Computer Science and Application 计算机科学与应用, 2026, 16(1), 374-387
Published Online January 2026 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2026.161031

Hans汉斯

YOLO11-Swin：一种面向复杂水下环境的目标检测模型

郑广海*，张　倩，张　薇

大连交通大学轨道智能工程学院，辽宁 大连

收稿日期：2025年12月20日；录用日期：2026年1月19日；发布日期：2026年1月27日

摘　要

水下目标检测在海洋资源开发与生态环境监测中至关重要，但水下图像的低对比度、色彩失真及复杂背景干扰为精准检测带来巨大挑战。为克服传统方法在特征提取与小目标识别上的局限，本文提出一种深度融合Swin Transformer与YOLO11架构的新型检测模型(A Novel Detection Model with Deep Integration of Swin Transformer and YOLO11 Architectures, YOLO11-Swin)。该模型以Swin Transformer作为主干特征提取网络，利用其分层设计与滑动窗口自注意力机制，有效捕获图像的全局上下文依赖关系，增强对模糊、遮挡目标的表征能力。在特征融合阶段，本文设计了一种跨层特征聚合机制(Cross-layer Feature Aggregation, CFA)，通过全局池化与自适应权重计算，引导不同尺度特征图进行高效信息交互，以解决特征金字塔中的语义间隙与尺度不匹配问题。此外，在各级特征图输出端嵌入卷积注意力模块(Convolutional Block Attention Module, CBAM)，通过串行的通道与空间注意力子模块，自适应地优化特征响应，突出目标区域并抑制背景噪声。针对水下数据集正负样本不均衡的问题，模型采用Focal Loss作为分类损失函数，以聚焦困难样本的训练，提升模型收敛速度与稳健性。在URPC数据集上的实验结果表明，YOLO11-Swin的mAP@50达到75.54%，相比基线YOLO11模型显著提升9.42%。特别地，对小目标(如扇贝)的检测平均精度(AP)提升10.16%，召回率(Recall)提高4.55%，充分验证了所提模型在复杂水下环境下的有效性与先进性。

关键词

水下目标检测，YOLO11，Swin Transformer，跨层融合，注意力机制

YOLO11-Swin: A Target Detection Model for Complex Underwater Environments

Guanghai Zheng*, Qian Zhang, Wei Zhang

College of Intelligent Rail Engineering, Dalian Jiaotong University, Dalian Liaoning

Received: December 20, 2025; accepted: January 19, 2026; published: January 27, 2026

*通讯作者。

文章引用: 郑广海, 张倩, 张薇. YOLO11-Swin: 一种面向复杂水下环境的目标检测模型[J]. 计算机科学与应用, 2026, 16(1): 374-387. DOI: 10.12677/csa.2026.161031

Abstract

Underwater object detection plays a crucial role in marine resource development and ecological environment monitoring. However, the low contrast, color distortion, and complex background interference of underwater images pose significant challenges to accurate detection. To overcome the limitations of traditional methods in feature extraction and small object recognition, this paper proposes a novel detection model with deep integration of Swin Transformer and YOLO11 architectures (referred to as YOLO11-Swin). This model adopts Swin Transformer as the backbone feature extraction network. Leveraging its hierarchical design and sliding window self-attention mechanism, it effectively captures the global contextual dependencies of images and enhances the representation capability for blurred and occluded objects. In the feature fusion stage, a Cross-layer Feature Aggregation (CFA) mechanism is designed. Through global pooling and adaptive weight calculation, it guides efficient information interaction among feature maps of different scales, thereby addressing the issues of semantic gaps and scale mismatches in the feature pyramid. Additionally, Convolutional Block Attention Module (CBAM) is embedded at the output end of feature maps at all levels. Via serial channel and spatial attention sub-modules, it adaptively optimizes feature responses, highlights object regions, and suppresses background noise. To tackle the problem of imbalanced positive and negative samples in underwater datasets, the model employs Focal Loss as the classification loss function. This focuses on the training of hard samples, improving the model's convergence speed and robustness. Experimental results on the URPC dataset demonstrate that the mAP@50 of YOLO11-Swin reaches 75.54%, which is a significant increase of 9.42% compared to the baseline YOLO11 model. Specifically, the Average Precision (AP) for small objects (e.g., scallops) is improved by 10.16%, and the Recall is increased by 4.55%. These results fully verify the effectiveness and advancement of the proposed model in complex underwater environments.

Keywords

Underwater Object Detection, YOLO11, Swin-Transformer, Cross-Layer Fusion, Attention Mechanism

1. 引言

水下目标检测是海洋工程与水下机器人技术中的核心任务，广泛应用于海洋生物资源(如海参、海胆等)勘探、水下结构监测及生态环境评估等领域，对于海洋资源的可持续开发与生态保护具有不可替代的现实意义。然而，复杂多变的水下成像环境为目标检测带来了严峻挑战。一方面，水体对光线的吸收与散射使图像普遍存在对比度低、色彩偏蓝绿色失真等问题；另一方面，悬浮颗粒引起的背景杂波掩盖了目标边缘[1]，尤其对于如幼体扇贝等小目标，易与背景发生混淆。此外，水下目标尺寸跨度大、形态差异显著，也进一步增加了特征匹配与尺度建模的难度，传统检测算法在此类复杂场景中表现不佳[2] [3]。

现有基于卷积神经网络(CNN)的目标检测方法(如 YOLO 系列、Faster R-CNN)在通用场景中取得了一定成果，但由于其核心构件——卷积操作本质上具有局部感受野限制，难以捕捉长距离的上下文依赖信息，因此在低对比度、小目标密集或复杂背景下的水下图像中常出现特征表达不足、定位偏差等问题。

为突破这一瓶颈，本文引入 Swin Transformer 的移位窗口自注意力机制[4]。该机制通过将自注意力计算限制于非重叠的局部窗口，再通过相邻层窗口移位操作实现跨窗口特征交互，既保留了线性计算复杂度，又增强了模型对局部细节与全局结构的建模能力。具体而言，Swin Transformer 在小目标边缘纹理建模和复杂背景区域的空间建模方面表现出显著优势，可有效弥补传统卷积模型在全局建模上的不足。此外，Swin Transformer 的金字塔式层级架构自然输出三种不同下采样率(8×、16×、32×)的特征图，分别对应于小目标的细节捕捉、中等目标的轮廓建模以及大目标的语义提取，满足水下目标检测在多尺度建模上的多样化需求[5]。基于此特性，本文将其深度融合至 YOLO11 框架中，提出 YOLO11-Swin 水下目标检测算法。

