计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2024,33(12):131−140 [doi: 10.15888/j.cnki.csa.009684] [CSTR: 32024.14.csa.009684]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于改进 RT-DETR 的水下目标检测①

张　路, 魏本昌, 魏鸿奥, 周龙刚

(湖北汽车工业学院 电气与信息工程学院, 十堰 442002)
通信作者: 魏本昌, E-mail: bc_david@163.com

摘　要: 水下目标检测技术在海洋探测中具有重要的现实意义. 针对水下场景复杂, 以及存在遮挡重叠导致目标特征提取有限的问题, 提出了一种适用于水下目标检测的 FERT-DETR 网络. 该模型首先提出了一种特征提取模块 Faster-EMA, 用于替换 RT-DETR 中 ResNet18 的 BasicBlock, 能够在有效降低模型参数量和模型深度的同时, 显著提升对水下目标的特征提取能力; 其次在编码部分使用级联群体注意力模块 AIFI-CGA, 减少多头注意力中的计算冗余, 提高注意力的多样性; 最后使用高水平筛选特征金字塔 HS-FPN 替换 CCFM, 实现多层次融合, 提高检测的准确性和鲁棒性. 实验结果表明, 所提算法 FERT-DETR 在 URPC2020 数据集和 DUO 数据集上比 RT-DETR 检测准确率提高了 3.1% 和 1.7%, 参数量压缩了 14.7%, 计算量减少了 9.2%, 能够有效改善水下复杂环境中不同尺寸目标漏检、误检的问题.

关键词: 计算机视觉; RT-DETR; FasterNet; 注意力机制; 高水平筛选特征金字塔

引用格式: 张路,魏本昌,魏鸿奥,周龙刚.基于改进 RT-DETR 的水下目标检测.计算机系统应用,2024,33(12):131–140. http://www.c-s-a.org.cn/1003-3254/9684.html

Underwater Target Detection Based on Improved RT-DETR

ZHANG Lu, WEI Ben-Chang, WEI Hong-Ao, ZHOU Long-Gang

(College of Electrical & Information Engineering, Hubei University of Automotive Technology, Shiyan 442002, China)

Abstract: Underwater target detection has practical significance in ocean exploration. This study proposes a FERT-DETR network suitable for underwater target detection to address the issues of complex underwater environments and limited target feature extraction due to occlusion and overlap. The proposed model first introduces a feature extraction module, Faster EMA, to replace the BasicBlock of ResNet18 in RT-DETR, which can significantly improve its capability to extract features of underwater targets while effectively reducing the number of parameters and depth of the model. Secondly, a cascaded group attention module, AIFI-CGA, is used in the encoding part to reduce computational redundancy in multi-head attention and improve attention diversity. Finally, a feature pyramid for high-level filtering named HS-FPN is used to replace CCFM, achieving multi-level fusion and improving the accuracy and robustness of detection. The experimental results show that the proposed algorithm, FERT-DETR, improves detection accuracy by 3.1% and 1.7% compared to RT-DETR on the URPC2020 and DUO datasets respectively, compresses the number of parameters by 14.7%, and reduces computational complexity by 9.2%. It can effectively avoid missed and false detection of targets of different sizes in complex underwater environments.

Key words: computer vision; RT-DETR; FasterNet; attention mechanism; high-level screening-feature fusion pyramid (HS-FPN)

① 基金项目: 湖北省教育厅项目 (B2019077)
收稿时间: 2024-05-08; 修改时间: 2024-05-29; 采用时间: 2024-06-04; csa 在线出版时间: 2024-10-25
CNKI 网络首发时间: 2024-10-25

海洋是世界上最大的资源库, 占据着地球表面绝大部分的面积, 蕴含着丰富的资源, 如石油、天然气、海产品等[1]. 为满足人类社会的快速发展, 人们不断对海洋进行开发. 近 10 年来, 相关水下机器人和探测技术得到了快速的发展, 例如配备了智能水下目标检测系统的自主潜水器[2]和遥控潜水器, 在海洋资源的开发以及保护的过程中, 起到了至关重要的作用. 随着计算机视觉的发展, 目标检测开始愈发重要, 而水下目标检测分为声学系统检测和光学系统检测[3], 同声学图像相比, 光学图像具备更高的分辨率以及信息量, 并且在获取的方法上更加轻易[4]. 因此, 基于光学系统的水下目标检测更加受到人们关注. 目标检测是计算机视觉的一个核心的分支, 现有的目标检测的方法, 大致可以分为两大类: 传统的目标检测方法和基于深度学习的目标检测方法[5].

传统的目标检测方法通常分为 3 个阶段: 区域选择、特征提取和特征分类[6]. 首先通过滑动窗口选择感兴趣的区域[7]. 然后, 应用尺度不变特征变换 (SIFT)[8]、定向梯度直方图 (HOG)[9]等多种特征提取算法提取每个感兴趣区域的特征. 最后, 采用支持向量机 (SVM) 等机器学习算法[10]对提取的特征进行分类, 判断窗口是否包含对象. 然而, 由于传统方法需要设计各种大小的窗口, 并且依赖于机器学习方法进行分类, 因此存在一些局限性.

