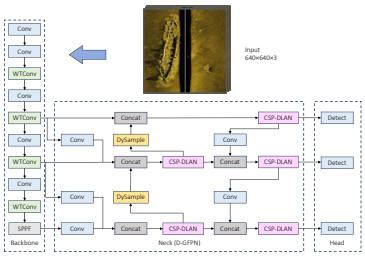
OEE | 光电工程 Opto-Electronic Engineering

Article

2026 年 6 月, 第 53 卷, 第 6 期

基于 YOLOv8-ESI 的侧扫声纳目标检测



谷剑峰[1,2], 刘锡祥[1,2*], 徐　浩[1,2]

[1]东南大学仪器科学与工程学院，江苏 南京 210096；

[2]微惯性仪表与先进导航技术教育部重点实验室，江苏 南京 210096

摘要：受水下复杂环境影响，侧扫声纳图像普遍存在强噪声干扰、几何畸变及特征模糊等问题，严重制约了传统检测模型的性能。此外，声纳图像数据集规模有限且标注质量不一，进一步加剧了模型训练的难度。针对上述问题，本文提出一种改进检测模型高效声呐图像 YOLOv8 模型 (YOLOv8-efficient sonar image, YOLOv8-ESI)。首先在骨干网络中引入小波卷积层 (WTConv)，以协同增强对图像低频结构信息与高频细节特征的提取能力；然后对颈部网络进行重构，设计了动态特征金字塔网络 (D-GFPN) 架构，融合高效动态上采样器 (DySample) 和跨阶段局部-双路径轻量聚合网络模块 (CSP-DLAN)，从而有效整合多尺度高层语义与空间细节信息；最后采用 WIoU 损失函数替代原 CIoU，通过抑制低质量样本产生的有害梯度，提升模型训练的鲁棒性。实验结果表明，YOLOv8-ESI 相比原版 YOLOv8n 在侧扫声纳数据集上性能有显著提升，mAP50 提高了 5.2%，参数减少了 10%，计算复杂度降低了 1.1 GFLOPs。本文设计的 YOLOv8-ESI 不仅实现了对侧扫声纳图像目标的精确检测，还实现了模型的轻量化设计，适配水下设备的实时检测需求。

关键词：侧扫声纳；YOLOv8n；小波卷积；动态上采样；损失函数

The English title and abstract appear at the end of the paper.

DOI: 10.12086/oee.2026.250370 | CSTR: 32245.14.oee.2026.250370 | 中图分类号：TP391.4 | 文献标识码：A

谷剑峰，刘锡祥，徐浩. 基于 YOLOv8-ESI 的侧扫声纳目标检测 [J]. 光电工程，2026，**53**(6): 250370

Gu J F, Liu X X, Xu H. Underwater object detection in side-scan sonar images based on YOLOv8-ESI[J]. *Opto-Electron Eng*, 2026, **53**(6): 250370

1　引　言

在海洋资源勘探等领域，对海底目标进行高效、精确的探测与识别具有至关重要的价值。侧扫声纳 (Side scan sonar, SSS) 作为一种主动式声学成像设备，因其探测范围广、效率高，已成为获取高分辨率海底信息的关键工具[1]。然而，受复杂水下环境影响，SSS 图像普遍存在强斑点噪声、几何畸变、目标特征模糊及分辨率不均等问题，这导致从 SSS 图像中可靠地识别目标 (如沉船、飞机残骸等) 成为了一个极具挑战性的任务。

针对 SSS 图像的目标检测研究主要经历了两个阶段。早期传统 SSS 图像目标检测主要有基于数理统计[2-3]、数学形态学[4] 以及像素分析的方法[5-6]。然而，这些方法都依赖于人工设计特征，难以自动学习 SSS 图像中复杂的深层特征表示。在面对现代海洋探测任务时，逐渐显现出适应性不足、泛化能力有限等问题。随着深度学习，特别是卷积神经网络在计算机视觉领域取得成功，基于深度学习的目标检测算法为 SSS 图像处理开辟了新的途径。主流算法可分为以 R-CNN 系列[7-8] 为代表的两阶段检测算法 (精度较高但速度慢)，以及以 YOLO

收稿日期：2025-11-28；修回日期：2026-03-04；录用日期：2026-03-10；网络出版日期：2026-06-25

基金项目：江苏省重点研发计划 (BF2024072)

*通信作者：刘锡祥，scliuseu@163.com

版权所有©2026 中国科学院光电技术研究所

系列[9]、SSD[10] 为代表的单阶段检测算法 (在速度与精度间取得更好平衡)。后者因对实时性应用的友好性，已成为当前水下目标检测的研究热点与主流改进框架。

然而，将通用的深度学习模型直接应用于 SSS 图像，仍面临一系列挑战，这些挑战与处理可见光等光学图像有显著区别。首先，声学成像过程引入的强烈斑点噪声远超光学图像中的常见噪声，导致图像信噪比极低，目标边界模糊，有效纹理信息被严重淹没，使得目标特征与背景噪声高度耦合。其次，SSS 图像中的目标常表现为特征稀疏、像素稀少的“亮区-阴影”对，且从近距到远距成像尺度差异巨大，导致小目标及关键的阴影特征在传统网络的前向传播与下采样过程中极易丢失。同时，受限于图像本身的低质量和目标特征模糊，人工标注框存在不可避免的位置偏移与尺寸误差，使得训练数据集中包含大量低质量样本，对模型优化方向产生误导性有害梯度。最后，算法常需部署于自主水下航行器等计算与能源均受限的平台，对模型的轻量化程度与实时推理效率提出了远高于一般场景的要求。这些相互关联的难题共同构成了 SSS 图像目标检测领域亟待攻克的核心瓶颈。

针对上述难点，YOLO 系列模型因其出色的速度-精度平衡特性，成为水下目标检测领域的主流检测模型。大量研究以 YOLO 各系列模型作为基准模型进行架构改进研究[11-12]，主要从以下几个方面推进：1) 轻量化设计：王妍梅等人[13] 压缩了 YOLOv3 网络的深度，将模型参数数量减少了 43%，但目标的召回率下降了 18%；高玉等人[14] 提出了基于 YOLOv7 的 GCP-YOLO 模型，在参数量和计算复杂度降低 4 倍的同时，准确率提升了 2.8%。2) 多尺度融合：袁明阳等人[15] 提出了上下文特征金字塔模块 (AC-FPN)，该模块聚合多尺度感受野，以增强上下文的信息提取；并且进一步引入上下文与内容双注意力机制，引导模型聚焦关键特征，从而提升目标定位精度。万大航等人[16] 通过引入多信息融合技术来改进 YOLOv8，使其用于灰度图像中的目标检测。3) 噪声抑制：郑林涵等人[17] 在 YOLOv8 中嵌入了空间信道增强机制 (ScEMA)，在 SonicDataset 中实现了 94.2% 的 AP (average precision)，但其性能随着检测距离的增加而出现显著衰减。

尽管上述研究在特定维度上取得了一定进展，但仍存在明显局限：首先，多数改进仅在网络结构层面进行经验性调整，未能深入结合 SSS 图像的物理成像机理 (如频域噪声特性) 进行针对性设计；其次，改进策略往往孤立地侧重轻量化、多尺度融合或噪声抑制中的单一目标，缺乏协同优化框架以同时应对模型效率、尺度鲁棒性与噪声鲁棒性等多重约束；最后，对 SSS 数据中普遍存在的标注噪声与低质量样本问题缺乏鲁棒的训练机制。因此，当前仍缺乏一套能够系统应对 SSS 图像目标检测核心挑战的端到端解决方案。

为此，本文提出一种改进的 YOLOv8-ESI 模型，其设计紧密围绕前述挑战，旨在提供一套端到端的系统性解决方案。本工作的核心贡献如下：

1) 在骨干网络引入 WTConv 模块，通过级联小波分解，将特征提取迁移至小波域，通过频带分解协同增强结构信息提取与噪声抑制能力。

2) 对颈部网络进行了系统性重构，提出动态特征金字塔网络 D-GFPN 的整体架构，优化了跨尺度连接与计算资源分配。引入高效动态上采样器 DySample，通过多层次点采样策略保留声纳图像的细腻特征。设计跨阶段局部 -双路径轻量聚合网络模块 CSP-DLAN，利用组卷积技术和分组重排机制优化信息流动，实现更高效的多尺度特征融合。

