张路, 魏本昌, 魏鸿奥, 等. 面向水下目标检测的改进 DETR 算法[J]. 智能计算机与应用,2025,15(10):46-53. DOI:10.20169/j.issn.2095-2163.251007

面向水下目标检测的改进 DETR 算法

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, 湖北 十堰 442002)

摘 要: 水下的目标检测是当前目标检测的一大热点,针对水下的低能见度和低光照条件带来的挑战,提出了一种改进的 DETR 算法。首先,使用分散注意力模块的 ResNeSt 骨干网络,提升了对水下数据特征的提取性能,剔除了冗余的背景信息;其次,引入了多尺度可变形注意力编码器,加强了对特征信息的聚合能力,使模型收敛速度更快,同时提高了对小目标的检测性能;然后,使用了 Smooth-L1 和 CIoU 结合的损失函数,使模型能够更快地收敛,并且达到更高的精度;最后,在扩充的 DUO 水下目标检测数据集上进行实验。结果表明所提的改进算法优于其他先进的目标检测算法,所提出的算法可以有效地应用于水下目标检测任务。

关键词: 计算机视觉; UOD; ResNeSt; DETR
中图分类号: TP391 **文献标志码**: A **文章编号**: 2095-2163(2025)10-0046-08

Improved DETR algorithm for underwater target detection

ZHANG Lu, WEI Benchang, WEI Hong'ao, ZHOU Longgang
(School of Electrical & Information Engineering, Hubei University of Automotive Technology, Shiyan 442002, Hubei, China)

Abstract: Underwater object detection is currently a hot topic in object detection. In response to the challenges brought by low visibility and low lighting conditions during underwater detection, an improved DETR algorithm is proposed. The ResNeSt backbone network using a distractor module has improved the performance of extracting underwater data features and eliminated redundant background information. The introduction of a multi-scale deformable attention encoder enhances the aggregation ability of feature information, resulting in faster model convergence speed and improved detection performance for small targets. The loss function combining Smooth-L1 and CIoU are used to enable the model to converge faster and achieve higher accuracy. Finally, experiments are conducted on the expanded DUO underwater object detection dataset, and it is demonstrated that the proposed improved algorithm outperforms other advanced object detection algorithms. The proposed algorithm can be effectively applied to underwater object detection tasks.

Key words: computer vision; UOD; ResNeSt; DETR

0 引 言

随着人工智能的发展,越来越多的场景课题和任务可以通过人工智能加以解决,随之而来的就是对人工智能算法和硬件设备的需求增长明显。而计算机视觉作为人工智能领域的一个重要分支,已广泛应用在人们的日常生活中。目标检测是计算机视觉的一项基本任务,其目的是对目标进行分类和定位,但是对于较为复杂的水下场景,目前的目标检测并没有取得令人满意的效果,这也在一定程度上阻碍了水下机器人的发展[1]。

海洋是地球上最重要的自然资源之一,其水下环境是一个庞大且复杂的生态系统,具有独特的物理和生物特性。而在水下目标检测的过程中,存在很多的困难和挑战,具体体现在以下 4 点:

首先,水下的场景普遍存在可见度低,即模糊;

其次,随着水下深度的增加,光照降低,能见度显著降低;

然后,很多时候水下场景呈现白色、绿色、蓝色,颜色较为相近且不易辨认;

基金项目: 湖北省教育厅项目(B2019077); 湖北汽车工业学院博士科研启动基金(BK201603)。
作者简介: 张 路(1999—),男,硕士研究生,主要研究方向:计算机视觉,深度学习。
通信作者: 魏本昌(1975—),男,博士,讲师,硕士生导师,主要研究方向:图像检索。Email:bc_david@163.com。
收稿日期: 2024-01-10

最后，水生杂质和水生植物会对水下生物造成遮挡。

这些问题导致其在色彩和纹理信息方面的表现不如大气光学图像[2]。

早期的目标检测依赖于人工提取图像特征，然而面对各种各样的场景和复杂的环境，传统的人工特征提取已经无法满足日益增长的需求。目前的目标检测主要分为两大类，分别是两阶段检测算法和单阶段检测算法。