深度学习的出现改变了目标检测领域, 它的高速度和通用性使其广泛应用于许多领域, 不需要人工干预, 减少了人为因素造成的错误. 目前, 基于深度学习的目标检测算法主要分为两大类: 基于区域提议的算法和基于回归的算法. 前一类又称为两阶段目标检测算法, 这一类的代表性算法包括 R-FCN (region-based fully convolutional networks)[11]和 R-CNN (region-CNN) 系列算法 (R-CNN[12], Fast R-CNN[13], Faster R-CNN[14], Mask-R-CNN[15], Cascade-R-CNN[16]等). Li 等人[17]将 Fast R-CNN 应用于水下图像的识别和检测, 使用新的数据集共 12 种水下生物, 相较于 R-CNN 平均精度有所提高. Zeng 等人[18]将对抗性遮挡网络 (AOV) 运用到 Faster R-CNN 当中, 开发了一种新的框架, 用于水下海产品的目标检测. 尽管基于区域的算法有较高的精度, 但它们往往速度较慢, 无法做到实时检测. 而后者基于回归的目标检测算法, 又称为一阶段目标检测算法, 可以从图像中直接预测目标的位置和类别, 而不需要进行额外的识别或提取区域的步骤, 突出了强大的实时处理的能力. 这一类的代表算法包括 SSD (single shot multibox detector)[19], 和 YOLO (you only look once)[20]算法家族. 强伟等人[21]将 ResNet 作为基础网络的 SSD 检测模型, 用于水下目标检测. Li 等人[22]开发了一种基于 YOLOv3 的浮游生物的检测网络, 该网络采用了密集连接的结构, 便于特征传递. 黄廷辉等人[23]提出了一种基于 F-CBAM 注意力机制的 YOLOv5 水下目标检测 FAttention-YOLOv5 模型. Wang 等人[24]利用加权 ghost-CSPDarknet 和简化的 PANet, 提出了一种轻量化的水下目标检测网络 LUO-YOLOX.

随着 Transformer 在计算机视觉领域的普及, Facebook 的研究人员巧妙地利用 Transformer 架构提出了一个新的目标检测器 DETR[25]. 它可以通过 Transformer 学习到全局特征, 将目标检测视为集合预测问题, 减少了很多人工先验知识, 不需要手工设计的组件, 如非极大值抑制 (NMS) 和锚框生成, 从而实现了端到端的目标检测. 然而, DETR 存在训练周期长, 收敛速度慢等问题, 且对于小目标的检测能力仍有很大的进步空间.

针对上述问题, 为解决水下目标的检测精度低, 对小目标的检测性能较差, 且实时性较差的问题, 本文提出了 FERT-DETR 目标检测架构, 更加轻量化且能够有效地提升精准度. 该方法的主要改进如下.

(1) 将 RT-DETR 中的骨干网络 ResNet18 替换为 Faster Block, 降低了网络参数量, 且在其中使用高效多尺度 (EMA) 注意力, 提升骨干网络的提取效率.

(2) 在 AIFI 中使用级联群体注意力模块 (CGA), 解决了多头注意力机制中计算冗余的问题, 提高计算效率, 通过增加模型深度提升模型容量.

(3) 将 CCFM 结构进行重写, 使用高水平筛选特征金字塔 (HS-FPN)[26], 实现跨尺度连接, 多层次融合. 增强了模型的特征表达能力.

1 改进的 RT-DETR 算法

DETR 算法于 2020 年推出, 是一款基于 Transformer 架构的端到端的目标检测器, 与传统的基于 CNN 的目标检测方法不同的是, 消除了很多手工设计的组, 如锚框生成和非极大值抑制, 大大简化了目标检测的流程. 但是 DETR 也存在许多不足: 相较于其他模型参数量较大, 收敛速度慢. 为解决这些问题, 基于 DETR 的改进不断被提出, Deformable DETR[27]提出了多尺度可变形注意力方法, 加速模型的收敛速度和降低算法的复

杂度, 同时也解决了小目标检测性能不足的问题; DAB-DETR[28]通过引入去噪思想来加快模型的收敛速度; 随后 DINO[29]更加进一步完善这套框架, 使其精度有了更高的提升. 但就“实时性”而言, DETR 系列依旧无法媲美 YOLO 系列. 2023 年, 百度推出了实时性的目标检测器 RT-DETR[30], 能够在保持高精度的同时提供实时性能, RT-DETR 的骨干网络采用 CNN 架构, 如流行的 ResNet 或者百度的 HGNet, 编码器组件采用高效的混合编码器, 通过内部尺度的相互解耦和跨尺度融合解决多尺度特征, 这种独特的结构降低了计算的成本, 解码器采用多层 Transformer 解码器, 可以在推理过程中灵活选择不同的解码器层数, 从而自适应的调整推理速度.

为了使 RT-DETR 更加适合水下任务, 解决海洋底栖息生物目标小而密集导致的探测精度低的问题, 以及进一步减少模型参数量和实时性问题. 本文提出了一种改进的 FERT-DETR 算法, 算法的流程图如图 1 所示. 检测模型的整体架构由一个骨干网络, 一个编码器, 一个解码器组成, 输入图像首先经过改进的 Faster-Net-EMA Block 输出最后 3 层特征图; 之后来到编码器部分, 由于最后一层 S5 特征层的信息较多, AIFI-CGA 会单独处理 S5 特征层, 处理后记为 F5, 之后将 S3、S4 和 F5 交由高水平筛选特征金字塔 HS-FPN 进行跨尺度特征融合; 最后来到解码器部分, 通过查询选择将预测头映射到置信度和边界盒上, 得到最终的检测结果.