YOLO11 作为单阶段检测器，采用 C3K2 模块替代 YOLOv8 的 C2f 模块，通过增加跨阶段特征交互路径，强化了不同层级特征的传递效率，尤其利于保留水下小目标的低层细节特征；其解耦检测头设计将分类与回归任务分离，有效缓解了在水下场景中，分类任务需聚焦目标与背景的语义差异，回归任务需精准定位模糊边缘的矛盾[6]；同时 YOLO11 优化了特征金字塔的通道分配策略，通过动态调整不同尺度特征的通道占比，提升了对尺度多变目标的适配性，这与水下目标尺度跨度大的特点高度契合。然而，YOLO11 原始主干网络仍采用 CSPDarknet 架构，依赖卷积操作的局部特征提取范式，无法突破全局建模能力的局限，且传统 FPN＋PAN 融合路径固化，导致小目标特征在跨尺度传递中易丢失。在水下低对比度、强杂波场景中，仍存在小目标漏检、模糊目标定位偏差等问题，无法全面解决水下检测的核心痛点[7]。

跨层特征融合与注意力机制的发展，为水下检测的性能提升提供了关键支撑。CMNet 提出的相邻层特征引导(ALFG)策略，通过高层语义信息引导低层特征学习，有效增强了低层级特征的语义表达，尤其对小目标检测增益显著；SWD-YOLO 通过动态卷积(DynamicConv)聚合多卷积核增强特征选择性，结合小波池化(WaveletPool)降低冗余，实现了模型轻量化与精度的协同提升，为水下资源受限场景提供了设计思路。注意力机制方面，CBAM (卷积块注意力模块)通过通道–空间双分支注意力自适应强化目标区域，其通道注意力的全局池化与空间注意力的 7 × 7 卷积，已被证实能有效过滤水下背景杂波 ；LSKA (大型可分离核注意力)通过分解大卷积核降低计算成本，为注意力模块的轻量化设计提供了核分解思路[8]。然而，现有跨层融合方法多局限于相邻层交互(如 CMNet)，全局适配性弱；注意力机制在水下场景中的应用仍缺乏针对性设计，难以充分抑制复杂杂波干扰[9]。

针对现有算法在水下检测中的短板 CNN 模型的局部建模局限、跨层融合的路径固化、注意力机制的场景适配不足，本文提出 YOLO11-Swin 水下目标检测算法，做出了以下改进：

1. 本文提出的水下目标检测算法 YOLO11-Swin。首次将 Swin Transformer 替代 YOLO11 原始卷积主干(CSPDarknet)，利用其移窗多头自注意力机制与金字塔式多尺度输出(8×、16×、32×下采样)，显著增强对水下模糊边缘、低对比度目标及复杂背景的全局建模能力，解决局部感受野受限问题；

2. 在 Swin Transformer 输出的三尺度特征层嵌入通道-空间双重注意力模块，通过通道权重的 softmax 竞争性归一化与空间显著性建模，自适应强化目标区域响应，引导模型更加聚焦于目标本身，抑制背景干扰，尤其提升小目标与部分遮挡目标的检测稳定性；

3. 打破 YOLO11 传统渐进式金字塔融合路径，设计全局池化与自适应权重引导的跨层交互结构，通过可学习参数动态调整多尺度特征权重，实现低层细节与高层语义的深度互补，有效应对水下目标尺度剧烈变化与边缘模糊问题；

4. 在 IoU 回归损失基础上引入 Focal Loss，通过动态加权机制抑制易分类样本梯度、强化难样本学习，解决水下场景中正负样本不平衡及小目标定位精度不足问题，提升训练稳定性与收敛效率。

2. YOLO11-Swin 算法设计

2.1. 整体架构

YOLO11-Swin 网络结构主要由三部分组成：主干网络(Backbone)、颈部网络(Neck)和检测头(Head)。其中，Swin Transformer 分别输出三个尺度(8×、16×、32×)的特征图，首先通过通道－空间双重注意力模块(Channel-Spatial Attention Module)进行特征增强，随后引入跨层特征融合机制(Cross-scale Feature Aggregation, CFA)实现多尺度信息融合。融合后的三尺度特征被分别输入至 YOLO11 解耦检测头，以并行完成目标的分类与回归任务。通过模块重构与机制优化，所设计网络能够有效应对水下图像中存在的低对比度、目标模糊及背景杂波等关键问题。网络整体结构如图 1 所示。