两阶段框架的过程与传统方案有些相似，首先生成候选区域，然后将其分类到不同的目标类别中。这种方法有着很高的准确性，但是检测速度较慢，对于实时性要求高的任务不太适用。基于两阶段的代表性模型有 R-CNN 系列[3-5]、空间金字塔池化网络(SPP-Net)[6]、R-FCN[6]、特征金字塔网络[7]。Li 等学者[8]将 Fast RCNN 应用于水下图像的识别与检测，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，此后又将 Faster RCNN[9]应用于水下图像识别与检测，再次提升平均精度。Zeng 等学者[10]通过将对抗性遮挡网络(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。

单阶段框架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，突显出了强大的实时处理的能力。强伟等学者[11]将 ResNet 作为基础网络的 SSD 检测模型，用于水下目标检测。Li 等学者[12]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。Li 等学者[13]将 YOLOv4 应用于水下目标检测，采用 k-means 算法重新聚类得到水下目标的边框。Lei 等学者[14]将 YOLOv5 用于水下环境，并结合水下环境的特点对其进行改进。黄廷辉等学者[15]提出一种基于 F-CBAM 注意力机制的 YOLOv5 水下目标检测网络模型 FAttention-YOLOv5 模型。Wang 等学者[16]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。但传统的深度学习卷积算法受制于卷积核感受野的大小和锚框的形状，缺乏对全局的学习能力，以至于对目标的检测精度并不高。

随着深度学习的不断发展，基于 Transformer 的目标检测网络引起了广泛的关注，引发了研究热潮，这种网络完全依赖于注意力机制来捕捉输入和输出之间的全局依赖关系，通过点积操作自适应调整权重参数，减少模型的学习偏差，因此 Transformer 拥有比 CNN 更加强大的全局建模能力和泛化能力。近年来，国内外学者也基于 Transformer 陆续提出了不同的目标检测的模型，如 CBNET[17]，DyHead[18]和 Swin-Transformer[19]等。Carion 等学者[20]提出了一个新的基于 Transformer 的目标检测模型、即 DETR 模型。不同于 Two-Stage 和 One-Stage 检测算法，DETR 模型将目标检测视为一个直接集合预测的问题，是一个端到端的检测模型。首先利用 CNN 骨干模型进行特征提取，再利用 Transformer 进行并行预测，DETR 模型的出现为解决目标检测任务提供了一种全新的思路。

1　改进的 DETR 检测算法

DETR 模型的出现打破了人们以往对目标检测的固定思维，通过采用端到端的方法，将 CNN 和 Transformer 相结合，DETR 由 CNN 提取特征，Transformer 编码器-解码器进行预测，将目标检测视为一个直接集合预测的问题，有效地消除了许多手工设计的组件的需要，如非极大值抑制和锚框生成。由于 Transformer 模型具备对全局特征进行关注和学习的能力，DETR 模型同样也具备强大的全局学习能力。本文提出的改进的 DETR 检测算法包含 3 个方面，分别是采用多尺度可变形注意力编码器、进行特征提取的 ResNeSt 骨干网络、以及将 Smooth-L1 和 CIoU 相结合作为损失函数。

1.1　Deformable DETR 模型[21]

DETR 采用注意力检测模块进行检测和结果输出。其中，注意力编码器(Encoder)作为注意力检测模块的编码部分，将输入的特征信息编码为一系列的特征向量序列，结合了可学习的位置编码，能够帮助模型实现更好的泛化以及捕捉位置信息，编码器通过一系列的 Transformer 编码层，其中包括自注意力机制和前馈神经网络，用来对输入的特征向量序列进行处理和转换。随后，与注意力解码器(Decoder)结合，接受来自编码器的特征向量序列，使得模型具备了全局建模的能力。通过堆叠编码器-解码器，可以将目标检测问题转换为集合预测的问题。然而，由于注意力模块初始化比较稀疏，需要长时间学习以实现收敛，不仅计算量大，且没有使用多尺度特征，导致对小目标的检测效果较差。由于本文是水下目标检测，小目标较多，为解决这一难题，本文采用了多尺度可变形注意力对 DETR 的注意力检测模块进行改进。

1.1.1 多尺度特征

由于 DETR 提取特征时并没有利用到多尺度的特征，导致其对小目标的检测效果较差。于是引进了可变形注意力。该模块可以很方便地处理多尺度特征(Multi-Scale Features)，使模型在不使用特征金字塔(Feature Pyramid Networks，FPN)[22]结构的情况下也能够有效地利用骨干网络提取到的多尺度特征图，进而降低 FPN 的计算量，以及减少小目标语义信息在下采样中的丢失，提升对小目标的检测性能。Backbone 生成多尺度特征如图 1 所示。