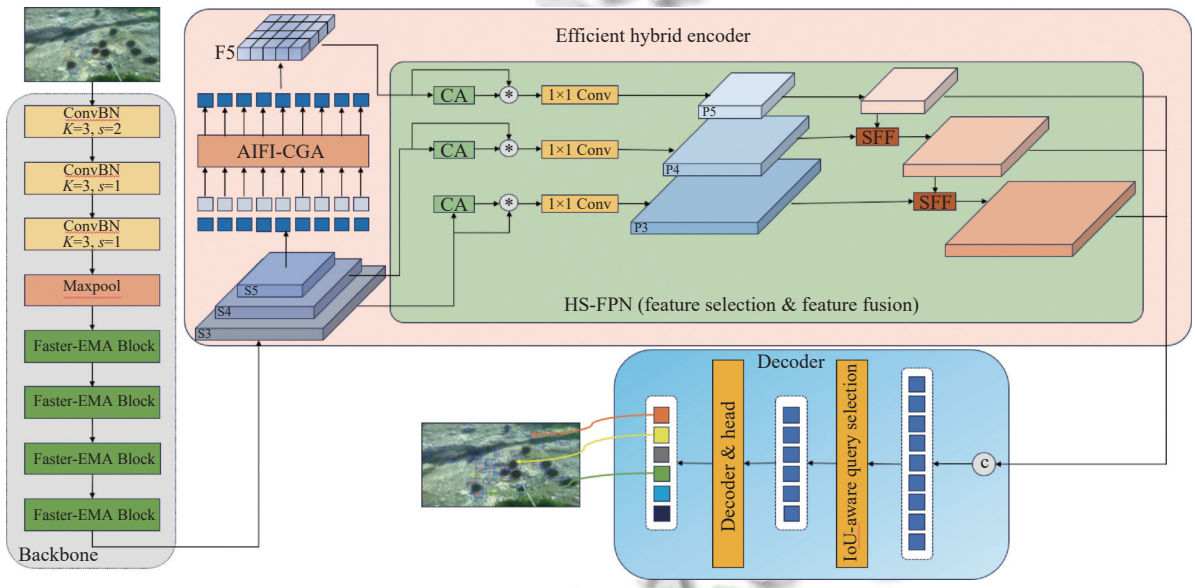


图 1 算法流程图

1.1 FasterNet 与 PConv 模块

为了设计一个更加轻量化快速的神经网络, 许多研究都是在降低 *FLOPs*, 例如 MobileNets[31], ShuffleNet[32], GhostNet[33]等, 但是 *FLOPs* 和延时 (*Latency*) 并不是遵循一一对应的关系, 上述的这些方法虽然减小了 *FLOPs*, 但会造成另一种缺点, 即内存访问 (memory access) 的增加, 对于小模型的训练时间往往是显著的. 所以, 计算一个模型的延时方法如式 (1) 所示.

$$Latency = \frac{FLOPs}{FLOPS} \tag{1}$$

其中, *FLOPs* 是总浮点运算数量, *FLOPS* 是每秒浮点运算数量, 用来衡量有效的运算速度. 因此在 FasterNet 网络当中, 提出了简单有效的部分卷积 PConv, 可以同时减少冗余计算和内存访问. PConv 部分卷积结构如图 2 所示, 其中*表示卷积操作.

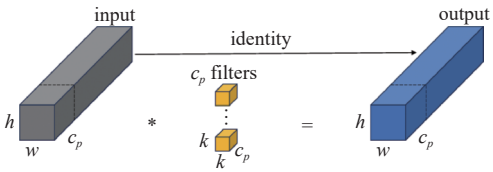


图 2 PConv 部分卷积结构图

假设输入和输出的特征映射具有相同的通道 c, 则 PConv 的 *FLOPs* 以及内存访问情况如式 (2)、式 (3) 所示:

$$h \times w \times k^2 \times c_p^2 \tag{2}$$

$$h \times w \times 2c_p + k^2 \times c_p^2 \approx h \times w \times 2c_p \tag{3}$$

其中, h和w为特征图的宽和高, k是卷积核的大小, c_p是常规卷积的通道数, 在实际的实现过程中, 一般有 $r = c_p/c = 1/4$, 所以 *FLOPs* 仅为常规卷积的 1/16, 内存访问量仅为常规卷积的 1/4. 由此可见, 在主干特征提取网络当中加入 PConv 卷积可以有显著减少计算量和内存访问, 从而使主干网络更加轻量化, 加快模型的训练速度.

1.2 EMA 注意力机制

由于水下生物的游动性以及环境原因, 图片中的水下生物存在重叠、遮挡情况, 导致出现错检漏检的情况. 为解决此类问题, 通过引入注意力模块动态调整模型对于图像中各个区域的权重分配, 增强模型对于目标区域的关注度, 提高检测能力. 近年来, 空间注意力和通道注意力的有效性已经得到了很好的证实. 为保证每个通道的信息并减少计算开销, 在 CA 注意力模块[34]上提出了一种新的高效多尺度注意力 (EMA) 模块[35]. EMA 通过特征分组、并行子网、跨空间学习等策略, 在卷积操作中学习有效的通道描述, 且不降低通道维数, 图 3 右侧展示了实现过程.

对于任意输入特征图$F \in R(c \times h \times w)$被划分为$g$ ($g < c$)个子特征, 用于学习不同的语义信息. EMA 采用 3 条平行路径提取分组特征图的注意力权重描述符. 具体来说, 其中两条路径通过 1×1 分支, 第 3 条路径使用 3×3 分支. 通过对通道方向上的跨通道信息交互进行建模, 实现对所有通道中的依赖关系的捕获, 同时有效地管理计算资源. 并行子结构有助于网络避免更多的顺序处理和大深度. 此外, EMA 还提供了不同空间维度方向的跨空间信息聚合方法, 实现了更丰富的特征融合.