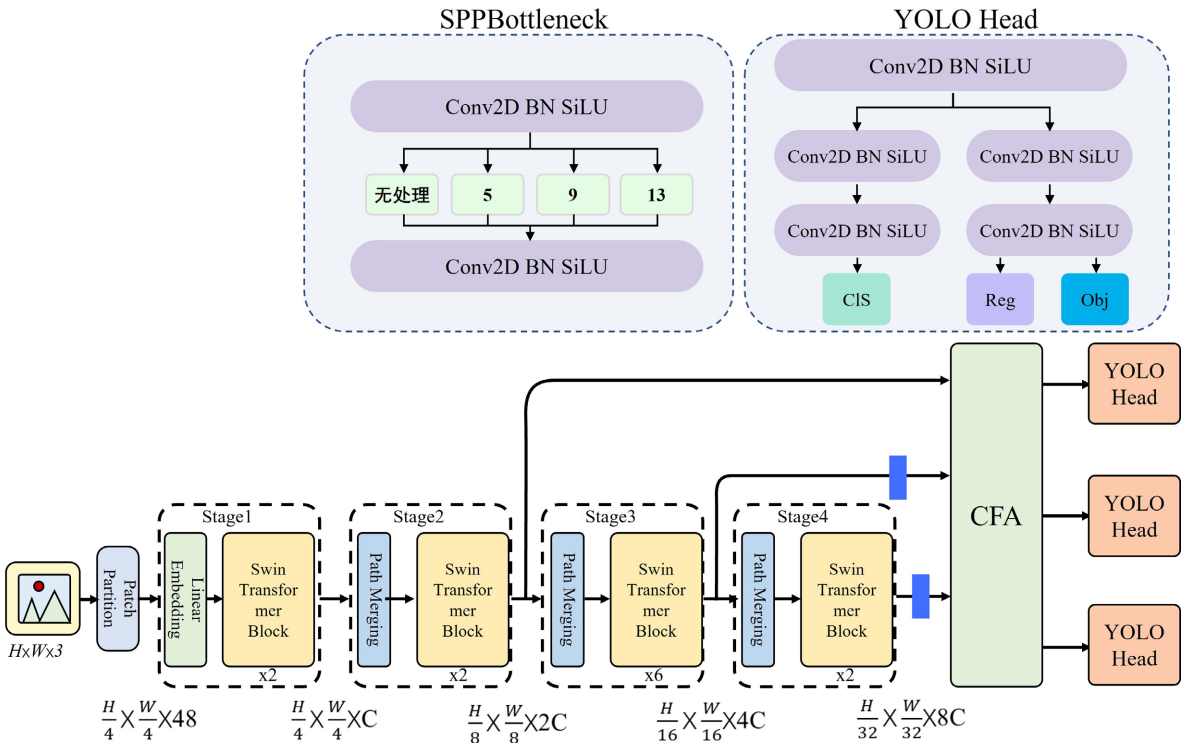

Figure 1. YOLO11-Swin modeling framework
图 1. YOLO11-Swin 模型框架

主干网络(Backbone)部分采用 Swin Transformer 替代 YOLO11 原始的 CSPDarknet 卷积结构，两者在特征提取范式上存在显著差异。CSPDarknet 借助卷积操作的局部感受野来提取图像细节，虽具备一定的水下目标局部纹理感知能力，但由于受限于卷积核大小，其感受野难以覆盖图像全局，导致模型在建模“目标与远端背景”以及“分散小目标之间”的长距离依赖关系方面存在不足。相较之下，Swin Transformer 引入移位窗口自注意力机制，将注意力计算限制在非重叠的局部窗口内，有效控制了计算复杂度(与图像尺寸呈线性关系)，同时通过相邻层间窗口位置的交错设置，实现跨窗口的信息交互。该机制不仅保留了局部细节的建模能力，还具备较强的全局语义建模能力，从而更适用于复杂水下环境中的目标特征表达。Swin Transformer 在 Stage2 至 Stage4 分别输出下采样倍数为 8×、16×和 32×的多尺度特征图，可有效对应水下目标检测中对“小目标细节(8×)－中等目标轮廓(16×)－大目标语义信息(32×)”的多层次需求。在颈部网络(Neck)设计中，引入跨尺度特征融合机制(Cross-scale Feature Aggregation, CFA)，突破传

统 FPN 与 PAN 所采用的渐进式特征融合路径。CFA 首先通过全局池化提取各尺度特征图的全局语义描述符，随后借助多层感知机(MLP)生成自适应融合权重，动态调节不同尺度特征的响应强度，实现低层细节与高层语义的深度互补。在检测头(Head)部分，本文沿用 YOLO11 的解耦检测头设计，将分类分支与回归分支结构上分离，提升模型在复杂场景下的判别能力与定位精度。同时，损失函数中引入 Focal Loss 与 EIoU Loss 的联合优化策略，有效缓解了水下目标检测中普遍存在的正负样本比例失衡问题。此外，为进一步抑制水下图像中常见的背景杂波干扰，在 Swin Transformer 的 Stage 2 至 Stage 4 中嵌入通道－空间双重注意力模块，用于增强目标区域特征并抑制冗余信息，如图 1 中蓝色模块所示。

2.2. 关键模块设计

2.2.1. 通道－空间双重注意力模块

为增强 Swin Transformer 主干网络对多尺度特征的表征能力，本研究设计了一种融合通道－空间双重注意力的多尺度注意力模块(结构如图 2 所示)。

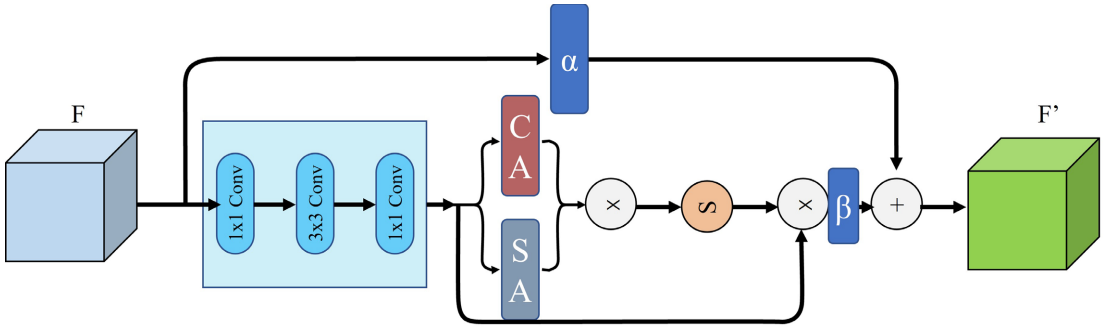

Figure 2. Channel-spatial dual attention block
图 2. 通道－空间双重注意力模块

该模块以 Swin Transformer 输出的多尺度特征图 $X \in \mathbb{R}^{H\times W\times C}$ (H、W 为特征图高宽，C 为通道数)为输入，分为通道注意力分支、空间注意力分支与轻量残差融合三部分，通道注意力分支采用“双池化－压缩扩展－竞争性归一化”策略，通过捕捉通道间依赖关系，突出目标特征所在通道(如扇贝边缘纹理通道)、抑制背景冗余通道(如悬浮颗粒噪声通道) [10]。首先对输入特征并行执行自适应最大池化与平均池化，提取能反映通道全局分布的统计特征，公式如下：

$$F_{\text{gap}} = \text{AdaptiveAvgPool2}d\left(X\right),\ F_{\text{gmp}} = \text{AdaptiveMaxPool2}d\left(X\right) \tag{1}$$

其中，AdaptiveAvgPool2d 与 AdaptiveMaxPool2d 分别为自适应平均池化与自适应最大池化操作，可将任意尺寸的输入特征压缩为 1 × 1 × C 维度，确保有效捕捉通道全局信息[11]。随后通过 1 ×1 卷积完成通道维度的压缩扩展，减少计算复杂度的同时强化通道间关联，公式为：

$$F_{\text{comp}} = ReLU\left(W_1\left(F_{\text{gap}} + F_{\text{gmp}}\right)\right),\ F_{\text{exp}} = W_2 F_{\text{comp}} \tag{2}$$

式中，$W_1 \in \mathbb{R}$，$W_2 \in \mathbb{R}$ 为 1 × 1 卷积核参数，先将通道数压缩至原维度的 1/4 以降低计算成本，经 ReLU 激活函数引入非线性后，再扩展回原通道数 C，有效捕捉通道间的复杂依赖关系[12]。最后采用 softmax 函数进行竞争性归一化，生成通道注意力权重 $Att_c \in \mathbb{R}$，相比传统 sigmoid 归一化，softmax 的竞争性机制能更显著区分关键与非关键通道，公式如下：

$$Att_c = \text{softmax}\left(F_{\text{exp}}\right) \tag{3}$$

Sigmoid 函数通过 $\sigma(x)=\frac{1}{1+e^{-x}}$ 将每个通道的权重独立映射至[0,1]区间，各通道权重互不干扰，若输入特征中存在多个通道携带相似的目标或背景信息，sigmoid 会对这些通道均赋予中等程度的权重，无法有效区分“关键通道”(如承载扇贝边缘纹理的通道)与“非关键通道”(如承载悬浮颗粒噪声的通道)。而 softmax 函数通过 $\text{soft max}(x_i)=\frac{e^{x_i}}{\sum_{j=1}^{C}e^{x_j}}$ 实现全局归一化，所有通道权重之和为 1，这意味着某一通道权重的提升必然伴随其他通道权重的降低。例如在水下图像中，若承载小目标细节的通道特征响应较高，softmax 会将其权重提升至 0.2~0.3 (远高于平均权重 $\frac{1}{C}$)，同时将背景噪声通道的权重压制至 0.01 以下，显著强化关键通道的贡献，sigmoid 函数的梯度在输入值接近 0 时最大，对特征差异的敏感度较低，当多个通道的特征响应差异较小时，sigmoid 输出的权重差异会进一步缩小，导致关键通道被淹没。而 softmax 通过指数函数放大特征响应的差异，若关键通道的特征响应比非关键通道高 1，经指数运算后差异会扩大至 $e^1\approx 2.718$ 倍，再通过全局归一化，关键通道的权重优势会被进一步凸显。