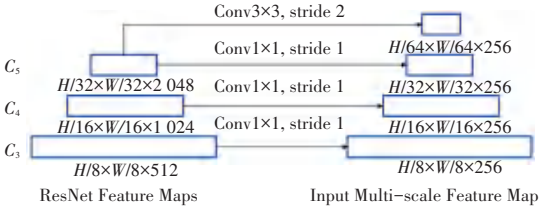

图 1 多尺度特征图

Fig. 1 Multi-scale feature map

1.1.2 多尺度可变形注意力

不同于普通的注意力编码器，可变形注意力编码器的查询向量(query)并不是和全局每个位置上的键向量(key)都计算注意力权重，而是对于每个查询向量(query)，仅在全局位置中采样部分位置的键向量(key)，并且值向量(value)也是基于这些位置插值得到，使模型在保留性能的同时降低计算量。因此本文采用的是多尺度可变形注意力(Multi-Scale Deformable Attention)检测模块，使模型可以有效地利用多尺度信息，以及减少计算量。Transformer 中的多头注意力机制的数学公式具体如下：

$$\mathrm{MultiHeadAttn}(z_q,x)=\sum_{m=0}^{M}W_m\Big[\sum_{k\in\Omega_k}A_{mqk}\cdot W'_m x_k\Big] \quad (1)$$

其中，z_q 表示查询向量(query)，由 x 经过线性变换生成；q 表示对应的索引；k 表示键向量(key)的索引；Ω_k 表示所有的 k 集合；m 表示是第几个注意力头部；W_m 表示对注意力施加在值向量(value)后的结果进行线性变换从而得到不同头部的输出结果；A_{mqk} 表示归一化的注意力权重。由此可知，在 Transformer 的多头注意力计算中，每个查询向量(query)都要与所有位置的键向量(key)计算注意力权重，并且对应施加在所有的值向量(value)上。可变形注意力机制的数学公式具体如下：

$$\mathrm{DeformAttn}(z_q,p_q,x)=\sum_{m=0}^{M}W_m\Big[\sum_{k=1}^{K}A_{mqk}\cdot W'_m x(p_q+\Delta p_{mqk})\Big] \quad (2)$$

这里多了 p_q 和 Δp_{mqk}。其中，前者表示 z_q 的位置(理解成坐标即可)，是 2D 向量，研究中称为参考点(reference points)；后者是采样集合点相对于参考点的位置偏移(offsets)。如可变形卷积一样，位置偏移 Δp_{mqk} 是可学习的，由查询向量(query)经过全连接层得到。并且，注意力权重也一样，直接由查询向量(query)经过全连接层得到。多尺度可变形注意力机制的数学公式具体如下：

$$\mathrm{MSDeformAttn}(z_q,\hat{p}_q,\{x^l\}_{l=1}^{L})=\sum_{m=0}^{M}W_m\Big[\sum_{l=1}^{L}\sum_{k=1}^{K}A_{mlqk}\cdot W'_m x^l(\phi_l(p_q)+\Delta p_{mlqk})\Big] \quad (3)$$

其中，L 表示共有 L 层特征；l 表示维度编号；$\hat{p}_q$ 表示归一化的参考点坐标；ϕ_l 表示将归一化后的特征坐标映射到第 l 层特征上去；x 表示特征图编号。式(3)中加入了多尺度特征。每个参考点在所有特征层都会有一个对应的坐标，从而方便计算在不同特征层进行采样的点的位置。算法的整体流程如图 2 所示。

