能，在几乎不增加推理时间的前提下，改进算法的平均检测精度（mAP）相较于 YOLOv5s 提高了 13.7%。试验结果表明，该改进算法可以在保持较快推理速度优势的同时，有效提高算法对水下小目标自主检测性能。

参考文献

[1] JUAN L，GWUN O. A comparison of SIFT，PCA-SIFT and SURF[J]. International Journal of Image Processing（IJIP），2009，3（4）：143-152.

[2] DALAL N，TRIGGS B. Histograms of oriented gradients for human detection[C]// 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego：IEEE，2005.

[3] PAUL M，HAQUE S M E，CHAKRABORTY S. Human detection in surveillance videos and its applications-a review[J]. EURASIP Journal on Advances in Signal Processing，2013，2013（1）：1-16.

[4] WEN X Z，SHAO L，XUE Y，et al. A rapid learning algorithm for vehicle classification[J]. Information Sciences，2015，295：395-406.

[5] CORTES C，VAPNIK V. Support-vector networks[J]. Machine Learning，1995，20：273-297.

[6] LI Z W，LIU F，YANG W J，et al. A survey of convolutional neural networks：analysis，applications，and prospects[J]. IEEE Transactions on Neural Networks and Learning Systems，2021，33（12）：6999-7019.

[7] GIRSHICK R，DONAHUE J，DARRELL T，et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]// 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus：IEEE，2014.

[8] GIRSHICK R. Fast R-CNN[C]// 2015 IEEE International Conference on Computer Vision. Santiago：IEEE，2015.

[9] 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，2016，39（6）：1137-1149.

[10] LIU W，ANGUELOV D，ERTAN D，et al. SSD：single shot multibox detector[C]// Computer Vision-ECCV 2016. Amsterdam：IEEE，2016.

[11] REDMON J，FARHADI A. YOLOv3：an incremental improvement[EB/OL]. [2018-03-26]. https：//pjreddie.com/media/files/papers/YOLOv3.pdf.

[12] ZHANG Y，GUO Z Y，WU J Q，et al. Real-time vehicle detection based on improved YOLOv5[J]. Sustainability，2022，14（19）：12274.

[13] 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，2014，37（9）：2389824.

[14] 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：IEEE，2017.

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

[16] JADERBER M，SIMONYAN K，ZISSERMAN A. Spatial transformer networks[J]. Advances in Neural Information Processing Systems，2015，28：2017-2025.

[17] PARK J，WOO S，LEE J Y，et al. BAM：bottleneck attention module[EB/OL]. [2018-07-17]. https：//joonyoung-cv.github.io/assets/paper/18_bmvc_bam.pdf.

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

（责任编辑：张曼莉）