3) 采用 WIoU 替换原 CIoU，通过动态非单调聚焦机制评估锚框质量，抑制低质量样本产生的有害梯度，提升训练鲁棒性。

上述改进通过协同设计，在显著提升精度的同时，实现了模型的整体轻量化，满足了水下平台对高效率、低功耗的实时检测需求。

2 相关工作

2.1 YOLOv8 模型

YOLO 系列算法因其在目标检测任务中兼具高效率与高精度而受到广泛认可。其中，YOLOv8 作为该系列于 2023 年发布的重要版本，在前代架构基础上进行了一系列融合与改进。如图 1 所示，YOLOv8 的整体框架延续了经典的 backbone-neck-head 三级结构。相较于 YOLOv5，其在 Backbone 与 Neck 部分的核心改进在于引入 C2f 模块以替代原有的 C3 模块。C2f 模块借鉴了 YOLOv7[18] 中的 ELAN 思想，通过多分支并行和跨层连接结构，使其能够捕获更全面的梯度流信息，以及进行更充分的信息融合。在 head 部分，YOLOv8 采用了解

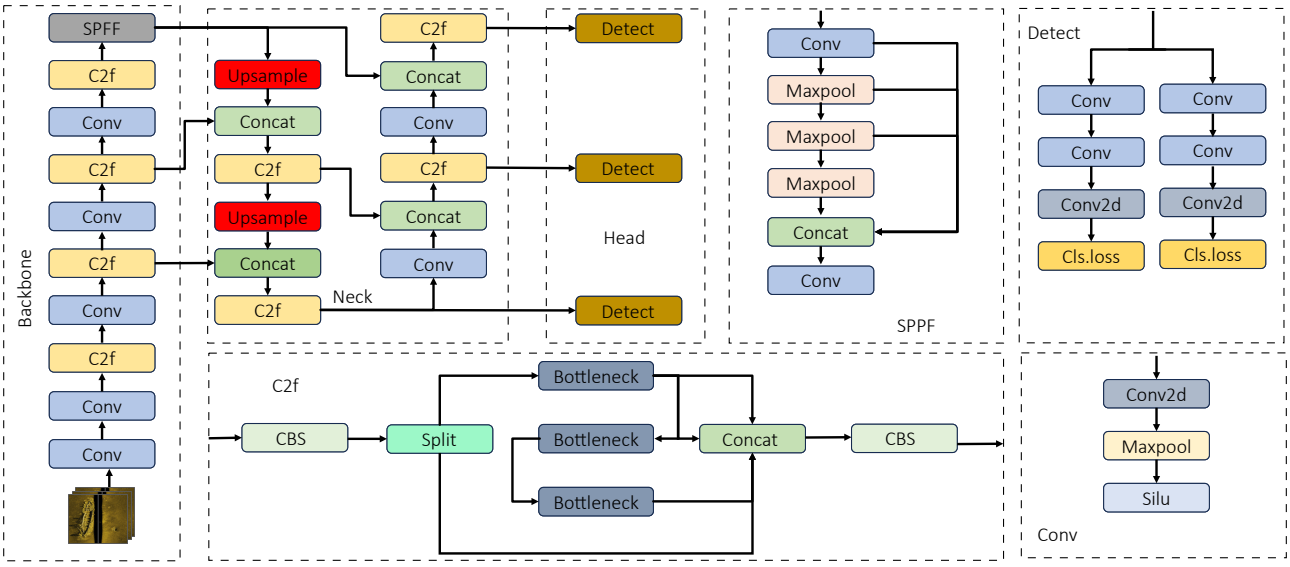


图 1 YOLOv8 网络结构图

Fig. 1 YOLOv8 network structure diagram

耦头设计，将分类 head 和检测 head 独立设计，显著提升了模型的收敛速度。在损失函数计算方面，YOLOv8 改进了损失函数，并采用 CIoU 优化边界框回归，提升了准确率和训练稳定性，并支持更高效的多任务联合优化[19]。

与 YOLO 系列的其他版本一致，YOLOv8 同样提供了五种不同尺度的预定义模型 (如 n、s、m、l、x)，以适配从移动端到数据中心等多种计算资源约束下的应用场景。鉴于 SSS 图像数据集的规模与特点，本研究选择计算量最小、速度最快的 YOLOv8n 模型作为基准模型进行改进。

2.2 基于深度学习的小波变换

小波变换作为一种时频域分析工具，通过在时域与频域同时提供自适应分辨率分析，有效克服了短时傅里叶变换中固定窗口尺寸的固有局限性。其窗口可以根据频率成分进行动态调整 (高频成分对应较窄的时域窗口，低频成分对应较宽的时域窗口)，这一特性使其成为时频信号分析和处理的最佳工具[20]。鉴于其优异的时频局部化能力，自 20 世纪 80 年代发展至今，小波变换已在信号处理、数据分析及医学成像等多个领域得到成功应用。近年来，研究者们开始探索将其多尺度分析能力与深度学习架构结合，以增强神经网络的特征提取与计算效率。例如，Wang 等人[21] 利用小波变换从心电图 (electrocardiogram，ECG) 信号的时频分量中提取多尺度特征；Gal 等人[22] 则将小波变换融入生成对抗网络的生成过程，有效提升了生成图像的质量与模型的计算性能；Finder 等人[23] 通过小波变换对卷积神经网络的特征图进行压缩，显著提升了网络前向传播的效率。作为该方向的进一步深化，该团队后续提出了可学习的 WTConv 层[24]，该层通过将小波变换无缝嵌入标准卷积操作，在不显著增加计算负担的前提下，有效扩大了卷积层的感受野，为构建轻量化的高性能网络提供了一种用户友好的解决方案。本文在骨干网络中引入 WTConv，正是为了利用其频域分析优势，协同增强模型对 SSS 图像低频结构信息与高频细节特征的提取能力，进而从源头提升特征质量，以应对声纳图像信噪比低、目标模糊的挑战。

2.3 特征金字塔

受测量机制和海洋环境的影响，SSS 声纳图像中不同目标的尺度差异较大。而 FPN (feature pyramid networks)[25] 作为一种专为多尺度特征提取而设计的深度学习架构，其核心思想是通过构建融合低级细节信息和高级语义信息的多尺度特征金字塔，来提高各尺度目标检测与分割的性能。图 2 展示了不同形式的 FPN。传统的 FPN 通过引入自上而下的路径将语义信息从高级特征图迁移到低级特征图，实现了有效的多尺度特征融合。然而，这种单向信息流存在局限性，主要是因为浅层特征信息可能在迁移过程中丢失，而自下而上的特征迁移

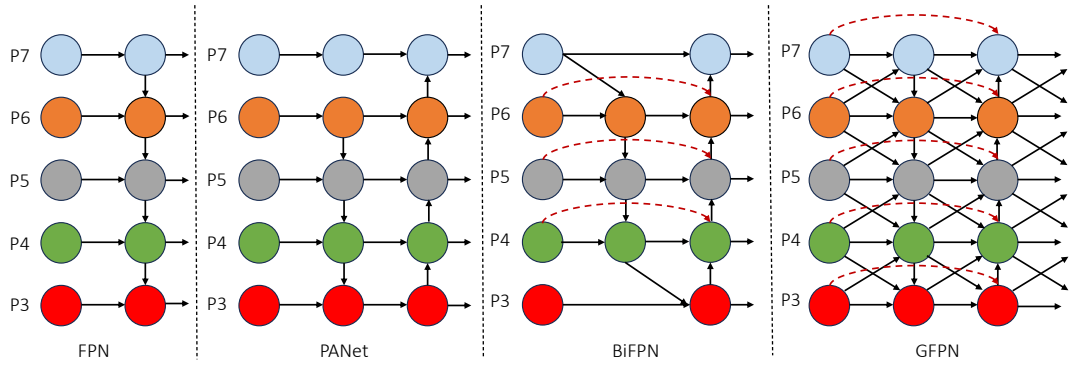


图 2 特征金字塔的不同设计

Fig. 2 Feature pyramid network evolution design

不能得到充分利用。为了解决这个问题，PANet[26] 引入了自下而上和自上而下的双向融合架构，通过加入额外的自下而上的路径聚合网络，大大增强了多尺度特征的融合能力。BiFPN[27] 通过精心设计更加多样化的层间连接以及加入跳跃连接，进一步完善了特征融合机制，从而增强了特征交互和融合效率。然而，BiFPN 中各个模块独立运行，网络过度堆叠可能导致梯度消失问题，对模型训练和整体性能产生不利影响。因此，文献 [28] 提出了 GFPN 网络结构。GFPN 采用 $\log_2 n$ 跳跃层连接，通过 Queen-Fusion 结构促进特征融合，使每个节点不仅接收来自前一个节点的输入，还接收来自该节点上下对角节点的输入。将多尺度特征合并到前一个层和当前层的两层特征中，可以更高效地传递信息，并扩展到更深的网络。但徐等人[29] 通过实验发现基于 GFPN 的模型延迟明显高于基于改进的 PANet 的模型，因此进一步优化 GFPN，提出 Efficient-RepGFPN 以满足实时检测的要求。Efficient-RepGFPN 在 GFPN 的 Queen-Fusion 基础上，引入了重参数化技术和更高效的层间连接策略，显著降低了计算复杂度和推理延迟。为简化表述，本文将基于重参数化技术优化后的 GFPN 结构统一称为 RepGFPN。本文提出的 D-GFPN是在 RepGFPN 的设计理念基础上，进一步引入动态上采样和跨阶段局部 - 双路径轻量聚合网络模块，以更好地适配 SSS 声纳图像的多尺度特征融合需求。