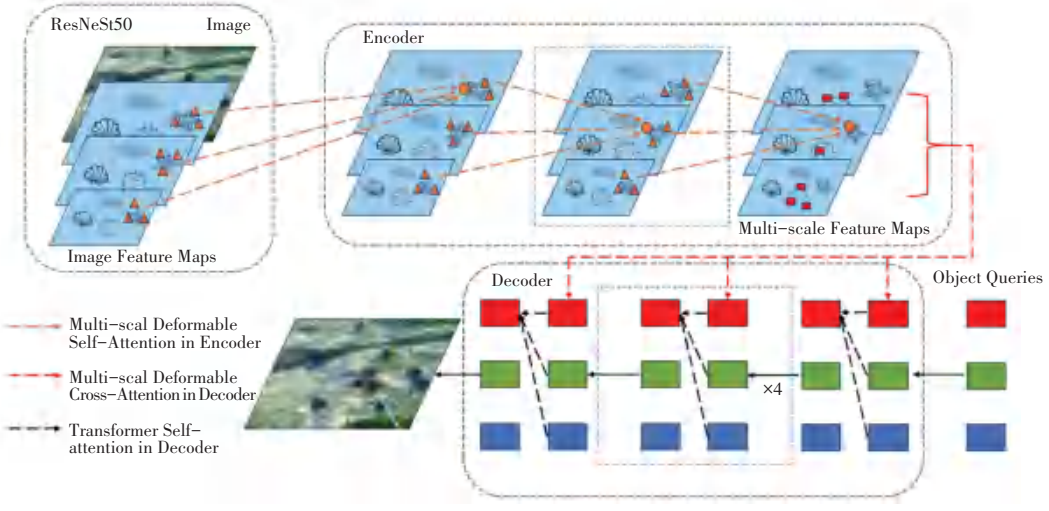

图 2 算法流程图

Fig. 2 Flowchart of the algorithm

1.2　改进的骨干网络

ResNet 网络模型[23]自 2015 年提出之后，许多学者将该网络模型当作基础模型架构应用于后续研究中。通过引入跳跃连接（shotcut connection 或者 skip connection）的概念，这种连接方式允许网络直接跳过一层或多层，将输入直接传递到输出，从而解决了网络深度增加时产生的梯度消失和退化问题。

传统的卷积神经网络通过堆叠许多卷积层和非线性激活函数来提取图像的特征，但由于网络层数的不断增加，网络面临的梯度问题也随之而来。ResNet 通过跳跃链接，即将某一层的输出直接与之后的某一层的输入相加，能够传递更多的信息和梯度，从而保持较浅层的特征信息能够传递到高层。

具体而言，ResNet 的残差函数通过如下公式来计算：

$$F = \boldsymbol{W}_2\sigma(\boldsymbol{W}_1 X) \tag{4}$$

其中，σ 表示 ReLU 激活函数；$\boldsymbol{W}_1$ 和 $\boldsymbol{W}_2$ 分别表示线性变换矩阵；X 表示输入图像。在此基础上、又推得：

$$Y = F(X, \{X_i\}) + X \tag{5}$$

将差异和原始输入相加，获得输出 Y。因此，很多后续的网络模型都借鉴了 ResNet 的思想，并将其作为骨干模型来构建更复杂的网络结构。

ResNeSt 网络模型的灵感来源于 ResNeXt[24]和 SE-Net[25]，SK-Net[26]，与 ResNet 相比，准确率有所提高，但参数量没有显著增加，由于 ResNet 等网络有限的感受野大小以及缺乏跨通道之间的相互作用，这些网络不能对特定图像进行有效分类。ResNeSt 中引入了分散注意力（Split-Attention）模块，这一模块使 Attention 能在特征图组间获取不同权重的特征，其中 ResNeSt 中 ResNeSt block 的结构如图 3 左侧所示。

其具体思想如下：将图像输入到网络当中，通过分支结构划分为由超参数 k 确定的组数（Cardinal group），并引入一个新的基数参数 r，从而将整个图像划分为 $G = k \times r$ 个子组，即将整个网络模型分组进行拆分处理。然后，对每一个子群分别进行 1×1 和 3×3 的卷积，汇总并输入到分散注意力（Split Attentions）当中。接下来，通过分散注意力机制对不同的特征赋予不同的权重，使每个通道的特征对最终的预测结果产生不同的影响。最后，将聚合输出的特征图与残差模块输出的特征图进行线性组合，得到最终的输出结果。其中，分散注意力（Split Attentions）的结构图如图 3 右侧所示。

图 3 的分散注意力模块中，首先聚合特征进行全局平均池化处理，得到通道的全局上下文信息，然后通过全连接层对各子组进行分类，并利用 Softmax 函数自适应计算各子组的权值。模块中 r-Softmax 函数公式见如下：