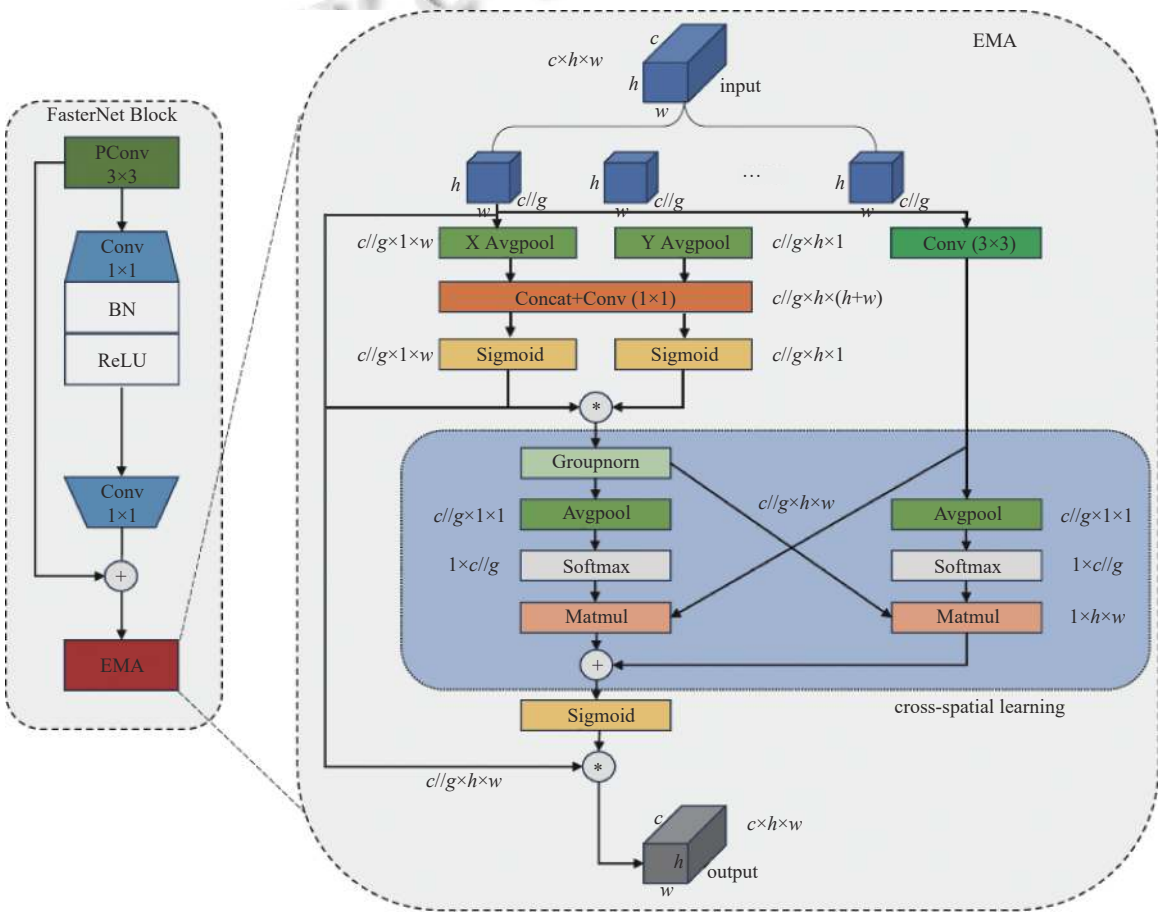


图 3 FasterNet-EMA 模块结构图

EMA 有效地解决了传统的注意力机制的缺点, 它显示了出色的计算效率和泛化能力, 利用 EMA 的灵活性和轻量化特性, 将其集成到了 FasterNet Block 当中, 形成了 FasterNet-EMA 模块, 如图 3 所示.

1.3 AIFI-CGA 模块

随着 Transformer 的发展, 从自然语言处理领域发展到计算机视觉领域, 具有较强的全局建模的能力, 其核心思想是自注意力机制. 但在不同头部之间的注意

力图具有高度的相似性, 导致计算冗余, 为解决该问题, 通过级联群体注意力 (cascaded group attention) 模块[36] 将完整特征切分为不同部分, 并将这些部分输入到不同的注意力头当中, 从而既节约了计算成本, 又提高了注意力的多样性, 最后, 将不同的注意力头的输出特征级联在一起, 得到最终结果. 其结构如图 4 所示.

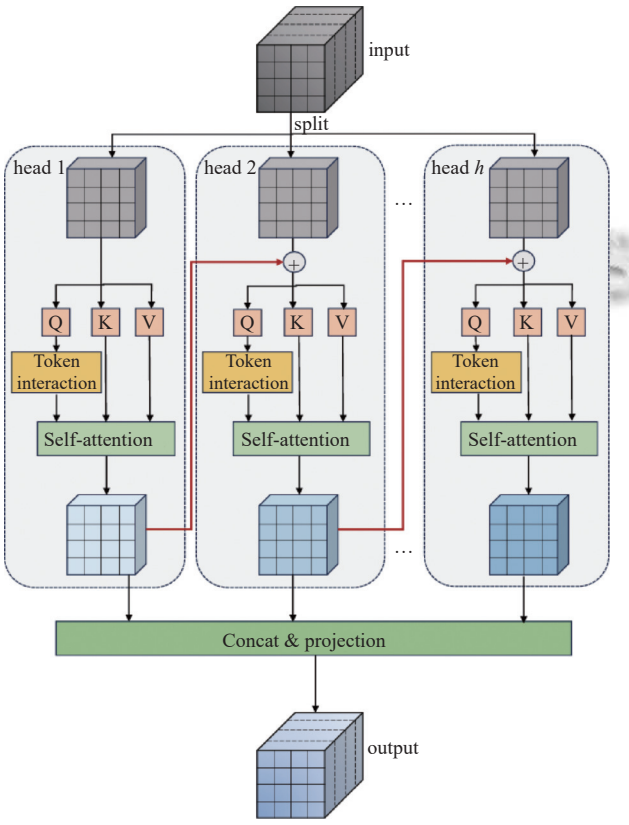


图 4 级联群体注意力模块结构图

通过将级联群体注意力模块加入 RT-DETR 的 AIFI 模块当中, 可以将不同的特征语义信息提供给每个注意力头, 以提高多样性; 还可以增加网络深度, 在不引入额外参数的情况下, 提高模型容量.

1.4 高水平筛选特征金字塔

在水下目标检测数据集当中, 不同生物之间存在大小差异, 并且相同的生物也会因为位置、拍摄角度的不同导致尺寸差异, 因此对于不同尺寸的水下生物的识别会存在一定困难. 为解决水下目标数据集存在的多尺度问题, 设计了基于 RT-DETR 的高水平筛选特征金字塔 (HS-FPN)[26], 来解决多尺度的特征融合问题. 其结构图如图 1 右侧 HS-FPN (feature selection & feature fusion) 所示. HS-FPN 可以分为特征选择 (feature selection) 和特征融合 (feature fusion) 两部分, 首先对于不同的特征尺度进行选择, 之后将高层信息和底层信息进行融合.

特征选择模块主要由通道注意力 (channel attention, CA) 和维度匹配模块两部分组成. CA 注意力根据通道的重要性, 选择性的保留和弱化特征图中的不同通道, 帮助筛选出低级特征和高级特征中的有用信息, 提高模型对目标特征的表达能力. 由于不同尺度的特征图具有不同的通道数, 因此在特征融合之前采用 1×1 卷积将每个尺度的特征图通道数缩小为 256, 以保证不同尺度的特征图能够进行维度匹配. 选择特征融合模块 (select feature fusion, SFF) 通过将高层特征作为权重来筛选底层特征中重要的语义信息, 能够有效地结合高级特征的语义信息和低尺度的细节信息, 实现有目的的特征融合. SFF 的结构图如图 5 所示.