3 算法改进

YOLOv8 作为当前主流的目标检测算法，已经在多个领域被广泛应用。但由于声纳图像成像质量低、纹理特征稀疏等问题，导致该模型优秀的检测性能在此类数据上很难实现有效泛化。针对这一挑战，本文提出了改进模型 YOLOv8-ESI，如图 3 所示。该模型通过引入 WTConv 模块、设计 D-GFPN 颈部网络，以及使用 WIoU 替换 CIoU 等技术手段，既进行了模型轻量化的架构设计，又实现了 SSS 图像目标检测精度的显著提升。

3.1 基于小波变换的 WTConv

在水下目标检测中，SSS 图像普遍存在高噪声干扰、形变失真以及“亮区-阴影”对特征稀疏等问题，标准卷积神经网络难以从图像中学习到判别性强的特征表示，进而导致对物体的检测精度不高。虽然可以通过在骨干网络中堆叠卷积或使用更深的残差网络来增强特征提取能力，但这些方法会引入大量额外参数，显著增加模型的计算负担。为此，本文提出一种面向声纳图像特性的轻量化特征增强方案：将 YOLOv8n 骨干网络中的 C2f 模块替换为小波卷积 (WTConv) 模块。该模块通过多尺度小波分解，将特征提取过程迁移至小波域：一方面，低频子带聚焦于目标整体结构与“亮区-阴影”对等宏观模式，增强模型对几何畸变的鲁棒性；另一方面，高频子带则捕捉边缘、纹理等细节，并利用小波阈值特性抑制高频噪声。这一设计在不显著增加计算开销的前提下，实现了感受野的有效扩大与频域特征的针对性增强，从而在提升检测精度的同时，保持了模型在水下实时场景下的部署可行性。

WTConv 的核心构建基于二维 Haar 小波变换。对于输入特征图 $\boldsymbol{X}$，首先通过步长为 2 的深度卷积，利用一组预先定义的、正交的 Haar 滤波器组对其进行分解。滤波器组成如下：

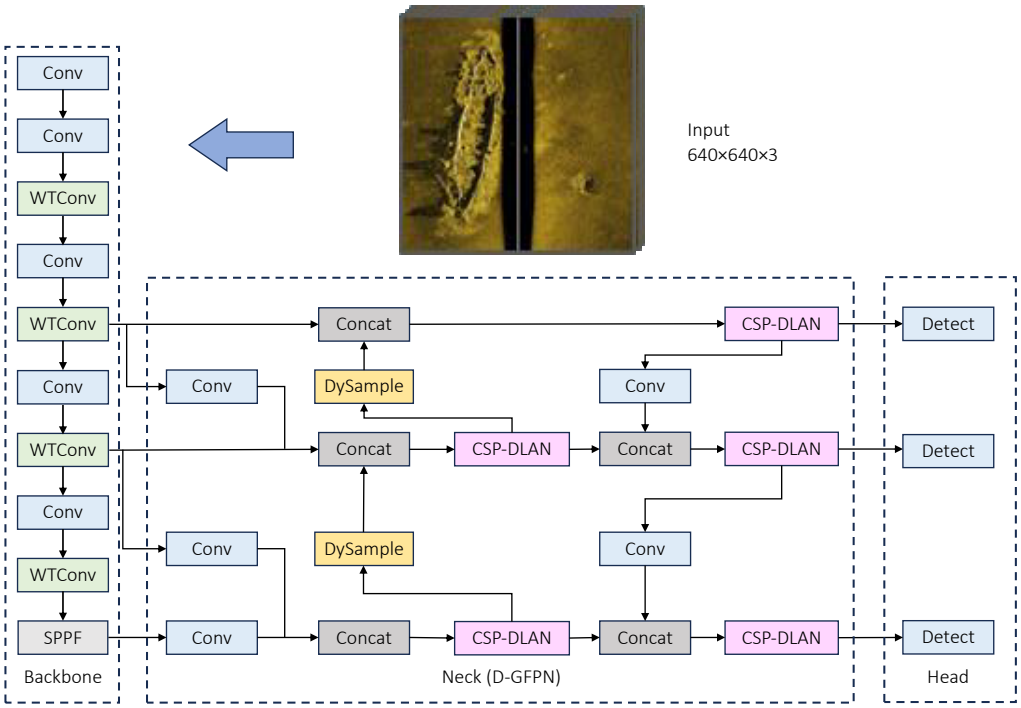


图 3 YOLOv8-ESI 网络结构图

Fig. 3 YOLOv8-ESI network structure

$$f_{\mathrm{LL}}=\frac{1}{2}\begin{bmatrix}1&1\\1&1\end{bmatrix},\quad f_{\mathrm{LH}}=\frac{1}{2}\begin{bmatrix}1&-1\\1&-1\end{bmatrix},$$
$$f_{\mathrm{HL}}=\frac{1}{2}\begin{bmatrix}1&1\\-1&-1\end{bmatrix},\quad f_{\mathrm{HH}}=\frac{1}{2}\begin{bmatrix}1&-1\\-1&1\end{bmatrix},\tag{1}$$

其中：f_{LL} 是低通滤波器，捕获图像低频分量；f_{LH}、f_{HL}、f_{HH}是一组高通滤波器，共同捕获图像高频分量。对于每个输入通道卷积的输出：

$$[\boldsymbol{X}_{\mathrm{LL}},\boldsymbol{X}_{\mathrm{LH}},\boldsymbol{X}_{\mathrm{HL}},\boldsymbol{X}_{\mathrm{HH}}]=Conv([f_{\mathrm{LL}},f_{\mathrm{LH}},f_{\mathrm{HL}},f_{\mathrm{HH}}],\boldsymbol{X}),\tag{2}$$

其中，$\boldsymbol{X}_{\mathrm{LL}}$是 $\boldsymbol{X}$ 的低频分量，而$\boldsymbol{X}_{\mathrm{LH}}$、$\boldsymbol{X}_{\mathrm{HL}}$、$\boldsymbol{X}_{\mathrm{HH}}$是其水平、垂直和对角高频分量。此过程将 $\boldsymbol{X}$ 在空间域的特征，高效转换至由 LL, LH, HL, HH 四个子带构成的小波域中进行表示，如图 4 所示。

随后对得到的各频率子带 (尤其是下采样后的低频子带) 施加小尺寸深度卷积核 (如 $k\times k$)。由于子带分辨率减半，小卷积核在其上的有效感受野等价于在原始输入上使用一个更大的卷积核。该过程可表述为

$$\boldsymbol{Y}=IWT\left(Conv\left(W,WT\left(\boldsymbol{X}\right)\right)\right),\tag{3}$$

其中：W 是可学习的深度卷积权重，WT 表示二维 Haar 小波变换操作，将空间域特征分解为四个频率子带；$Conv$ 表示深度卷积操作；IWT 表示逆小波变换操作，将处理后的频域特征重建回空间域；$\boldsymbol{Y}$ 为最终输出的特征图。这一设计在显著降低参数量的同时，保障了特征提取的全局感知能力。

为进一步增强多尺度特征提取能力，WTConv 支持对低频分量 $\boldsymbol{X}_{\mathrm{LL}}$ 进行递归级联分解，构建多层次的小波金字塔。基于小波变换的线性性质，不同层级处理后的高频特征可通过逆变换线性叠加，实现跨尺度特征融合：

$$\boldsymbol{Z}^{i}=IWT\left(\boldsymbol{Y}^{i}_{\mathrm{LL}}+\boldsymbol{Z}^{i+1},\boldsymbol{Y}^{i}_{\mathrm{H}}\right),\tag{4}$$