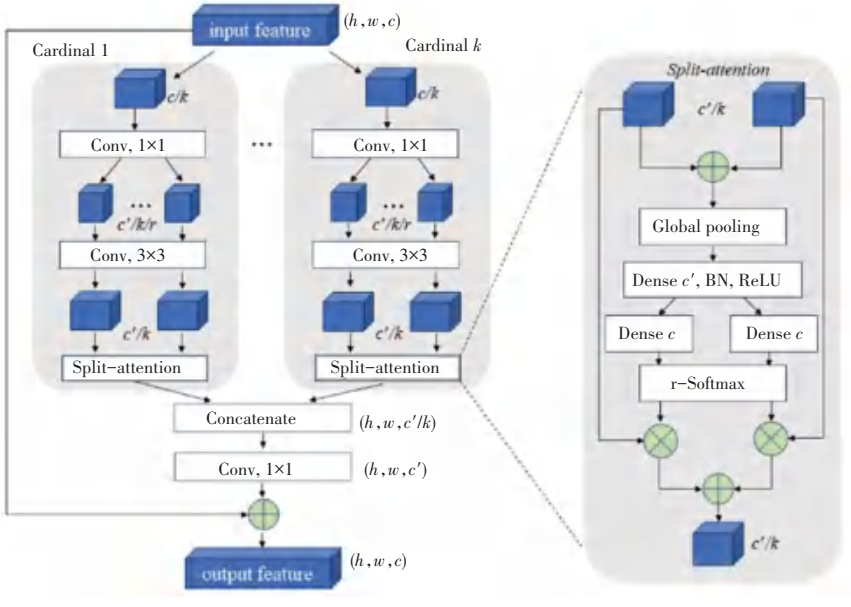

图 3　ResNeSt 骨干网络

Fig. 3　Backbone network of ResNeSt

$$a_i^k(c)=\begin{cases}\dfrac{\exp(\zeta_i^c(s^k))}{\sum_{j=1}^{R}\exp(\zeta_j^c(s^k))}, & \text{if } R>1\\ \dfrac{1}{1+\exp(-\zeta_i^c(s^k))}, & \text{if } R=1\end{cases} \tag{6}$$

其中，$a_i^k(c)$ 表示第 i 个子群的特征权值，ζ_i^c 表示由密集连接层、全连接层和激活函数组成的注意力权重。

1.3 损失函数的优化

本文是针对水下生物的检测与定位。为了优化模型，使损失函数更快地收敛以达到最优，本文提出了将 Smooth-L1 损失函数和 CIoU 损失函数相结合作为回归损失对检测边框进行预测回归，Smooth-L1 损失函数可以更快地收敛，CIoU 损失函数可以更好地优化目标检测的精度，两者结合能够在训练过程中更好地平衡目标检测的精度和鲁棒性。

1.3.1 Smooth-L1 损失函数

Smooth-L1[27]结合了 *L*1 损失函数和 *L*2 损失函数的优点，其定义公式为：

$$L_{\text{Smooth-L1}}(b_{\sigma(i)},\hat{b}_i)=\begin{cases}\dfrac{1}{2}\sum_{i=0}^{N}(b_{\sigma(i)}-\hat{b}_i)^2, & \text{if } |b_{\sigma(i)}-\hat{b}_i|<1\\ \sum_{i=0}^{N}|b_{\sigma(i)}-\hat{b}_i|-0.5, & \text{other}\end{cases} \tag{7}$$

由式(7)可知 Smooth-L1 Loss 是一个分段函数，综合了 *L*1 Loss 和 *L*2 Loss 两个损失函数的优点，即在 x 较小时采用平滑的 *L*2 Loss，使得模型在训练的过程中更加容易收敛，加快模型的训练速度，提高训练效率。在 x 较大时采用稳定的 *L*1 Loss，这有助于减小梯度爆炸的风险，提高了模型的训练稳定性。