空间注意力分支通过“特征压缩–上下文建模–显著性生成”流程，削弱背景杂波干扰、强化目标空间区域响应[13]。首先沿通道维度对输入特征执行平均池化与最大池化，保留空间维度的多尺度信息并压缩通道维度，公式为：

$$F_{avg}^{s}=\text{AvgPool2}d(X,1),F_{\max}^{s}=\text{MaxPool2}d(X,1) \tag{4}$$

将池化结果沿通道维度拼接，得到 H × W × 2 的空间特征，融合平均池化的全局信息与最大池化的局部峰值信息，公式如下：

$$F_{\text{cat}}=Concat\left(\left[F_{avg}^{s},F_{\max}^{s}\right]\right) \tag{5}$$

为适配水下目标的不规则轮廓，采用 7 × 7 卷积捕捉空间局部上下文依赖，再通过 sigmoid 函数生成空间注意力权重 $Att_s\in\mathbb{R}$，将权重映射至[0,1]区间，对目标区域赋予高权重、背景区域赋予低权重，公式为：

$$F_{conv}=\text{Conv2d}(F_{\text{cat}},kernerl\_size=7,padding=3),\ Att_s=sigmoid(F_{CONV}) \tag{6}$$

为避免注意力机制过度改变 Swin Transformer 主干特征的分布，模块采用轻量残差融合策略，将注意力增强特征与原始特征线性叠加，既维持全局特征稳定性，又精准强化目标局部细节，通过消融实验确定原始特征与注意力特征的权重比为 0.9:0.1，以平衡特征稳定性与细节增强。公式如下：

$$X_{out}=0.9X+0.1\times\left(Att_c\otimes\left(Att_s\otimes X\right)\right) \tag{7}$$

其中，$\otimes$表示逐元素乘法，0.9 为原始特征权重，0.1 为注意力增强特征权重。实验表明，该融合策略能有效提升水下小目标(扇贝) AP 1.18%，同时避免模型过拟合至背景杂波特征[14]。

2.2.2. 跨层融合机制(CFA)

针对传统 YOLO 系列依赖的 FPN + PAN 结构在特征传递路径固定、跨尺度信息交互能力不足等问题，本文设计了一种跨层融合机制，以实现动态特征重组。相比于 FPN，CFA 在信息交互的直接性与特征调控的动态性两个维度上实现了显著优化。传统 FPN 通过“高层特征上采样–与相邻低层特征拼接–卷积融合”的渐进式融合路径，导致跨尺度信息需经多阶段间接传递。在复杂水下环境中，该方式容易造成小目标边缘模糊、尺度剧烈变化场景下的信息损失；此外，FPN 对各尺度特征的权重分配固定，依赖卷积核的隐式调控，难以适应水下目标分布动态变化的特性，例如小目标密集区域或低对比度背景区域，易出