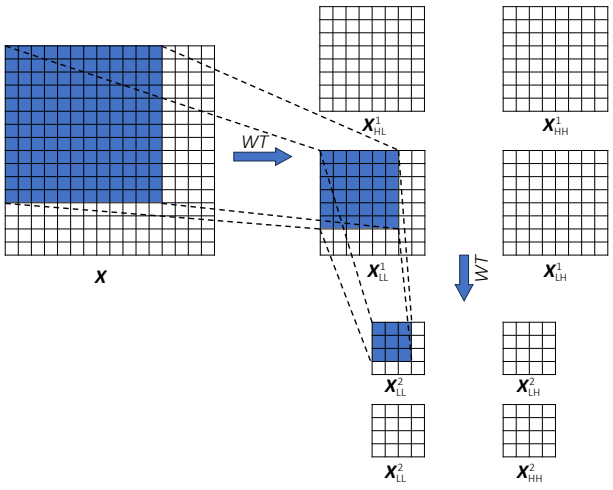


图 4 小波域中的卷积

Fig. 4 Convolution process in wavelet domain

其中：$\boldsymbol{Z}^i$是从级别 i 及以上的聚合输出。该机制使模型能够自适应整合从粗到细的频域信息，进而通过抑制 SSS 图像中的噪声频带、强化有效信号，来提升模型的鲁棒性。本文将该模块集成至 YOLOv8n 骨干网络，在不增加计算负载的前提下，显著提升了模型对 SSS 图像的特征提取能力，实现了检测精度与推理效率的协同优化。

3.2 面向侧扫声纳图像检测的颈部网络 (D-GFPN)

为应对 SSS 图像中目标尺度差异悬殊、细节信息模糊等挑战，并满足水下实时检测对模型轻量化的严苛要求，本文对 YOLOv8n 的颈部网络进行了系统性重构，提出了 D-GFPN，其结构如图 5 所示。YOLOv8 采用的 FPN-PAN 结构虽然通过双向路径融合实现了跨尺度信息传递，但面对分辨率有限、细节模糊的 SSS 图像时，受输入特征尺度影响，不能有效平衡不同尺度间的融合权重，导致多尺度特征利用率低下。因此，本文提出一种新的颈部网络结构 D-GFPN。D-GFPN 以经过重参数化优化的 RepGFPN(即 efficient-RepGFPN) 为基准结构，继承了 GFPN 的相邻层级的跨尺度连接与同尺度下的跨层连接，促进各种层次和尺度信息充分交互，增强目标特征的传递能力。同时，引入动态尺度感知的通道分配机制，根据不同尺度特征图的信息密度自适应配置通道维度，替代原 GFPN 的均一通道策略，实现计算资源的最优分配。并且通过优化跨层连接路径，减少冗余计算量，解决了原 GFPN 带来的过多参数量和计算负担的问题。该设计使得网络能够更有效地融合 SSS 图像中模糊的边缘细节与多尺度目标特征，在显著降低计算复杂度的同时，获得了比原 GFPN 及传统 FPN-PAN 结构更优的特征融合质量，为后续检测头提供了判别力更强的多尺度特征表示。

3.3 动态上采样 (DySample)

特征上采样是目标检测网络中的一项关键操作，旨在恢复特征图的空间分辨率以促进多尺度特征融合。YOLOv8n 采用最近邻插值作为上采样方法，该方法通过简单复制最近的像素值来实现尺寸扩展。因此在处理纹理复杂的 SSS 图像时，容易导致边缘模糊与细节信息丢失，从而削弱特征金字塔的融合质量。为了改善这一局限性，本文引入融合了多层次点采样策略的上采样算子 DySample[30]，为不同尺度的特征图提供更高质量的上采样结果。相比于传统点采样策略，多层次点采样策略在初始阶段采用轻度分裂，之后随着特征图尺度的减小，增加分裂的层次和深度，使其能够在不同尺度下最大程度地捕捉到目标的细节特征，并且保证特征图在多尺度融合过程中的一致性和连贯。

DySample 具体结构如图 6 所示，图中 $\boldsymbol{X}$、$\boldsymbol{X}_2$、$\boldsymbol{O}$、$\boldsymbol{G}$ 和 δ 分别代表输入特征、上采样特征、生成偏移、原始采样网格和 sigmoid 函数。对给定特征图 $\boldsymbol{X}\in\mathbb{R}^{C\times H\times W}$ 与采样集 $\boldsymbol{S}\in\mathbb{R}^{2g\times sH\times sW}$，通过 *grid_sample* 函数，依据 $\boldsymbol{S}$ 中的位置对 $\boldsymbol{X}$ 进行重采样，得到上采样特征 $\boldsymbol{X}_2\in\mathbb{R}^{C\times sH\times sW}$，该过程可表示为

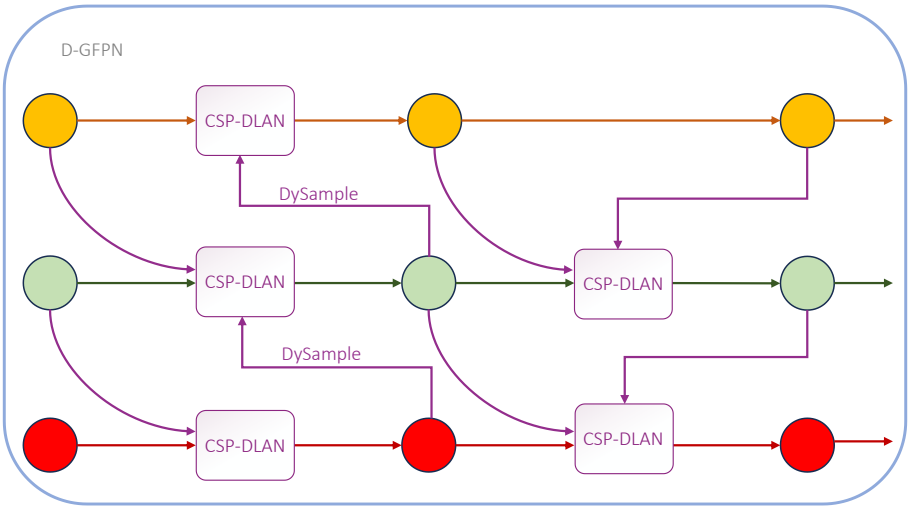


图 5 D-GFPN 网络结构图

Fig. 5 The network structure of D-GFPN

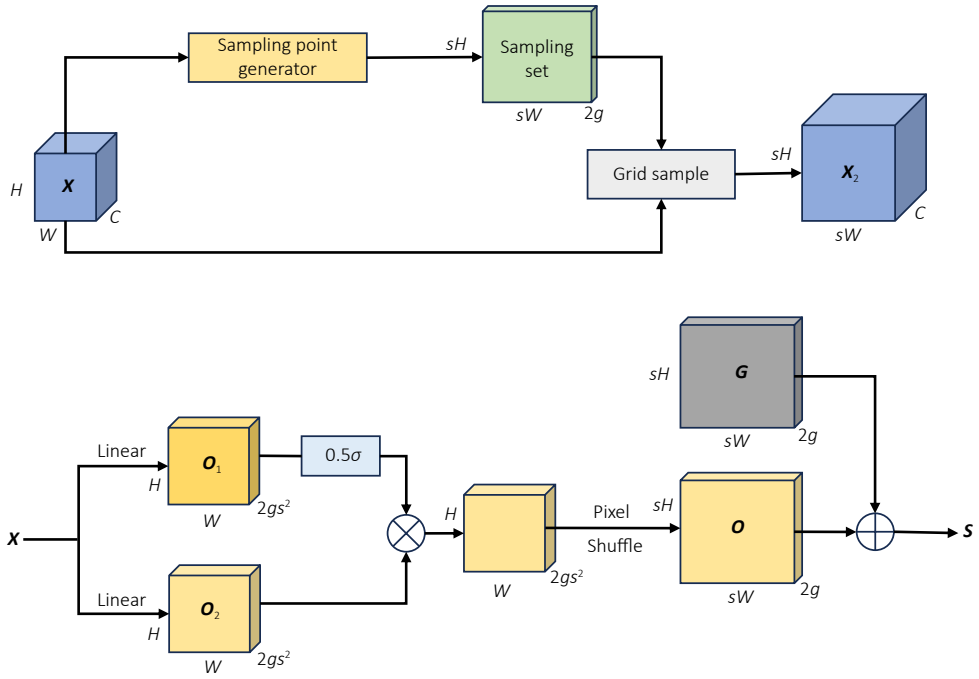


图 6 DySample 模块结构

Fig. 6 Sampling based dynamic upsampling and module designs in DySample

$$\boldsymbol{X}_2 = grid_sample(\boldsymbol{X},\boldsymbol{S})\ . \tag{5}$$

DySample 收到多通道张量信息 $\boldsymbol{X}$ 后，利用两个并行线性变换层输出偏移量 $\boldsymbol{O}_1,\boldsymbol{O}_2 \in \mathbb{R}^{2gs^2\times H\times W}$。然后将 $\boldsymbol{O}_1$ 设置为动态因子 0.5δ，$\boldsymbol{O}_2$ 采用默认采样因子，二者结合后经像素重组，重塑为偏移量 $\boldsymbol{O}\in\mathbb{R}^{2g\times sH\times sW}$。其中偏移范围由静态和动态范围因子所决定。最后，通过结合偏移量 $\boldsymbol{O}$ 和原始数据 $\boldsymbol{G}$ 生成采样集 $\boldsymbol{S}$。该过程的计算表达式如下所示：