1.3.2 CIoU 损失函数

DETR 模型中使用的是 GIoU 损失函数[29]，该损失函数考虑到了预测框和真实框之间的覆盖程度、包围盒的大小以及相应的偏移量，对部分重叠的目标框有较好的鲁棒性。GIoU 定义公式如下：

$$\text{GIoU}=\text{IoU}-\frac{C-(A\cup B)}{C} \tag{8}$$

其中，A 表示预测框；B 表示真实框；C 表示能同时包住 A 与 B 的集合。但是当 AB 两个框完全重叠的情况下，不能反映出实际情况，GIoU 会退化成 IoU，A 和 B 的相对位置关系会无法区分，且由于 GIoU 仍然严重依赖 IoU，因此在 2 个垂直方向，误差很大，很难收敛。因此针对上述问题提出了 DIoU 损失函数[29]。DIoU 定义公式如下：

$$\text{DIoU}=\text{IoU}-\frac{\rho^2(b,b^{\text{gt}})}{c^2} \tag{9}$$

其中，b，b^{gt} 分别表示预测框和真实框的中心点；c 表示能够同时包含预测框和真实框的最小外接矩形的对角线距离；ρ 表示计算 2 个中心点间的欧式距离。

仍需一提的是，DIoU 损失能够将预测框和真实框之间的距离、重叠率以及尺度都考虑进去。DIoU 的惩罚项是基于 2 个边框之间的欧氏距离，能够有效地避免 GIoU 在 2 个边框距离较远时产生较大的外包框，所以 DIoU 的收敛速度更快、更稳定。虽然 DIoU 能够对最小化预测框和真实框的中心点距离实现加速收敛，但是预测框和目标框之间的长宽比的一致性也是非常重要的。针对以上问题，提出了 CIoU 损失函数，CIoU 定义公式如下：

$$\text{CIoU}=\text{IoU}-\left(\frac{\rho^2(b,b^{\text{gt}})}{c^2}+\alpha\nu\right) \tag{10}$$

其中，α 表示权重函数，ν 表示度量长宽比的相似性。CIoU 损失函数同时考虑到了重叠面积、中心点长度和长宽比带来的问题。

通过将 Smooth-L1 损失函数和 CIoU 损失函数相结合的方法，从而提升了模型的训练效率和性能。

2 实验结果及分析

2.1 实验数据集与评价标准

本文在已公开的探测水下物体数据集(Detecting Underwater Objects, DUO)[30]上对本文的模型进行实验验证。文献[30]在水下机器人专业大赛(Underwater Robot Professional Contest, URPC)于 2017~2020 年公布的数据集的基础上，删减重复和极其相似的图像并且进行重新标注，构建了水下目标检测数据集 DUO，该数据集一共包含 7 782 张图像，具体就是：训练集 6 671 张图像，测试集 1 111 张图像，其中包含 4 种生物，分别是海参、海胆、扇贝、海星。对该数据集使用翻转、剪裁、去雾等数据增强方法，将数据集的数量进行扩充，对训练集和测试集的数量分别扩充一倍。部分图片如图 4 所示。

2.2 实验设备与评价指标

本文是运行在 AMD EPYC 7371 处理器、运行内存为 24 GB、GPU 为 RTX A5000 的硬件平台上进行训练和测试模型，操作系统是 Ubuntu20.04，软件环境为 Pytorch1.12.0。在模型训练过程中，采用预训练模型的骨干网络以加速训练。为了更好地与其

他的检测网络进行性能比对，本文采用 AP（Average Precision）来衡量模型的检测精度的指标，在评估过程中，考虑了不同 IoU 阈值范围，分别设定为 0.5（记作 AP_{50}）以及 0.50 到 0.95 的范围（记作 $AP_{50:95}$）。

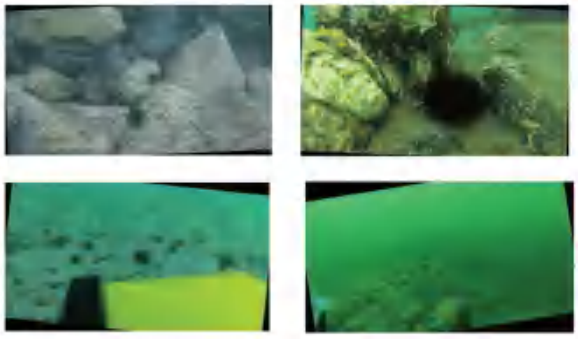

图 4　部分训练集图片

Fig. 4　Partial images in training set

2.3　消融实验

本文设置了消融实验，进一步验证所提出的每一个改进是否有效。将 DETR 作为基准模型，每组模型在 DUO 数据集上进行训练，对比每项改进对模型的检测性能影响。训练网络时 batch_size 设置为 2，初始学习率为 0.000 1，实验结果见表 1。