现“大目标语义冗余掩盖小目标细节”或“背景杂波干扰目标特征”等问题[15] (结构如图 3 所示)。

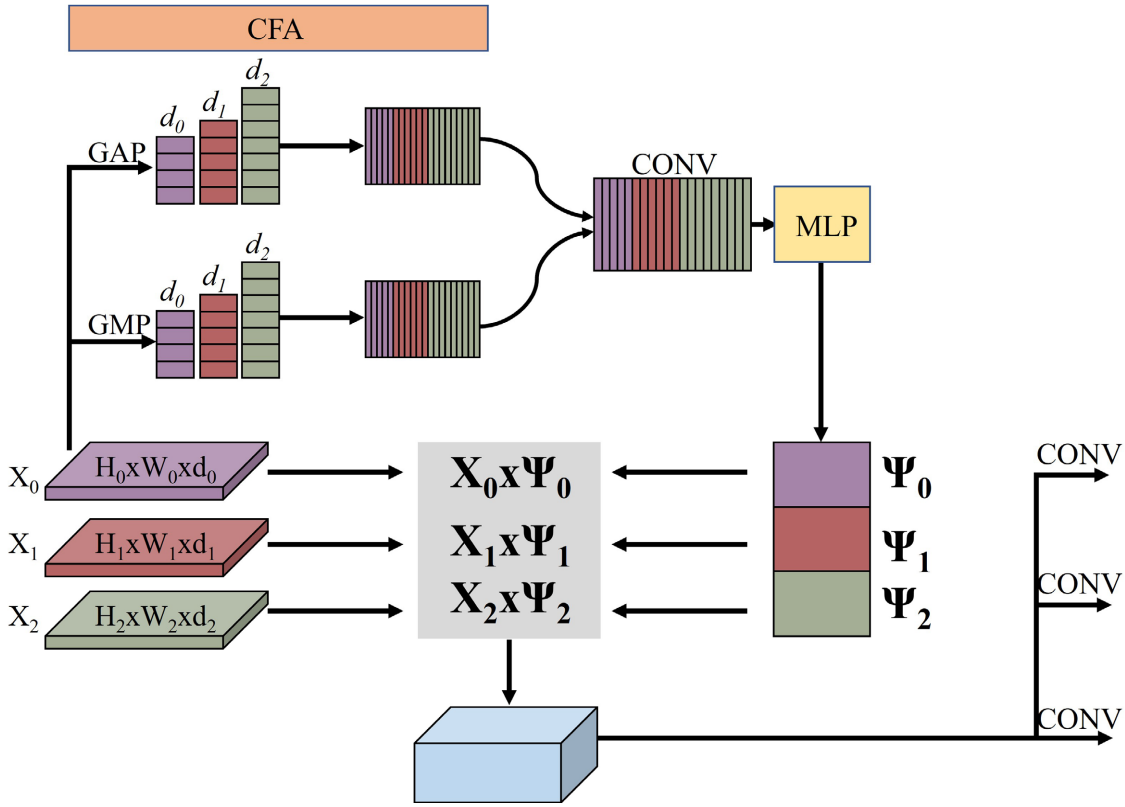

Figure 3. CFA module diagram
图 3. CFA 模块图

为突破上述限制，CFA 采用“全局感知–动态加权–直接调控”的融合策略。首先，对 Swin Transformer 输出的三个尺度特征图(8×、16×、32×，对应 Stage2 至 Stage4)进行 1 × 1 卷积，统一通道维度至 256，以消除通道不一致带来的交互障碍，避免因维度差异引起的特征权重不均衡[16]。随后，并行执行全局平均池化(Global Average Pooling, GAP)与全局最大池化(Global Max Pooling, GMP)：其中 GAP 提取通道维度上的统计信息，增强目标的整体语义表达；GMP 捕捉空间维度上的局部响应峰值，突出边缘细节。二者结合，有效覆盖水下目标从“全局轮廓”到“局部纹理”的多层次特征需求。接着，将 GAP 与 GMP 得到的全局描述符(均为 1 × 1 × 256)加权拼接后输入多层感知机，以生成自适应融合权重。该权重可根据输入图像中目标的尺度分布与对比度水平，动态评估各尺度特征的重要性，并用于直接调控原始三尺度特征的响应强度。通过这种方式，可实现无需中间层传递的低层细节与高层语义的实时深度互补[17]。这种直接调控机制使得低尺度(8×)的小目标细节可与高尺度(32×)的全局语义信息建立直接关联，有效减少多阶段传递过程中的信息损耗，提升小目标检测性能。

从特征权重分配来看，传统 FPN 对不同尺度特征的权重分配固定(仅通过 3 × 3 卷积核隐式调整)，无法适配水下目标的动态分布特性。例如在清澈水下环境中，小目标细节清晰，8×尺度特征应占更高权重；而在浑浊环境中，大目标轮廓更依赖 32×尺度的语义信息，固定权重会导致特征适配性不足。CFA 的动态加权设计则通过全局描述符与 MLP 的结合，实现权重的像素级自适应调整，其核心在于将各尺度特征的全局信息(GAP 与 GMP)作为权重生成的依据，而非依赖预设规则。具体而言，当输入图像中目标与背

景灰度差小于 10 (低对比度场景)时，MLP 会提升 32×尺度特征的权重 ω_2 (通常增加 0.1~0.2)，利用高层语义信息辅助目标定位；当图像中小目标数量占比超过 60%时，8×尺度特征的权重 ω_0 会显著提升，强化细节特征的贡献。这种动态调整机制与水下场景的多样性高度契合，而 FPN 的固定权重设计无法实现此类适配，在 URPC 数据集实验中，CFA 对扇贝(小目标)检测的 AP 提升 6.35%，进一步验证其对水下复杂场景的适配能力与有效性[18]。

2.3. 损失函数改进

针对水下目标检测中普遍存在的正负样本极度不平衡、小目标定位困难以及边界框回归置信度偏移等挑战，本文在传统 IOU 系列回归损失函数的基础上，引入焦点损失函数(Focal Loss)，构建融合型损失函数(Focal Loss + EIoU Loss)，以同步提升模型的定位精度与训练稳定性。IOU (Intersection over Union)损失函数通过几何重叠区域衡量预测框与真实框之间的差异，能够较为准确地刻画定位误差。然而，在训练初期或背景复杂场景中，IOU 损失容易受到正负样本不平衡的影响，尤其对低质量预测框的梯度更新极为敏感，导致优化过程不稳定。此外，IOU 的计算需遍历每个预测框与对应真实框的交并区域，在样本量较大时计算开销显著，且初始阶段大量无重叠或低重叠预测框的存在会进一步加剧训练震荡，降低模型的收敛速度。

为缓解上述问题，本文引入焦点损失函数(Focal Loss)以增强模型对难分类样本的学习能力。该损失函数首先通过 Sigmoid 函数将模型输出转换为概率值，并根据真实标签划分正负样本后计算对数损失；随后引入焦点调制因子，通过指数加权机制抑制易分类样本的损失贡献，同时放大难分类样本的梯度信号。此外，通过调节负样本权重，有效压制背景干扰区域的误导性梯度，从而提升模型的判别能力。最终，本文将焦点损失与 EIoU 回归损失进行加权融合，构建统一的目标检测损失函数，在分类与回归任务中实现协同优化。该设计不仅提升了模型对小目标和边界不确定目标的鲁棒性，还在复杂水下场景中显著提高了整体检测性能。

当 Focal Loss 权重为 0.3、EIoU Loss 权重为 0.7 时，模型 mAP@50 最高(75.54%)，公式如下：

$$L_{\text{total}} = 0.3L_{\text{Focal}} + 0.7L_{EIoU} \tag{8}$$

其中，L_{total} 为改进的 IoU 损失，相比传统 IoU 损失，进一步考虑边界框的宽高比差异，公式为：

$$L_{EIoU} = 1 - IoU + \frac{\rho^2\left(b, b^{gt}\right)}{\left(w^c\right)^2 + \left(h^c\right)^2} + \frac{\rho^2\left(w, w^{gt}\right)}{\left(w^c\right)^2} + \frac{\rho^2\left(h, h^{gt}\right)}{\left(h^c\right)^2} \tag{9}$$

式中，b 和 b^{gt} 分别为预测框和真实框的中心坐标，w, h，w^{gt}，h^{gt} 分别为预测框和真实框的宽高，$\left(w^c\right)$、$\left(h^c\right)$ 为预测框和真实框的最小外接矩形的宽高，ρ 为欧氏距离。

L_{Focal} 为焦点损失，公式为：

$$L_{\text{Focal}} = -\sum_{i=1}^{N} y_i \left(\left(1 - p_i\right)^{\gamma} \log\left(p_i\right)\right) + \left(1 - y_i\right) p_i^{\gamma} \log\left(1 - p_i\right) \tag{10}$$

其中，y_i 为样本标签(1 表示正样本，0 表示负样本)，p_i 为模型预测的正样本概率，γ 为焦点因子，经实验设置 $\gamma = 2$ 时效果最佳。

该设计既保留 IOU 对几何误差的高敏感性，又通过分类难度动态加权机制增强判别能力，有效缓解样本稀疏与背景杂波干扰。尤其在处理小目标或模糊边界时，焦点损失弥补了 IOU 梯度稀疏的缺陷，显著提升训练稳定性、收敛效率及检测精度，为水下复杂场景目标检测提供了鲁棒的优化方案。

3. 实验结果与分析

3.1. 实验平台

实验的操作系统是 Windows 11 专业版，处理器是 AMD Ryzen 9 5950X (16 cores)，运行内存是 264 GB，GPU 模型是 RTX 3090，实验是在 PyTorch 1.7.1 深度学习框架，Cuda 11.0 架构，和 Python 版本是 3.9。训练参数：batch size 设置为 4，epoch 设置为 100，初始学习率为 0.01，最终学习率为 0.01，输入图像的大小自动缩放到 640 x640，不使用预训练权重值，其他参数为默认值。