$$\boldsymbol{O} = Linear(\boldsymbol{X}), \tag{6}$$

$$\boldsymbol{S} = G + \boldsymbol{O}\ . \tag{7}$$

通过该设计，使上采样过程能够自适应地融合多层次上下文信息，显著提升 SSS 图像在特征融合中的细节保持能力，为检测头提供更丰富的空间特征表示。

3.4 跨阶段局部 - 双路径轻量聚合网络模块 (CSP-DLAN)

在原始的广义特征金字塔网络 (GFPN) 中，特征融合模块依赖于多个 3×3 卷积层。这些卷积层虽然能够有效提取多尺度特征，但同时也带来了显著的参数量和计算负担，难以满足 SSS 图像对低延迟、轻量化检测的工程需求。为此，本文设计了一种轻量化特征融合模块 CSP-DLAN，其架构如图 7 所示。CSP-DLAN 模块采用了通道拆分-处理-融合的拓扑结构。输入特征首先沿通道维度分为两部分：一部分保留原始特征信息，另一部分通过串联的双 DualConv 单元进行深度特征变换。每个 DualConv 单元由并行的 1×1 卷积和 3×3 分组深度可分离卷积构成，二者输出沿通道维度拼接。这种设计既保留了 3×3 卷积捕捉空间细节的能力，又通过 1×1 卷积实现高效的通道交互[31]。为提升组间信息流动性，模块在每次 DualConv 后引入通道重排层，将特征按组进行周期性移位。该操作在不增加参数的前提下，促进了不同组别特征间的交互，有效增强了特征的跨组融合能力。在 SSS 图像检测中，CSP-DLAN 的 3×3 卷积支路可以通过局部感受野来抑制高频随机噪声，1×1 卷积支路可以通过跨通道交互强化低频结构信息，并且分组重排机制使模块能够区分真实目标反射与噪声模式。这些设计有效解决了 SSS 图像检测中特征融合的核心难题，并且有效满足了水下平台对轻量化、实时化检测的工程需求。

3.5 基于动态非单调聚焦机制的损失函数 (WIoU)

在目标检测任务中，损失函数是衡量预测框与真实框之间差异的关键工具，其目的是通过优化这些差异来提高目标检测的精度。最常见的损失函数有 IoU[32]、GIoU[33]、DIoU[34] 等。YOLOv8 n 采用 CIoU，在 IoU 的

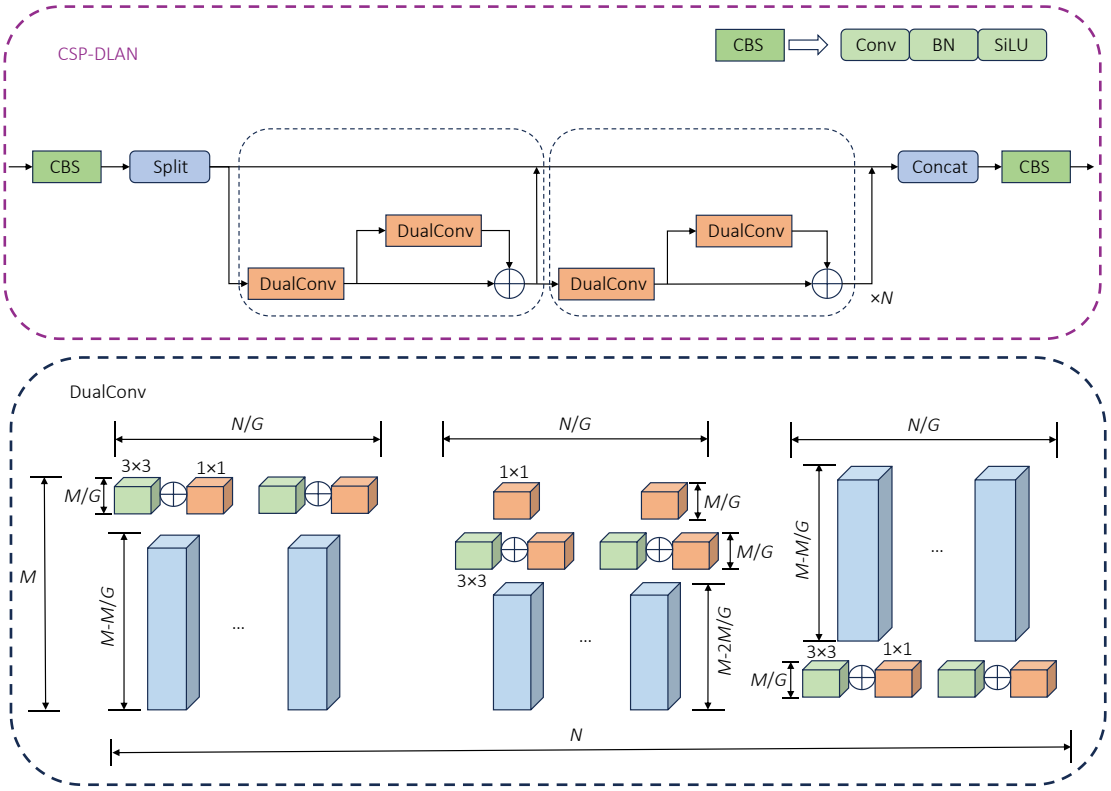


图 7 CSP-DLAN 模块结构图

Fig. 7 Design of CSP-DLAN module integrated with DualConv

基础上进一步引入边界框的中心点距离和宽高比作为优化目标，以提高目标检测的精度。然而，上述所有边界框损失函数都假设训练数据由高质量样本组成，其主要关注点在于优化边界框损失的拟合精度。这一假设与 SSS 图像的标注现实严重不符：受限于图像固有的低分辨率、强噪声及目标模糊性，标注过程中不可避免地引入边界偏移、尺寸误差及部分遗漏等低质量样本。过分强调对这些低质量样本的边界框回归，不仅无法提升模型的检测能力，反而可能降低其检测能力。因此，本文引入具有动态非单调聚焦机制的 WIoU[35]。该机制利用“离群度”评估锚框质量，并在此基础上建立均衡的梯度增益分配策略。WIoU 能够动态降低低质量样本对梯度更新的影响，同时聚焦于中等质量样本的优化，从而在质量不均的 SSS 图像数据集上实现更稳定、更有效的训练，最终提升检测器的整体性能。

WIoU 共有三种形态，具体定义如下：WIoUv1 在基础 IoU 损失中引入距离注意力，旨在当锚框与目标框对齐良好时，有效抑制不必要的几何约束。其构造如下：

$$L_{\mathrm{IoU}} = 1 - IoU = 1 - \frac{W_i H_i}{S_{\mathrm{u}}}, \tag{8}$$

$$S_{\mathrm{u}} = wh + w_{\mathrm{gt}} h_{\mathrm{gt}} - W_i H_i, \tag{9}$$

$$L_{\mathrm{WIoUv1}} = R_{\mathrm{WIoU}} L_{\mathrm{IoU}}, \tag{10}$$

$$R_{\mathrm{WIoU}} = \exp\left(\frac{(x-x_{\mathrm{gt}})^2 + (y-y_{\mathrm{gt}})^2}{\left(W_{\mathrm{g}}^2 + H_{\mathrm{g}}^2\right)^*}\right), \tag{11}$$