表 1 中，DETR 为原模型的实验结果，在没有任何改进的情况下准确度不高；对比实验 A 可知，通过使用分散注意力模块的 ResNeSt 骨干网络，有效提升了其准确度。对比实验 B 可知，添加多尺度可变形注意力编码器，模型的准确度得到了不小的提升。对比实验 C 可知，将改进的 Smooth-L1 和 CIoU 结合的损失函数替换掉原来的损失函数可知，优化后的损失函数使网络进一步收敛至更高的检测精度，且能够使网络模型更加快速地收敛。

表 1　消融实验及结果

Table 1　Ablation experiment and results

实验编号	Epoch	改进的 ResNeSt-50	多尺度可变形注意力编码器	优化的损失函数	AP_{50}	$AP_{50:95}$	AP_s	AP_M	AP_L
DETR	300				0.758	0.421	0.400	0.559	0.576
实验 A	50	√			0.794	0.562	0.412	0.601	0.595
实验 B	50		√		0.826	0.619	0.473	0.635	0.611
实验 C	50			√	0.766	0.450	0.395	0.565	0.580
实验 D	50	√	√	√	**0.834**	**0.622**	**0.480**	**0.638**	**0.613**

2.4　横向对比实验

将本文算法与其他常见的目标检测模型以及专门的水下目标检测进行对比实验。每个模型在扩充的 DUO 水下目标检测数据集上进行实验，为了进一步验证本文对小目标的检测性能，采用 MSCOCO 数据集[31]中的评价 APs 来衡量像素点小于 32×32 的小目标的检测性能。实验结果见表 2。

表 2　横向对比实验及结果

Table 2　Horizontal comparison experiment and results

模型	AP_{50}	$AP_{50:95}$	AP_s
Faster R-CNN	0.748	0.538	0.466
Cascade R-CNN	0.756	0.557	0.430
RetinaNet	0.706	0.493	0.386
FCOS	0.779	0.546	0.403
Boosting R-CNN	0.791	0.581	0.493
YOLOv5m	0.781	0.432	0.425
YOLOX	0.831	0.618	0.428
DETR	0.758	0.421	0.400
Anchor-DETR	0.823	0.607	0.435
DAB-DETR	0.821	0.605	0.431
Deformable DETR	0.826	0.619	0.473
本文	**0.834**	**0.622**	**0.480**

为了更好地比较本文所使用的方法与 DETR 原模型在水下目标检测方面的效果，图 5 展示了两者的对比结果。在左侧是 DETR 模型的检测效果，在右侧是本文的检测效果。可以看出，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。本文方法可以更好地应用于水下目标检测。

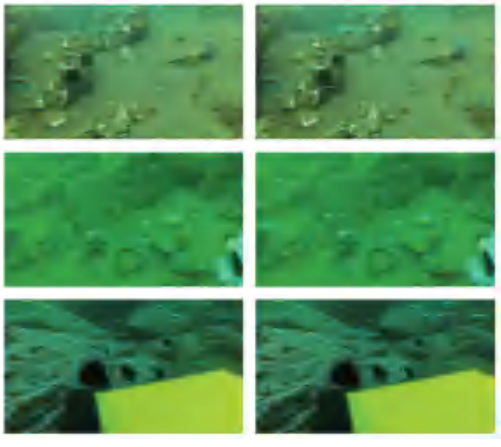

图 5　Deformable DETR 和本文测试效果图

Fig. 5　Deformable DETR and test results in this article

3 结束语

水下目标检测复杂多变,本文以 DETR 为基础框架,提出了改进的 DETR 检测算法。将小数据集进行扩充,更好地发挥基于 Transformer 的端到端的目标检测效能。采用改进的骨干网络 ResNeSt 网络作为特征提取网络,提高了模型的检测精度。采用可变形注意力机制使模型能够快速收敛,以及提高对小目标的检测能力。在训练过程中,本文采用了由 Smooth-L1 和 CIoU 组合成的损失函数,以加快模型的收敛速度和提高检测精度。实验结果表明,与原算法及其他先进算法相比,该算法不仅能够更快速地在训练阶段达到更高的精度,而且具有更高的平均检测精度。

参考文献

[1] 李柯泉,陈燕,刘佳晨,等. 基于深度学习的目标检测算法综述[J]. 计算机工程,2022,48(7):1-12.

[2] 董金耐,杨淼,谢卓冉,等. 水下图像目标检测数据集及检测算法综述[J]. 海洋技术学报,2022,41(5):60-72.