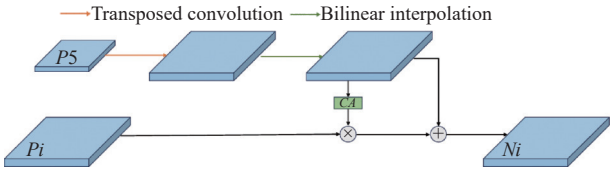


图 5 SFF 模块结构图

对于输入的高层特征 $f_{\text{high}} \in R^{(C\times H\times W)}$ 和低层特征 $f_{\text{low}} \in R^{(C\times H_1\times W_1)}$, 首先将高层特征经过转置卷积 (T-Conv), 随后为了统一高层特征和低层特征的尺寸, 使用双线性插值对高层特征进行下采样或者上采样, 得到 $f_{\text{att}} \in R^{(C\times H_1\times W_1)}$, 之后使用 CA 注意力模块将高层特征作为注意力权重筛选低层特征, 最后将筛选的高层和低层特征进行融合, 得到 $f_{\text{out}} \in R^{(C\times H_1\times W_1)}$. 式 (4)、式 (5) 表示了特征融合的过程.

$$f_{\text{att}} = BL(T-\text{Conv}(f_{\text{high}})) \tag{4}$$

$$f_{\text{out}} = f_{\text{low}} \times CA(f_{\text{att}}) + f_{\text{att}} \tag{5}$$

2 实验及结果及分析

2.1 实验细节

2.1.1 数据集

本文使用的数据集为 2020 年全国水下机器人专业竞赛 (URPC2020) 目标识别组数据集和 DUO (detecting underwater objects) 探测水下物体数据集[37], 两个数据集均由 4 种水下生物组成, 分别是: 海星、扇贝、海胆、海参. URPC2020 数据集共有 7 543 张图片, DUO

数据集共有 7782 张图片. 根据训练集、验证集和测试集按照 7:1:2 的比例进行划分数据集. 部分图片如图 6 所示.

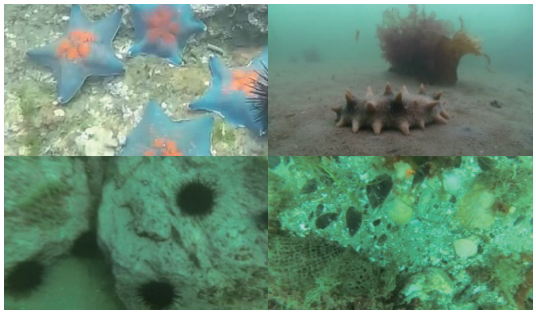


图 6 部分训练集图片

2.1.2 实验环境和参数设置

本实验所使用的硬件是 Intel(R) Core(TM) i5-12400F (CPU), NVIDIA GeForce RTX 4060 (8 GB) (GPU), 16 GB 内存, Windows 操作系统, 软件环境为 PyTorch 1.12.0, CUDA 10.2, 代码运行环境为 Python 3.9. 输入图像大小为 640×640, 运行 100 epoch, 优化器为 AdamW, 学习率为 0.000 1.

2.1.3 模型评价指标

本实验使用精准度 (P)、召回率 (R)、平均精度均值 (mAP)、参数量 (params) 和浮点运算次数 (GFLOPs) 共 5 个指标客观的评价模型性能. 精准率表示预测为正样本的那些数据里预测正确的数据个数, 计算公式如式 (6) 所示. 召回率表示真实为正例的那些数据里预测正确的数据个数, 计算公式如式 (7) 所示. 平均精度均值 (mAP) 是目标检测中最常用的评估指标之一, mAP 是一个综合指标, 他在不同的平均准确率 (AP) 之下计算平均准确率. 计算公式如式 (8)、式 (9) 所示.

$$P=\frac{TP}{TP+FP} \tag{6}$$

$$R=\frac{TP}{TP+FN} \tag{7}$$

$$AP=\int_0^1 P(R)\mathrm{d}R \tag{8}$$

$$mAP=\frac{1}{c}\sum_{i=1}^{c}AP(i) \tag{9}$$

其中, TP表示检测结果中目标正确的个数, FP表示检测结果中目标错误的个数, FN表示正确目标中缺失目标的个数, $AP(i)$表示第i类目标的检测精度, 由P和R计算得出.

2.2 实验结果及其分析

2.2.1 消融实验结果及其分析

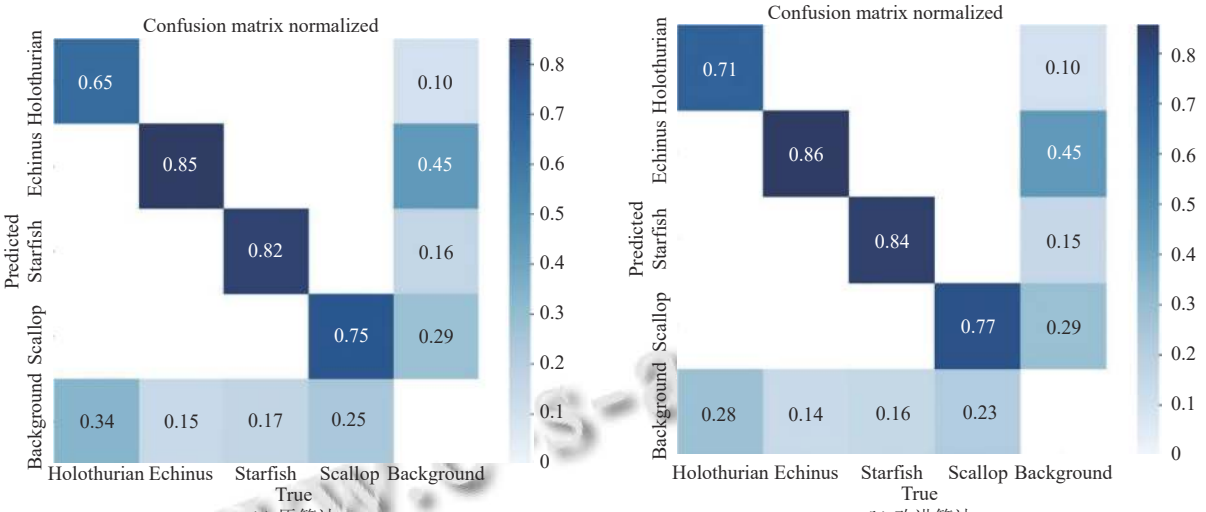
本文采用了多种方法来改进 RT-DETR 算法, 为了说明改进方法的有效性, 在 URPC2020 数据集上进行消融实验, 将 RT-DETR 作为基准模型对比分析改进的结果, 并对实验结果进行分析, 实验结果汇总如表 1 所示.