3.2. 实验数据集

本研究采用 URPC 水下目标检测基准数据集进行模型训练与验证。该数据集包含 5543 张真实水下场景图像，涵盖海参(holothurian)、海胆(echinus)、扇贝(scallop)、海星(starfish) 4 类典型海洋生物目标，共标注超过 5000 个实例。由于水下成像受光照衰减、色彩失真及浑浊度影响，数据集中图像普遍存在低对比度、蓝绿色偏移、纹理模糊等退化现象，显著增加检测难度。为保障实验严谨性，数据集按 8:1:1 比例划分为训练集、验证集与测试集，确保模型在独立样本上评估泛化性能。针对上述退化问题，本研究设计针对性数据增强策略：通过随机色彩抖动(亮度、对比度、饱和度调整)模拟不同水深光照条件，结合高斯模糊与噪声叠加仿真悬浮颗粒散射效应，提升模型对低质量图像的鲁棒性。该策略通过扩展训练样本分布，强化模型对复杂水下环境的适应能力，为后续检测任务奠定数据基础。

3.3. 评价指标

选取参数量、计算量、精度、召回率和 mean Average Precision(mAP@50)作为模型的评价指标。其中，mAP@50 由精度 P 和召回率 R 计算得出。

精度 P 为：

$$p = \frac{TP}{TP + FP} \tag{11}$$

召回率 R 为：

$$R = \frac{TP}{TP + FN} \tag{12}$$

mAP@50 为：$AP = \int_0^1 dR$

$$mAP@50 = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{13}$$

其中 TP 为判断正确的阳性样本数，FP 为错误检测到的样本数，FN 为遗漏的样本数，AP 为由精度 P 和 R 组成的关于轴的曲线面积；mAP 为所有 AP 的平均值，mAP@50 中的 i 表示当前类别。当 mAP@50 较高时，这意味着模型被训练得更好。

3.4. 消融实验

为验证 YOLO11-Swin 算法中各核心模块对水下目标检测性能的提升作用，本文基于 URPC 数据集，以“YOLO11-Swin 主干网络 + 数据增强”作为基准，设计系列消融实验，逐步引入通道－空间双重注意力模块、跨层融合机制及融合型损失函数。实验结果表明，各模块均能有效缓解水下检测中的关键难题，且存在显著的协同增益。

仅采用数据增强策略时，通过色彩抖动模拟水深光照变化，并结合高斯模糊及噪声叠加模拟悬浮颗粒散射，模型的 mAP@50 提升了 1.44%，召回率提升了 5.59%，验证了数据层面优化对提升水下场景鲁棒性的基础作用。引入通道–空间双重注意力模块后，mAP@50 进一步提升了 0.81%，小目标扇贝的 AP 提升了 1.18%，表明该模块通过通道竞争性归一化与空间显著性建模，有效抑制背景杂波干扰，强化了小目标特征表示[3]。

叠加跨层融合机制后，mAP@50 再提升 1.57%，其中扇贝 AP 大幅提升 6.35%，充分证明了 CFA 通过全局描述符提取与自适应权重引导，突破传统 FPN + PAN 的刚性融合路径，显著缓解了水下目标尺度失配问题。最终集成融合型损失函数后，mAP@50 达到 75.54% (较基准提升 4.52%)，召回率提升 3.96%，低对比度目标海参的 AP 提升 4.96%，表明 Focal Loss 对难分类样本的梯度强化以及 EIoU Loss 对定位误差的精细度量，有效解决了水下场景中正负样本极度不平衡及小目标定位精度不足的问题。

总体来看，各模块的逐步集成在模型参数量适度增加(从 94.34 M 增至 100.81 M)的前提下，实现了检测精度与鲁棒性的同步提升，尤其在水下小目标及低对比度目标的识别性能上表现突出，充分验证了所设计算法的针对性和有效性。部分实验结果及数据可视化示例如表 1 和图 4 所示。

Table 1. Ablation experiment
表 1. 消融实验

	YOLO11-swin				
数据增强		√	√	√	√
注意力模块			√	√	√
跨层融合				√	√
损失函数改进					√
Parameters/M	94.34	94.34	98.56	100.81	100.81
mAP@50/%	69.58	71.02	71.83	73.40	75.54
P/%	86.79	81.66	81.39	82.51	81.89
R/%	52.46	58.05	58.95	58.64	62.60