式中：W_{g}，H_{g} 是最小边界框尺寸 (图 8)。为了防止 R_{WIoU} 产生阻碍收敛的梯度，W_{g}，H_{g} 与计算图分离 (上标*表示此操作)，消除了阻碍收敛的因素，不引入长宽比等新指标。接着仿照 Focal Loss 为 v1 构造了单调聚焦系数$L_{\mathrm{IoU}}^{\gamma^*}$，从而得到 WIoUv2，公式如下：

$$L_{\mathrm{WIoUv1}} = L_{\mathrm{IoU}}^{\gamma^*} L_{\mathrm{WIoUv1}}, \gamma > 0, \tag{12}$$

$$\frac{\partial L_{\mathrm{WIoUv2}}}{\partial L_{\mathrm{IoU}}} = L_{\mathrm{IoU}}^{\gamma^*} \frac{\partial L_{\mathrm{WIoUv1}}}{\partial L_{\mathrm{IoU}}}, \gamma > 0, \tag{13}$$

$$L_{\mathrm{WIoUv2}} = \left(\frac{L_{\mathrm{IoU}}^*}{\overline{L_{\mathrm{IoU}}}}\right)^{\gamma} L_{\mathrm{WIoUv1}}. \tag{14}$$

锚框的离群度用β表示，使用β构造一个非单调聚焦系数并将其应用到 WIoUv1 得到 WIoUv3。

$$\beta = \frac{L_{\mathrm{IoU}}^*}{\overline{L_{\mathrm{IoU}}}} \in [0, +\infty), \tag{15}$$

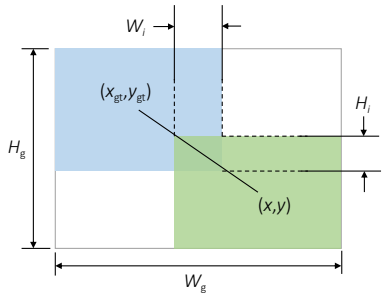


图 8 IoU 示意图

Fig. 8 IoU example diagram

$$L_{\mathrm{WIoUv3}} = rL_{\mathrm{WIoUv1}},\ r = \frac{\beta}{\delta\alpha^{\beta-\delta}}. \tag{16}$$

4 实验与结果分析

4.1 实验环境

为验证本文所提出方法的可行性和有效性，进行进一步实验，具体的实验环境硬件与软件配置如表 1 所示，模型训练所使用的关键超参数设置如表 2 所示。

表 1 实验环境配置

Table 1 Experimental environment configuration

环境名	配置
CPU	Intel Core i5-13600KF
GPU	NVIDIA GeForce RTX 4070 SUPER，12GB
Pytorch	2.4.1
Python	3.8.20
Cuda	12.6
os	Windows 11

表 2 网络模型训练参数

Table 2 Network model training parameters

参数名称	设置值
优化器	SGD
初始学习率	0.01
动量	0.937
权重衰减	0.0005
批次大小	16
训练轮次	200
学习率调度	Cosine Annealing

4.2 数据集选择

鉴于 SSS 图像数据的稀缺性，单一公开数据集样本量有限，易导致实验出现过拟合或欠拟合现象。为解决这一问题，本文整合了开源数据集 Seabed Objects-KLSG 和 SONAR-2019，并补充了通过水下实验测量以及网络公开的高质量 SSS 图像，构建了自建数据集 SSSD (side-scan sonar dataset)。该数据集图像分辨率主要分布在 300 pixel×300 pixel 至 800 pixel×500 pixel 之间，包含沉船、飞机残骸和人类遗体三个主要检测类别，其中沉船图像 602 张，飞机残骸图像 279 张，人类遗体图像 75 张，共计 956 张原始图像。数据集按照 7∶2∶1 的比例划分为训练集、验证集和测试集。为防止模型训练过程中出现过拟合现象并提升其泛化能力，本文对原始训练集采用了多种数据增强操作进行样本扩充。具体增强手段包括：水平翻转、随机旋转、随机亮度，以及添加高斯噪声等。通过引入这些几何与光度变换，有效模拟了水下成像的多种不确定性，丰富了训练样本的多样性。数据增强后，进一步剔除了部分变化异常的图像，最终训练集包含 2926 张沉船图像。图 9 展示了经过操作后的图像和原始的图像对比，第一列为原始图像，第 2~5 列为经过操作后的图像。

4.3 模型评估指标

为了验证改进模型的有效性，本文从精确率、召回率、平均精度均值、 GFLOPs、参数量以及推理时间几大指标来客观地评价改进后模型的性能。从检测精度与模型轻量化程度两个核心维度进行综合评估，各大指标计算公式如下：

$$P = \frac{TP}{TP+FP}, \tag{17}$$

$$R = \frac{TP}{TP+FN}, \tag{18}$$

$$AP = \int_0^1 p(r)\mathrm{d}r, \tag{19}$$

$$mAP = \frac{1}{n}\sum_{i=1}^{n} AP(i). \tag{20}$$

上述各评估指标中，*TP* (true positive) 表示正确检测的正样本数量；*FP* (false positive) 表示误检为正样本的负样本数量；*FN* (false negative) 表示未被检测出的正样本数量，即被模型漏检的真实目标实例数。在此基础上，精确率 *P* 表示预测为正类的样本中实际为正类的条件概率，反映了预测结果的可靠性；召回率 *R* 表示模型对真实正类样本的覆盖能力，表征检测的完备性；AP (average precision) 为某一类别下 *P*-*R* 曲线与坐标轴围成的面积，衡量模型在该类别上的综合检测性能；mAP (mean average precision) 则为所有类别 AP 值的平均值，是

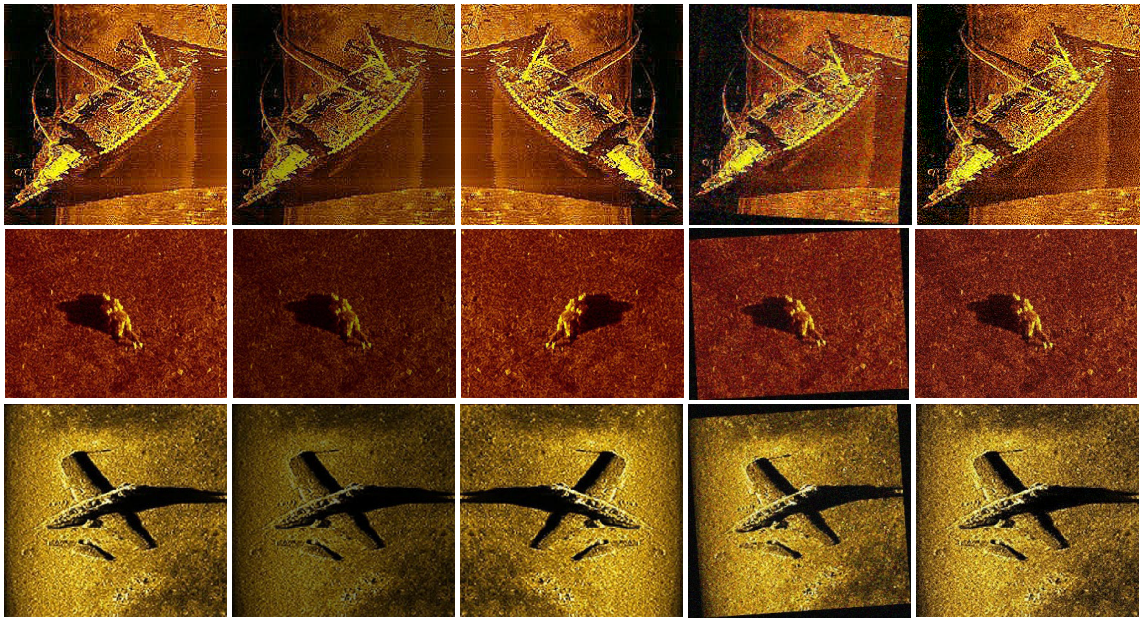


图 9 数据增强前后效果对比示例

Fig. 9 Examples of data augmentation effects

衡量模型在多类别目标检测任务中整体性能的核心指标。

4.4 消融实验

为系统评估 YOLOv8-ESI 模型中各改进模块的贡献，本研究设计并进行了一系列消融实验。实验以未做任何改动的原始 YOLOv8n 模型作为基线，采用逐步累加的验证策略，依次引入 WTConv 模块、D-GFPN 颈部网络以及 WIoU 损失函数，以量化分析各项技术对最终检测性能的影响。实验结果汇总于表 3。图 10 展示了各模块独立引入及组合使用的性能对比。