表 1 消融实验结果

实验编号	Faster-EMA	CGA	高水平筛选特征金字塔	参数量 ($\times10^6$)	计算量 (GFLOPs)	精准率 (%)	召回率 (%)	mAP (%)
1	—	—	—	20.18	58.6	81.0	75.8	82.8
2	√	—	—	17.2	53.2	82.1	77.2	84.4
3	—	√	—	20.01	58.7	82.0	77.3	84.3
4	—	—	√	18.32	54.8	81.7	77.4	84.1
5	√	√	—	17.03	53.3	83.6	78.1	85.6
6	—	√	√	18.15	54.8	83.3	78.2	85.2
7	√	—	√	15.34	49.3	83.5	78.0	85.3
8	√	√	√	15.17	49.4	83.1	78.5	85.9

实验 1 为原始的 RT-DETR 的算法, 实验 2 在原始模型的基础上, 将 R18 的 Block 替换为 Faster-EMA, 模型的参数量以及计算量分别下降 2.98×10^6 和 5.4, mAP 提升了 1.6 个百分点, 表明 Faster-EMA 能够使模型更加轻量化, 且对于图像特征提取效率更高. 实验 3 在编码器中引入 CGA 模块, 在参数量和计算量上无明显变化, mAP 提升了 1.5 个百分点, 表明改进的网络能够更准确地关注图像的重要信息, 过滤掉背景信息, 提升模型的表现能力. 实验 4 将 CCFM 替换为 HS-FPN, 参数量和计算量分别下降 1.86×10^6 和 3.8, mAP 提升了 1.3 个百分点, 表明使用多尺度特征融合, 将每个尺度上的信息进行选择融合, 更好的捕捉目标在不同尺度上的表现, 提升模型的性能. 实验 5、6、7 为同时引入两个改进模块, 组合使用对于检测性能的提升要高于使用单个改进模块, 表明各个模块之间的改进不会产生冲突. 实验 8 为本文的改进算法, 参数量和计算量分别下降 5.01×10^6 和 9.2, mAP 提升了 3.1 个百分点, 显著降低了参数量以及计算量. 综上所述, 本文所提方法能够在基准模型之上得到有效的改进, 在提高了平均检测精度的同时, 降低了参数量和计算量.

在实验过程中, 使用图 7 混淆矩阵更加清晰地看出原模型与改进后的模型对水下 4 种目标的检测效果, 其中对角线代表检测正确的比例, 横坐标代表真实标签, 纵坐标代表预测标签, 可以看出改进后的模型在各类别检测正确的比例都有提升, 检测为背景的比例也有所减少, 表明改进模型的检测能力有所提高.



(a) 原算法 (b) 改进算法

图 7 原算法与改进算法混淆矩阵

改进前后的模型在训练过程中的损失曲线如图 8 所示. 其中虚线为原始的 RT-DETR 算法, 实线为改进后的 FERT-DETR 算法, 可以看出改进后的算法损失值下降得更快, 且收敛效果更好.

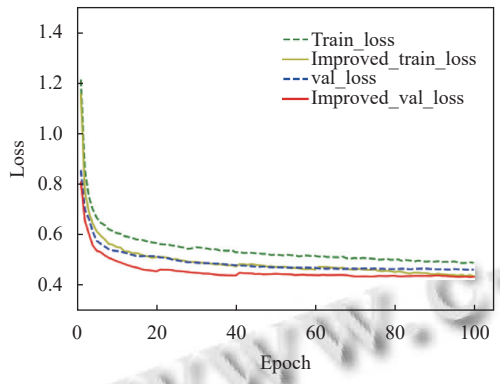


图 8 损失曲线对比图

2.2.2 不同模型实验结果对比

将本文所提的改进 FERT-DETR 算法同其他主流目标检测算法进行对比, 探究算法的优越性, 每个模型在 URPC2020 数据集和 DUO 数据集上进行实验, 实验结果如表 2 所示.

2.2.3 不同模型实验结果对比

为了更好地比较本文所使用的方法与原模型在水下目标检测方面的效果, 图 9 展示了两者的热力图对比结果. 可以看出, RT-DETR 模型存在较多的错检和漏检情况, 而本文的方法相对较少. 通过热力图可以看出, 在加入注意力机制后的改进算法能够对目标更加敏感. 对于环境复杂, 图片模糊的情况下, 引入多尺度融合的特征网络能够更好地检测出模糊、遮挡的目标. 因此本文可以更好地应用于水下目标检测.