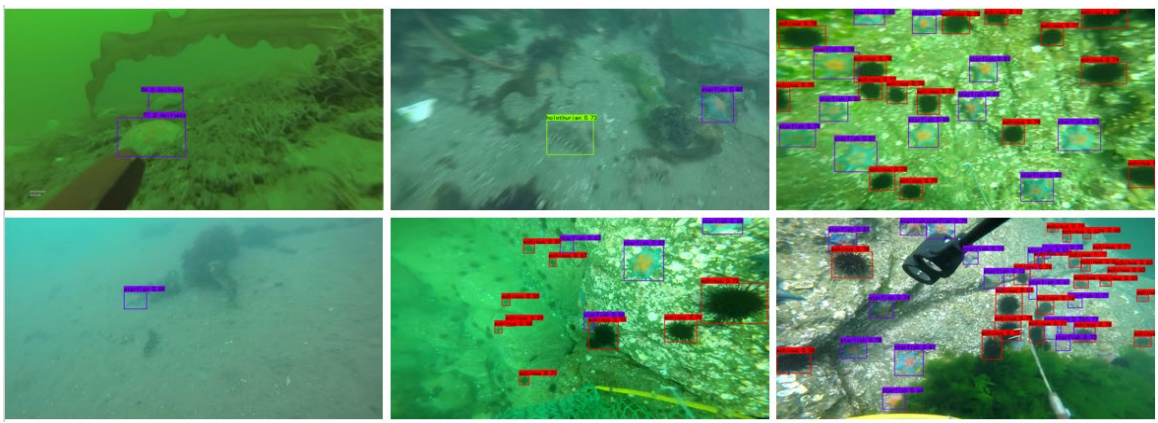

Figure 4. Example of YOLO-Swin detection visualization
图 4. YOLO-Swin 检测可视化示例

3.5. 详细对比实验

为了达到更全面的实验结果和评价，我们对平均精度(mAP@50)，精度和召回率一系列评价指标进行

了详细的对比实验。

对于平均精度，引入通道－空间双重注意力模块后，mAP@50 提升至 71.83%，提升 0.81%，该模块通过强化目标相关通道与空间区域响应，有效抑制水下杂波干扰，使小目标(扇贝) AP 从 63.44%提升至 64.62%；进一步集成跨层融合机制后，mAP@50 增至 73.40%，再提升 1.57%，大幅缓解尺度失配问题，其中扇贝 AP 因多尺度特征深度交互显著提升至 70.97%，海胆(高对比度目标) AP 从 85.38%提升至 87.99% [1] [3]；最终引入融合型损失函数后，mAP@50 达到 75.54%，较基准累计提升 4.52%，损失函数对难分类样本的梯度强化与定位误差的精细度量，使低对比度目标(海参) AP 从 54.76%提升至 56.46%，海星(轮廓清晰目标) AP 从 80.52%提升至 83.99% [2] [5]与现有主流算法对比，YOLO11-Swin 的 mAP@50 显著优于 Faster R-CNN (39.12%)、CenterNet (55.90%)等传统模型，较同系列的 YOLO11 (66.12%)提升 9.42%、较 YOLOv8 (72.56%)提升 2.98%，尤其在小目标检测上表现突出，扇贝 AP 较基准提升 10.16%，充分验证了该算法在复杂水下环境中对不同类型目标的高精度识别能力，实验结果如表 2 所示。

Table 2. Comparison of average accuracy of four types of underwater targets before and after improvement
表 2. 改进前后四类水下目的平均精度对比

YOLO11-swin					
数据增强		√	√	√	√
注意力机制			√	√	√
跨层融合(CFA)				√	√
损失函数改进					√
echinus (AP/%)	85.87	85.38	86.41	87.99	88.13
holothurian (AP/%)	54.74	54.76	55.06	51.50	56.46
scallop (AP/%)	58.57	63.44	64.62	70.97	73.60
starfish (AP/%)	79.14	80.52	81.21	83.14	83.99

Table 3. Comparison of accuracy of four types of underwater targets before and after improvement
表 3. 改进前后四类水下目标的精度对比

YOLO11-swin					
数据增强		√	√	√	√
注意力机制			√	√	√
跨层融合(CFA)				√	√
损失函数改进					√
echinus (P/%)	90.04	86.73	86.34	87.24	86.59
holothurian (P/%)	82.89	77.02	76.37	75.72	76.07
scallop (P/%)	87.88	81.29	80.91	81.62	80.83
starfish (P/%)	86.36	81.58	81.94	85.45	84.05

对于精度 P，引入通道－空间注意力模块后，海胆(echinus)精度微降 0.39% (从 86.73%至 86.34%)，海星(starfish)精度提升 0.36% (从 81.58%至 81.94%)，整体精度小幅降至 81.39%，表明注意力机制在强化目标特征的同时，可能对部分高置信度样本的判别阈值产生微调；叠加跨层融合机制后，海星精度显著提升 3.51% (至 85.45%)，海胆精度回升 0.90% (至 87.24%)，整体精度提升至 82.51%，验证了多尺度特征交互对目标特征完整性的增强作用；最终引入损失函数改进后，海胆精度微降 0.65% (至 86.59%)，海参(holothurian)精度微升 0.35% (至 76.07%)，整体精度略降至 81.89%，但难分类样本的识别稳定性显著提升。综合来看，各模块对海星等轮廓清晰目标的精度提升更为明显(累计提升 2.47%)，而对海参等低对比度目标的精度影响较小，体现了改进策略在精度与召回率之间的动态平衡[19]。实验结果如表 3 所示。

对于召回率 R，引入通道－空间注意力模块后，整体召回率提升 0.90%至 58.95%，扇贝(scallop)召回率提升 0.83%，表明注意力机制增强了对小目标的捕捉；叠加跨层融合机制后，整体召回率微降 0.31%，但海星(starfish)召回率保持稳定；最终引入损失函数改进后，整体召回率显著提升 3.96%至 62.60%，扇贝召回率大幅提升 11.83%，海参(holothurian)召回率提升 5.82%，验证了损失优化对难分类样本召回能力的强化作用。实验结果如表 4 所示。

Table 4. Comparison of recall rates for four types of underwater targets before and after improvement
表 4. 改进前后四类水下目标的召回率对比

YOLO11-swin					
数据增强		√	√	√	√
注意力机制			√	√	√
跨层融合(CFA)				√	√
损失函数改进					√
echinus (R/%)	72.36	76.67	77.23	79.02	80.23
holothurian (R/%)	36.87	37.01	39.55	31.19	37.01
scallop (R/%)	34.84	46.58	47.41	52.96	59.24
starfish (R/%)	65.75	71.93	71.71	71.38	73.92

不同模型对比实验表明，YOLO11-Swin 在水下目标检测任务中表现最优。与 CenterNet (mAP@50 = 55.9%)、Faster R-CNN (39.12%)等经典模型相比，所提算法 mAP@50 达 75.54%，分别提升 19.64%和 36.42%，且召回率(62.60%)显著高于上述模型，验证了深度学习方法对水下复杂场景的适应性；与同系列单阶段模型相比，较 YOLOv11 (66.12%)提升 9.42%，较 YOLOv8 (72.56%)提升 2.98%，在参数量(100.81 M)适度增加的情况下，实现了精度与鲁棒性的平衡。这一结果得益于 Swin Transformer 的全局建模能力与跨层融合机制对水下低对比度、多尺度目标的针对性优化，凸显了算法在实际水下检测场景中的应用优势。参数量增加主要源于 Swin Transformer 的多头自注意力机制，但其全局建模能力带来的精度提升(+9.42% mAP@50)显著优于计算成本增加，符合复杂水下场景对检测鲁棒性的需求[20]。对比实验结果如表 5 所示，本次测试随机选取了四张图片，展示了不同算法的结果，证明了 YOLO11-Swin 在海底检测方面取得的显著进步，如可视化图 5 所示。

Table 5. Comparison of different algorithms
表 5. 不同算法的对比实验

算法模型	mAP@50/%	P/%	R/%	Parameters/M
centernet	55.9	96.11	17.6	32.67
Faster R-CNN	39.12	35.84	45.13	136.76
yolov8	72.56	88.79	47.03	11.14
yolov11	66.12	88.76	45.56	9.46
Ours	75.54	81.89	62.60	100.81

Figure 5. Visualization detection results of different algorithms
图 5. 不同算法的可视化检测结果