[3] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2014: 580-587.

[4] HE Kaiming, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN [C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE, 2017: 2961-2969.

[5] REN Shaoqing, HE Kaiming, 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.

[6] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, 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.

[7] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 2117-2125.

[8] LI Xiu, SHANG Min, QIN Hongwei, et al. Fast accurate fish detection and recognition of underwater images with fast R-CNN [C]//OCEANS 2015-MTS/IEEE Washington. Piscataway, NJ: IEEE, 2015: 1-5.

[9] LI Xiu, SHANG Min, HAO Jing, et al. Accelerating fish detection and recognition by sharing CNNs with objectness learning [C]//OCEANS 2016-Shanghai. Piscataway, NJ: IEEE, 2016: 1-5.

[10] ZENG Lingcai, SUN Bing, ZHU Daqi. Underwater target detection based on faster R-CNN and adversarial occlusion network[J]. Engineering Applications of Artificial Intelligence, 2021, 100: 104190.

[11] 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[J]. 西北工业大学学报,2020,38(4):747-754.

[12] LI Yan, GUO Jiahong, GUO Xiaomin, et al. Toward in situ zooplankton detection with a densely connected YOLOV3 model [J]. Applied Ocean Research, 2021, 114: 102783.

[13] LI Bing, LIU Bin, LI Shuofeng, et al. Underwater target detection based on improved YOLOv4[C]//Proceedings of 2022 41st Chinese Control Conference (CCC). Piscataway, NJ: IEEE, 2022: 7012-7017.

[14] LEI Fei, TANG Feifei, LI Shuhan. Underwater target detection algorithm based on improved YOLOv5 [J]. Journal of Marine Science and Engineering, 2022, 10(3): 310.

[15] 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[J]. 微电子学与计算机,2022,39(6):60-68.

[16] WANG Zhuo, CHEN Haojie, QIN Hongde, et al. Self-supervised pre-training joint framework: Assisting lightweight detection network for underwater object detection[J]. Journal of Marine Science and Engineering, 2023, 11(3):18.

[17] LIU Yudong, WANG Yongtao, WANG Siwei, et al. CBNet: A novel composite backbone network architecture for object detection [J]. Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(7): 11653-11660.

[18] DAI Xiyang, CHEN Yinpeng, XIAO Bin, et al. Dynamic head: Unifying object detection heads with attentions[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 7373-7382.

[19] LIU Ze, LIN Yutong, CAO Yue, et al. Swin transformer: Hierarchical vision transformer using shifted windows [C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE, 2021: 10012-10022.

[20] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers[C]//Proceedings of European Conference on Computer Vision. Cham: Springer, 2020: 213-229.

[21] ZHU Xizhou, SU Weijie, LU Lewei, et al. Deformable DETR: Deformable transformers for end-to-end object detection[EB/OL]. (2021-03-18). https://arxiv.org/abs/2010.04159.

[22] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 2117-2125.

[23] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 770-778.

[24] XIE Saining, GIRSHICK R, DOLLÁR P, et al. Aggregated residual transformations for deep neural networks [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 1492-1500.

[25] HU Jie, SHEN Li, SAMUEL A, et al. Squeeze-and-excitation networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 7132-7141.

[26] LI Xiang, WANG Wenhai, HU Xiaolin, et al. Selective kernel networks [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 510-519.

[27] GIRSHICK R. Fast R-CNN [C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway, NJ:

IEEE，2015：1440-1448.

[28] REZATOFIGHI H，TSOI N，GWAK J Y，et al. Generalized intersection over union：A metric and a loss for bounding box regression[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway，NJ：IEEE，2019：658-666.

[29] ZHENG Zhaohui，WANG Ping，LIU Wei，et al. Distance-IoU loss：Faster and better learning for bounding box regression[J]. Proceedings of the AAAI Conference on Artificial Intelligence，2020，34(7)：12993-13000.

[30] LIU Chongwei，LI Haojie，WANG Shuchang，et al. A dataset and benchmark of underwater object detection for robot picking[C]//Proceedings of 2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). Piscataway，NJ：IEEE，2021：1-6.

[31] LIN T Y，MAIRE M，BELONGIE S，et al. Microsoft coco：Common objects in context[C]//Proceedings of the 13th European Conference on Computer Vision(ECCV 2014). Cham：Springer，2014：740-755.