表 2 横向对比实验结果

模型	参数量 ($\times 10^6$)	计算量 (GFLOPs)	URPC2020			DUO		
			精准率 (%)	召回率 (%)	*mAP* (%)	精准率 (%)	召回率 (%)	*mAP* (%)
Faster R-CNN	42.23	180.3	78.2	70.2	79.6	76.6	69.2	77.3
YOLOv5	25.09	64.2	82.5	77.1	84.9	81.1	74.3	82.7
YOLOv6	52.00	161.3	79.9	72.5	80.2	78.4	70.9	78.8
YOLOv7	36.71	106.7	81.7	77.4	84.1	80.1	75.2	81.6
YOLOv8	25.88	78.9	82.6	78.1	85.2	79.6	75.9	82.9
DETR	41.00	86.2	75.7	68.2	76.8	76.7	69.5	75.8
Deformable DETR	34.01	78.4	80.9	73.9	81.5	81.9	72.8	79.4
RT-DETR	20.18	58.6	81.0	75.8	82.8	84.0	74.2	81.9
Ours	15.17	49.4	83.1	78.5	85.9	85.4	75.0	83.6

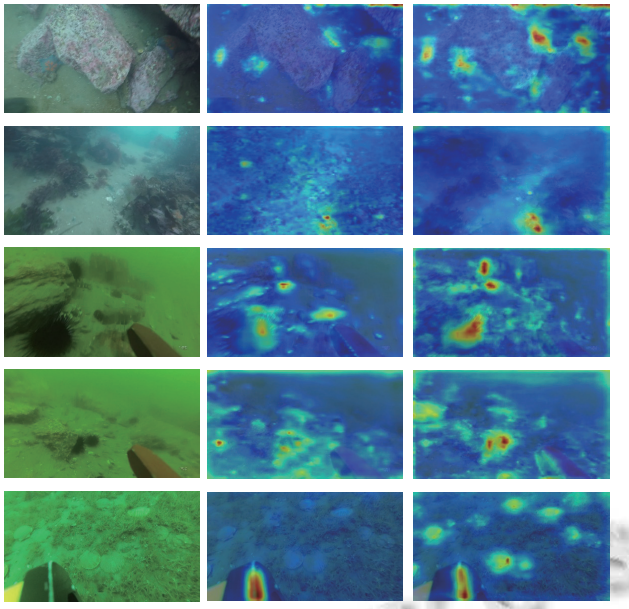


图 9 热力图对比

2.2.4 不同模型实验结果对比

图 10 展示了 FERT-DETR 模型对比 RT-DETR 原模型的检测效果, 从图中可以看到, 原模型对于水下目标检测由于图像模糊、遮挡、小目标等因素存在较多的错检、漏检和重复检测的情况. 而改进的 FERT-DETR 也存在部分错检漏检情况, 但明显优于原模型的检测效果. 总体来看, 本文的改进能够显著解决水下图像的模糊遮挡等问题, 增强检测精度, 更加适合于水下目标检测.

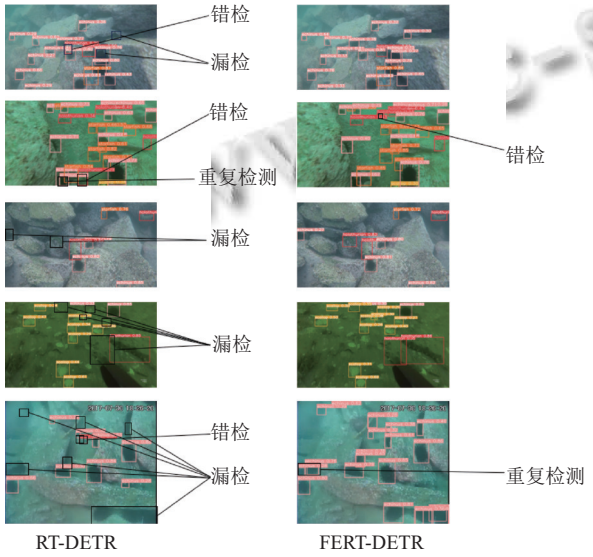


图 10 总体检测实验图对比

3 结论与展望

本文针对水下目标模糊和遮挡等问题, 在 RT-DETR 网络模型的基础上提出了 FERT-DETR. 该方法通过更换骨干网络, 使用更加轻量化的 FasterNet Block, 又在其骨干网络的基础上加入 EMA 注意力机制, 使模型对目标更加敏感, 增强特征提取能力. 之后提出 AIFI-CGA, 既节约了计算成本, 又提高了注意力的多样性以及模型容量. 最后通过优化特征融合网络, 使用高水平筛选特征金字塔 HS-FPN 替换 CCFM, 能够有效地提升模型对遮挡目标的检测性能. 通过一系列的实验, 证明本文所提算法的有效性. 本文方法仍有改进空间, 基于 Transformer 的目标检测最大问题来自计算量, 所以未来工作可以在模型规模和检测速度上进行优化.

参考文献

1 Khan A, Fouda MM, Do DT, *et al*. Underwater target detection using deep learning: Methodologies, challenges, applications and future evolution. IEEE Access, 2024, 12: 12618–12635 [doi: 10.1109/ACCESS.2024.3353688]

2 Sahoo A, Dwivedy SK, Robi PS. Advancements in the field of autonomous underwater vehicle. Ocean Engineering, 2019, 181: 145–160. [doi: 10.1016/j.oceaneng.2019.04.011]

3 Ghafoor H, Noh Y. An overview of next-generation underwater target detection and tracking: An integrated underwater architecture. IEEE Access, 2019, 7: 98841–98853. [doi: 10.1109/ACCESS.2019.2929932]

4 Liu K, Liang YQ. Enhancement of underwater optical images based on background light estimation and improved adaptive transmission fusion. Optics Express, 2021, 29(18): 28307–28328. [doi: 10.1364/oe.428626]

5 Shi J, Zhuo X, Zhang C, *et al*. Research on key technologies of underwater target detection. Proceedings of the 7th Symposium on Novel Photoelectronic Detection Technology and Applications. Kunming: SPIE, 2021. 1128–1137. [doi: 10.1117/12.2586895]

6 Fu HX, Song GQ, Wang YC. Improved YOLOv4 marine target detection combined with CBAM. Symmetry, 2021, 13(4): 623. [doi: 10.3390/sym13040623]

7 Felzenszwalb PF, Girshick RB, McAllester D, *et al*. Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(9): 1627–1645. [doi: 10.1109/TPAMI.2009.167]