4. 结论

本文提出的 YOLO11-Swin 算法，融合 Swin Transformer 的全局建模能力、通道－空间双重注意力机制、跨层特征融合策略以及融合型损失函数优化，有效缓解了水下目标检测中存在的“局部感受野受限、特征融合路径刚性、背景杂波干扰严重、正负样本分布极度不平衡”等关键问题。实验结果表明，YOLO11-Swin 在 URPC 数据集上取得了显著性能提升，mAP@50 达到 75.54%，相比原始 YOLO11 提升 9.42%。其中，小目标类别(如扇贝)的 AP 提高 10.16%，召回率提高 4.55%；低对比度目标(如海参)的 AP 提升 4.96%，充分验证了该算法在复杂水下环境中的鲁棒性与适应性，为实际水下作业提供了高效可靠的目标检测方案。尽管本算法在检测精度上取得显著提升，其模型参数量(100.81 M)相较传统 YOLO 系列仍偏高。未来可通过引入知识蒸馏等模型压缩技术，将模型体积压缩至 50 M 以下，同时保持 mAP@50 不低于 73%；结合 TensorRT 等推理加速框架，进一步提升推理速度，以适配水下机器人等资源受限平台的部署需求。此外，针对极端浑浊水域(如能见度低于 1 m)的挑战场景，可探索融合“光学图像 + 声呐信息”的多模态感知机制，提升感知完整性与检测鲁棒性。同时，引入自监督学习与无监督域适应技术，有望

在减少标注依赖的同时增强模型在未知水下环境(如深海热泉区、极地冰层下等)中的泛化能力，为广泛的水下应用场景提供理论与技术支撑。

参考文献

[1] Wang, W., Sun, Y.F., Gao, W., Xu, W., Zhang, Y. and Huang, D. (2024) Quantitative Detection Algorithm for Deep-Sea Megabenthic Organisms Based on Improved YOLOv5. *Frontiers in Marine Science*, **11**, Article 1301024. https://doi.org/10.3389/fmars.2024.1301024

[2] Ge, Z., Liu, S., Wang, F., *et al.* (2021) YOLOX: Exceeding YOLO Series in 2021. arXiv: 2107.08430.

[3] Li, B., Li, X., Li, S., Zhang, Y., Liu, K., Ma, J., *et al.* (2024) Cross-Layer Feature Guided Multiscale Infrared Small Target Detection. *IEEE Geoscience and Remote Sensing Letters*, **21**, 1-5. https://doi.org/10.1109/lgrs.2024.3358953

[4] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., *et al.* (2021) Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Montreal, 10-17 October 2021, 9992-10002. https://doi.org/10.1109/iccv48922.2021.00986

[5] Monterroso Muñoz, A., Moron-Fernández, M., Cascado-Caballero, D., Diaz-del-Rio, F. and Real, P. (2023) Autonomous Underwater Vehicles: Identifying Critical Issues and Future Perspectives in Image Acquisition. *Sensors*, **23**, Article 4986. https://doi.org/10.3390/s23104986

[6] Ferrari, V. (2018) Computer Vision—ECCV 2018: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VII, in Lecture Notes in Computer Science, Vol. 11211. Springer International Publishing AG.

[7] Yang, C., Zhang, C., Jiang, L. and Zhang, X. (2024) Underwater Image Object Detection Based on Multi-Scale Feature Fusion. *Machine Vision and Applications*, **35**, Article No. 124. https://doi.org/10.1007/s00138-024-01606-3

[8] Shen, X., Wang, H., Cui, T., Guo, Z. and Fu, X. (2023) Multiple Information Perception-Based Attention in YOLO for Underwater Object Detection. *The Visual Computer*, **40**, 1415-1438. https://doi.org/10.1007/s00371-023-02858-2

[9] Hu, X., Liu, Y., Zhao, Z., Liu, J., Yang, X., Sun, C., *et al.* (2021) Real-Time Detection of Uneaten Feed Pellets in Underwater Images for Aquaculture Using an Improved YOLO-V4 Network. *Computers and Electronics in Agriculture*, **185**, Article ID: 106135. https://doi.org/10.1016/j.compag.2021.106135

[10] Li, X., Zhao, Y., Su, H., Wang, Y. and Chen, G. (2025) Efficient Underwater Object Detection Based on Feature Enhancement and Attention Detection Head. *Scientific Reports*, **15**, Article No. 5973. https://doi.org/10.1038/s41598-025-89421-2

[11] Jia, J., Fu, M., Liu, X. and Zheng, B. (2022) Underwater Object Detection Based on Improved EfficientDet. *Remote Sensing*, **14**, Article 4487. https://doi.org/10.3390/rs14184487

[12] Chen, L., Yang, Y., Wang, Z., Zhang, J., Zhou, S. and Wu, L. (2023) Underwater Target Detection Lightweight Algorithm Based on Multi-Scale Feature Fusion. *Journal of Marine Science and Engineering*, **11**, Article 320. https://doi.org/10.3390/jmse11020320

[13] Lyu, Z., Peng, A., Wang, Q. and Ding, D. (2022) An Efficient Learning-Based Method for Underwater Image Enhancement. *Displays*, **74**, Article ID: 102174. https://doi.org/10.1016/j.displa.2022.102174

[14] Liu, K. (2023) Underwater Object Detection Using TC-YOLO with Attention Mechanisms. *Sensors*, **23**, Article No. 2567. https://doi.org/10.3390/s23052567

[15] Zhang, M., Xu, S., Song, W., He, Q. and Wei, Q. (2021) Lightweight Underwater Object Detection Based on YOLO V4 and Multi-Scale Attentional Feature Fusion. *Remote Sensing*, **13**, Article 4706. https://doi.org/10.3390/rs13224706

[16] Cheng, L., Zhou, H., Le, X., Chen, W., Tao, H., Ding, J., *et al.* (2024) An Improved Underwater Object Detection Algorithm Based on YOLOv5 for Blurry Images. 2024 12*th International Conference on Intelligent Computing and Wireless Optical Communications* (*ICWOC*), Chongqing, 21-23 June 2024, 42-47. https://doi.org/10.1109/icwoc62055.2024.10684955

[17] Ahlawat, V. (2022) An Efficient Algorithm for Collision Avoidance between a Solar Array Satellite and Space Debris. *International Journal of Research in Science and Technology*, **12**, 14-24. https://doi.org/10.37648/ijrst.v12i04.004

[18] Li, X., Lv, C., Wang, W., Li, G., Yang, L. and Yang, J. (2022) Generalized Focal Loss: Towards Efficient Representation Learning for Dense Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **45**, 3139-3153. https://doi.org/10.1109/tpami.2022.3180392

[19] Yang, Q., Meng, H., Gao, Y. and Gao, D. (2023) A Real-Time Object Detection Method for Underwater Complex Environments Based on FasterNet-YOLOv7. *Journal of Real-Time Image Processing*, **21**, Article No. 8. https://doi.org/10.1007/s11554-023-01387-4

[20] Guo, A., Sun, K. and Zhang, Z. (2024) A Lightweight YOLOv8 Integrating FasterNet for Real-Time Underwater Object Detection. *Journal of Real-Time Image Processing*, **21**, Article No. 49. https://doi.org/10.1007/s11554-024-01431-x