实验结果表明，本文提出的三大部分改进在不同程度上提高了网络性能。具体而言，通过独立引入各改进模块，模型的检测精度均获得明显提升，这验证了各模块在增强特征提取能力、优化信息融合机制以及提高模型稳定性等方面的有效性。更为重要的是，当多个模块协同工作时，模型性能呈现出更为显著的提升，这表明各模块之间相辅相成，能够有效增强模型对不同尺度特征和水下复杂场景的适应能力。在全面整合所有改进策略后，YOLOv8-ESI 模型展现出最优的性能表现。相较基线模型，mAP50 指标提升了 5.2%，mAP50-95 指标提高了 3.8%，这一结果充分证明了改进后的模型在多尺度目标检测任务中具有更强的适应性和检测精度。值得注意的是，在性能提升的同时，改进后的模型参数量减少了 10%，计算量降低了 1.1 GFLOPs，这一结果表明本研究在模型轻量化方面也取得了显著成效。

后文对实验结果进行了更深入的定性和定量分析，重点介绍了每个模块的具体功能，讨论了每个模块如何通过不同的机制影响网络的性能，从而进一步揭示了这

表 3 消融实验

Table 3 Ablation experiment

YOLOv8n	WTConv	D-GFPN	WIoU	*P*/%	*R*/%	Params/M	GFLOPs	mAP50/%	mAP50-95/%
√				84.1	72.6	3.01	8.1	83.6	59.6
√	√			84.1	80.4	2.71	7.3	86.2	62.1
√		√		85.6	78.4	2.98	7.8	87.3	62.2
√			√	86.9	79.2	3.01	8.1	84.7	59.5
√		√	√	82.1	77.2	2.98	7.8	85.7	60.8
√	√		√	84.1	80.4	2.71	7.3	86.2	63.1
√	√	√		85.7	81.1	**2.69**	**7.0**	87.7	62.3
√	√	√	√	**88.8**	**84.3**	**2.69**	**7.0**	**88.8**	**63.4**

些模块在 YOLOv8-ESI 中的关键作用。

4.4.1 WTConv 模块

消融实验结果显示，引入 WTConv 模块使模型的 mAP50 和 mAP50-95 分别提高了 2.6% 和 2.5%。验证了该模块在提高检测准确性方面的有效性。为了深入分析 WTConv 模块对模型特征聚焦机制的影响，本研究选取了测试集中不同类别的 SSS 图像 (如沉船、飞机残骸等) 进行热图可视化对比。图 11(b) 和图 11(c) 比较了引入 WTConv 前后的热力图。图中网络对区域的关注度从高到低分别显示为红色区域、黄色区域，以及蓝色区域。如图 11 所示，相较于原始 YOLOv8 模型，引入 WTConv 的改进模型在注意力分布上表现出质的飞跃。传统模型的热图响应呈现明显的弥散特征，目标区域 (红色) 与背景 (黄色/蓝色) 边界模糊，尤其在声纳图像特有的高噪声区域出现大量虚假激活，这直接导致检测过程中出现误报和漏检现象。而采用 WTConv 架构的模型则展现出精确的特征选择能力，其热图红色区域严格限定在目标真实轮廓范围内，背景噪声得到有效抑制，几乎完全呈现低响应状态的蓝色。这种改善源于 WTConv 独特的多尺度分析机制：低频子带 (LL) 专注于增强目标的整体

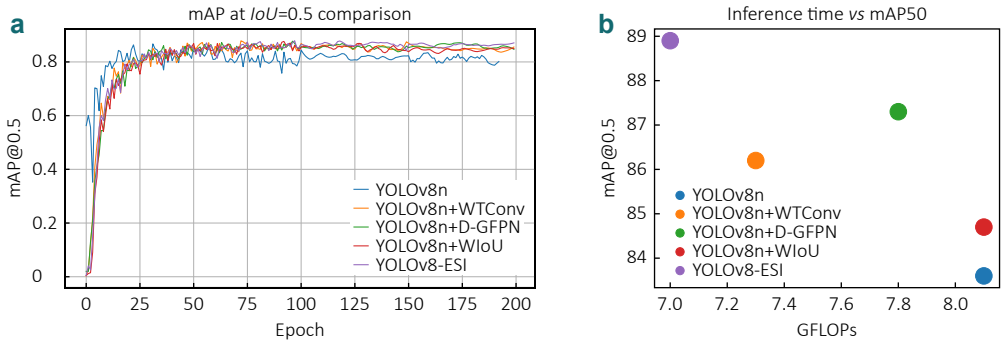


图 10 消融实验对比图。(a) 各改进模块对 mAP50 性能的影响；(b) 各改进模块对模型计算量的影响

Fig. 10 Comparison of ablation study results. (a) Impact of each improved module on mAP50 performance; (b) Impact of each improved module on model computational complexity

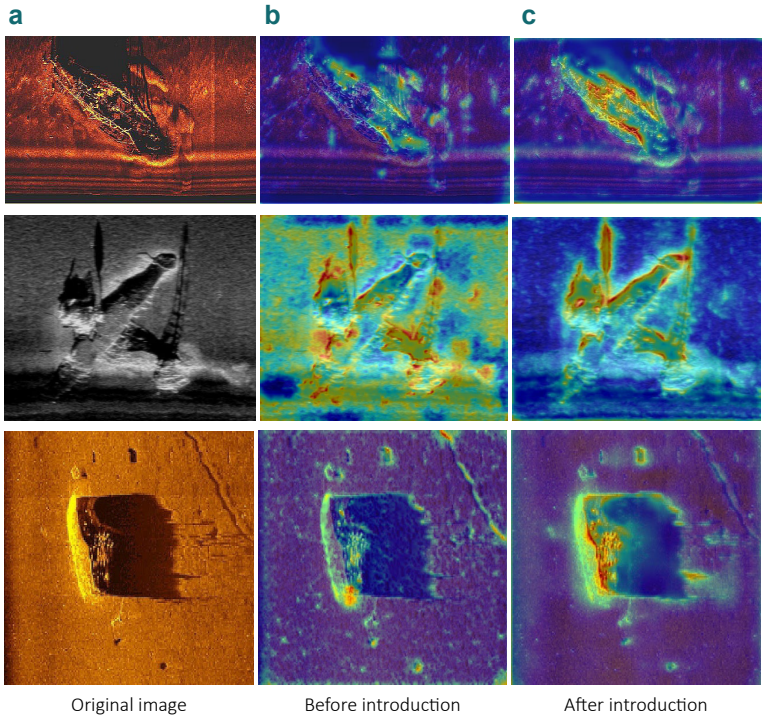


图 11 热图可视化对比。(a) 原始图像；(b) YOLOv8n 热力图；(c) YOLOv8n+WTConv 热力图

Fig. 11 Comparison of heatmap visualizations. (a) Original images; (b) Heatmaps of YOLOv8n; (c) Heatmaps of YOLOv8n+WTConv

结构表征，使模型对声纳目标的宏观特征保持稳定感知；同时高频子带 (HH) 通过自适应阈值处理，有效过滤了由水下环境引入的随机噪声。因此，与初始 YOLOv8n 模型相比，引入 WTConv 可以使模型注意力更精准地聚焦于有效特征区域，从而提升检测精度并降低误检率，从而实现更精确的目标定位。

4.4.2 D-GFPN

为了评估颈部网络 D-GFPN 中各种模块的贡献，本文也进行了一系列的消融实验。从表 4 实验结果可得，相较于基线模型，引入 RepGFPN 的特征金字塔网络结构 (基于重参数化优化后的 GFPN 结构)，动态上采样方式 DySample，以及跨阶段局部 - 双路径轻量聚合网络模块 CSP-DLAN，都对网络的性能带来了一定的提升效果。并且整合所有改进策略后得到的颈部网络展现出了最优的性能表现，在保持最低的参数和计算量的同时，让模型实现了最高的检测性能。

为了进一步验证 D-GFPN 颈部结构的优越性，本文选择了几种主流模型的颈部结构进行比较实验，如表 5 所示。图 12 直观展示了不同颈部网络在检测精度与模型规模之间的权衡关系，PANet 结构是 YOLOv8n 的原始颈部结构，相比之下， D-GFPN 在参数数量和浮点计算量均有降低的情况下，准确率提高了 1.6 %。HS-FPN 是一种基于不同尺度级别的特征金字塔，虽然结构参数较小，但其精度比 D-GFPN 低 1.2%。对比 RepGFPN 颈部网络 (包含原网络特征融合模块)，D-GFPN 虽然 mAP50 相同，但是参数数量和计算的浮点数分别比 RepGFPN 颈部网络降低了 9.1% 和 8.3%，并且 mAP50-95 提高了 2%。因此，实验证明了 SSS 图像目标检测任务中，D-GFPN 优于常见的颈部结构。以上结果充分证明，D-GFPN 通过其动态通道分配与优化连接设计，能够更高效地融合 SSS 图像中尺度差异大、细节模糊的特征，在提升检测精度的同时更好地满足了水下实时检测的轻量化需求。