8 Lowe DG. Distinctive image features from scale-invariant

keypoints. International Journal of Computer Vision, 2004, 60(2): 91–110. [doi: 10.1023/B:VISI.0000029664.99615.94]
9 Dalal N, Triggs B. Histograms of oriented gradients for human detection. Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego: IEEE, 2005. 886–893. [doi: 10.1109/CVPR.2005.177]
10 Hearst MA, Dumais ST, Osuna E, *et al*. Support vector machines. IEEE Intelligent Systems and Their Applications, 1998, 13(4): 18–28. [doi: 10.1109/5254.708428]
11 Dai JF, Li Y, He KM, *et al*. R-FCN: Object detection via region-based fully convolutional networks. Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona: Curran Associates Inc., 2016. 379–387.
12 Girshick R, Donahue J, Darrell T, *et al*. Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014. 580–587. [doi: 10.1109/CVPR.2014.81]
13 Girshick R. Fast R-CNN. Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago: IEEE, 2015. 1440–1448. [doi: 10.1109/ICCV.2015.169]
14 Ren SQ, He KM, Girshick R, *et al*. Faster R-CNN: Towards real-time object detection with region proposal networks. Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal: MIT Press, 2015. 91–99.
15 He KM, Gkioxari G, Dollár P, *et al*. Mask R-CNN. Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017. 2980–2988. [doi: 10.1109/ICCV.2017.322]
16 Cai ZW, Vasconcelos N. Cascade R-CNN: Delving into high quality object detection. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 6154–6162. [doi: 10.1109/CVPR.2018.00644]
17 Li X, Shang M, Qin HW, *et al*. Fast accurate fish detection and recognition of underwater images with Fast R-CNN. Proceedings of the OCEANS 2015-MTS/IEEE Washington. Washington: IEEE, 2015. 1–5. [doi: 10.23919/OCEANS.2015.7404464]
18 Zeng LC, Sun B, Zhu DQ. Underwater target detection based on Faster R-CNN and adversarial occlusion network. Engineering Applications of Artificial Intelligence, 2021, 100: 104190. [doi: 10.1016/j.engappai.2021.104190]
19 Liu W, Anguelov D, Erhan D, *et al*. SSD: Single shot multibox detector. Proceedings of the 14th European Conference on Computer Vision. Amsterdam: Springer, 2016. 21–37. [doi: 10.1007/978-3-319-46448-0_2]
20 Redmon J, Divvala S, Girshick R, *et al*. You only look once: Unified, real-time object detection. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 779–788. [doi: 10.1109/CVPR.2016.91]
21 强伟, 贺昱曜, 郭玉锦, 等. 基于改进 SSD 的水下目标检测算法研究. 西北工业大学学报, 2020, 38(4): 747–754. [doi: 10.3969/j.issn.1000-2758.2020.04.008]
22 Li Y, Guo JH, Guo XM, *et al*. Toward in situ zooplankton detection with a densely connected YOLOv3 model. Applied Ocean Research, 2021, 114: 102783. [doi: 10.1016/j.apor.2021.102783]
23 黄廷辉, 高新宇, 黄春德, 等. 基于 FAttention-YOLOv5 的水下目标检测算法研究. 微电子学与计算机, 2022, 39(6): 60–68. [doi: 10.19304/J.ISSN1000-7180.2021.1261]
24 Wang Z, Chen HJ, Qin HD, *et al*. Self-supervised pre-training joint framework: Assisting lightweight detection network for underwater object detection. Journal of Marine Science and Engineering, 2023, 11(3): 604. [doi: 10.3390/jmse11030604]
25 Carion N, Massa F, Synnaeve G, *et al*. End-to-end object detection with Transformers. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 213–229. [doi: 10.1007/978-3-030-58452-8_13]
26 Chen YF, Zhang CY, Chen B, *et al*. Accurate leukocyte detection based on deformable-DETR and multi-level feature fusion for aiding diagnosis of blood diseases. Computers in Biology and Medicine, 2024, 170: 107917. [doi: 10.1016/j.compbiomed.2024.107917]
27 Zhu XZ, Su WJ, Lu LW, *et al*. Deformable DETR: Deformable Transformers for end-to-end object detection. arXiv:2010.04159, 2021.
28 Liu SL, Li F, Zhang H, *et al*. DAB-DETR: Dynamic anchor boxes are better queries for DETR. Proceedings of the 10th International Conference on Learning Representations. OpenReview.net, 2022.
29 Zhang H, Li F, Liu SL, *et al*. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. arXiv:2203.03605, 2022.
30 Zhao YA, Lv WY, Xu SL, *et al*. DETRs beat YOLOs on real-time object detection. arXiv:2304.08069, 2024.

31 Howard AG, Zhu ML, Chen B, *et al*. MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861, 2017.

32 Zhang XY, Zhou XY, Lin MX, *et al*. ShuffleNet: An extremely efficient convolutional neural network for mobile devices. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 6848–6856. [doi: 10.1109/CVPR.2018.00716]

33 Han K, Wang YH, Tian Q, *et al*. GhostNet: More features from cheap operations. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 1577–1586. [doi: 10.1109/CVPR42600.2020.00165]

34 Hou QB, Zhou DQ, Feng JS. Coordinate attention for efficient mobile network design. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 13708–13717. [doi: 10.1109/CVPR46437.2021.01350]

35 Ouyang DL, He S, Zhang GZ, *et al*. Efficient multi-scale attention module with cross-spatial learning. Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing. Rhodes Island: IEEE, 2023. 1–5. [doi: 10.1109/ICASSP49357.2023.10096516]

36 Liu XY, Peng HW, Zheng NX, *et al*. EfficientViT: Memory efficient vision Transformer with cascaded group attention. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 14420–14430. [doi: 10.1109/CVPR52729.2023.01386]

37 Liu CW, Li HJ, Wang SC, *et al*. A dataset and benchmark of underwater object detection for robot picking. Proceedings of the 2021 IEEE International Conference on Multimedia & Expo Workshops. Shenzhen: IEEE, 2021. 1–6. [doi: 10.1109/ICMEW53276.2021.9455997]

(校对责编: 张重毅)