4.4.3 WIoU

WIoU 具有三种不同的形态，为了验证哪种形态更加适用于 SSS 图像目标检测任务，本文进行了一组对比实验。在基线模型 (集成 WTConv 模块和 D-GFPN 颈部网络架构) 的基础上，通过控制变量实验对比分析了 CIoU、WIoUv1、WIoUv2 和 WIoUv3 四种边界框回归损失函数的检测效能如图 13 所示。表 6 的实验结果表明，采用 WIoUv3 的 YOLOv8-ESI 模型展现出最优的检测性能，其 mAP50 和 mAP50-95 分别达到 88.8% 和 63.4%，较基线模型均提升 1.1% 。同时，从模型对各类别目标的检测结果可以看出，WIoUv3 对人类遗体的检测准确度提升较多，达到了 1.7%。不难分析，受限于声学成像分辨率和复杂水下环境，人体目标在 SSS 图像中通常呈现小尺寸特征，同时伴随显著的噪声干扰和边界模糊效应，导致数据集标注过程中不可避免地引入低质量样本。而 WIoUv3 利用“离群度” (outlier degree) 度量机制有效降低了低质量样本产生的不利梯度影响。各类别的检测实验也进一步表明，WIoUv3 十分适用于声纳图像目标检测任务。

4.5 不同算法模型对比实验

在验证本文所提算法性能的过程中，为确保实验结果的客观性和可比性，本研究在统一的实验环境及数

表 4 颈部网络消融实验

Table 4 Ablation study on the neck network

Neck	Params/M	GFLOPs	mAP50/%	mAP50-95/%
YOLOv8n	3.01	8.1	83.6	59.6
+ RepGFPN	3.32	8.4	87.1	62.2
+ RepGFPN + CSP-DLAN	2.98	7.9	85.3	60.8
+ DySample	3.02	8.2	84.7	60.1
+RepGFPN+DySample+CSP-DLAN	2.98	7.9	87.3	62.2

表 5 颈部网络对比实验

Table 5 Comparative experiments on the neck network

Neck	Params/M	GFLOPs	mAP50/%	mAP50-95/%
YOLOv8n	3.01	8.1	83.6	59.6
BiFPN	2.94	8.1	86.5	61.1
HS-FPN	1.93	6.9	86.1	60.8
RepGFPN	3.29	8.5	87.3	60.2
D-GFPN (Ours)	2.98	7.8	87.3	62.2

据集条件下，与 YOLO 系列目标检测算法进行了对比实验。所选取的对比算法涵盖了 YOLO 系列的多个版本，具体包括 YOLOv3 -tiny、YOLOv5s、YOLOv6n、YOLOv8n、YOLOv9c、YOLOv10n、YOLOv11n。此外，本研究还选取了其他几个主流算法进行比较，如 Faster-RCNN、Cascade-RCNN、RT-DETR、SCR-YOLOv8[36]、CSC-YOLO[37]，比较结果如表 7 所示。

由表 7 可知，Faster RCNN 和 Cascade-RCNN 不仅参数量大，模型复杂而且在该数据集上的适应性欠佳，检测精度较差。而 YOLO 系列和 RT-DETR 算法在参数

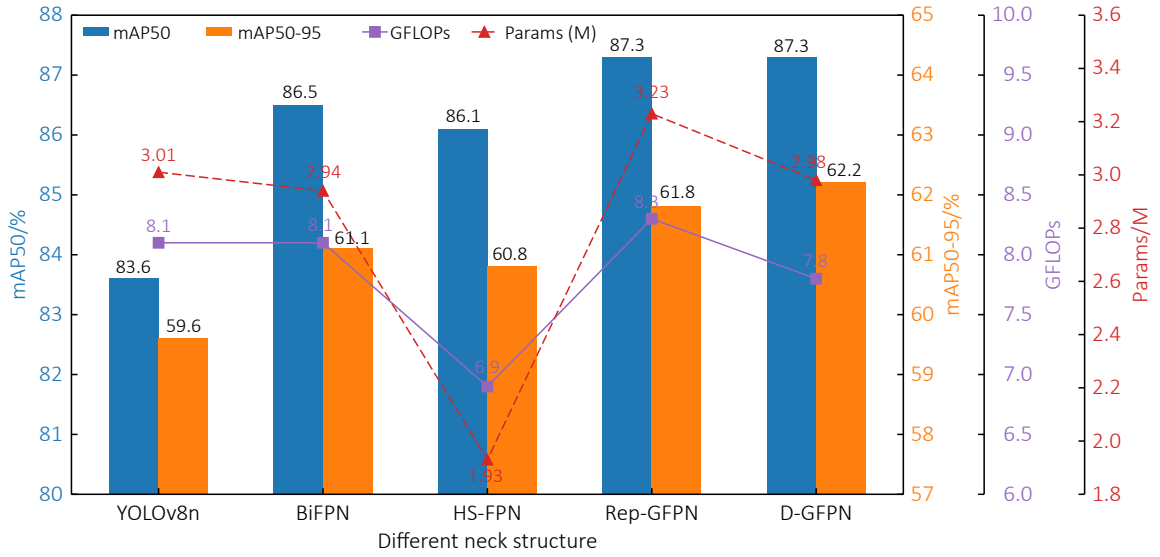


图 12 不同颈部网络对检测精度和模型大小的影响

Fig. 12 Impact of different neck networks on detection accuracy and model size

表 6 边界框回归损失函数对比实验 (单位：%)

Table 6 Comparative experiments on bounding box regression loss functions (Unit: %)

Neck	AP			mAP50	mAP50-95
	Ship	Plane	Body		
CIoU	95.1	87.8	80.2	87.7	62.3
WIoUv1	95.4	88.2	81.3	88.3	62.1
WIoUv2	94	89.2	80.6	87.9	63.3
WIoUv3	96.4	88.1	81.9	88.8	63.4

表 7 对比实验

Table 7 Comparative experimental results

Model	*P*/%	*R*/%	Params/M	GFLOPs	mAP50/%	mAP50-95/%	Inference time/ms
Faster-RCNN	85.0	75.0	41.39	208	82.1	60.1	2.3
Cascade-RCNN	83.7	79.6	69.3	236	82.6	60.3	2.4
YOLOv3-tiny	81.2	70.2	12.13	18.9	80.5	53.0	2.8
YOLOv5s	83.8	74.2	7.02	30.39	79.0	56.8	2.3
YOLOv6n	88.1	79.3	4.34	11.4	86.3	62.2	1.9
YOLOv8n	84.1	72.6	3.01	8.1	83.6	59.6	1.2
YOLOv9c	82.5	81.6	25.5	103.1	86.4	61.6	5.8
YOLOv10n	85.4	75.6	2.71	8.4	85.8	61.7	1.6
YOLOv11n	86.8	78.1	2.59	6.4	87.3	62.2	1.7
RT-DETR	87.3	80.1	19.8	56.9	86.6	61.4	1.2
SCR-YOLOv8	88.6	78.4	1.80	6.2	86.5	60.5	1.7
CSC-YOLO	87.3	80.1	2.40	6.3	86.6	61.4	1.5
YOLOv8-ESI (Ours)	88.8	84.3	2.69	7.0	88.8	63.4	1.3

数量和计算效率方面具有一定优势，但在检测精度方面仍有提升空间。相比之下，YOLOv8-ESI 算法在各项参数上均优于上述算法。同时，与新提出的 SSS 图像检测模型 (SCR-YOLOv8，CSC-YOLO) 相比，虽然参数数量和计算量略有增加，但在准确率、召回率、mAP 等检测精度指标上也获得了显著的提升。图 14 中的气泡图进一步可视化了每个网络的性能，其中靠近左上角的圆圈表示更好的准确性和速度。YOLOv8-ESI 位于左上角，由红色圆圈表示，展示了其在检测精度、推理速度和轻量化设计方面的整体优势。综合评测表明，YOLOv8-ESI 算法兼具高性能和高效率，在保持较低参数数量和计算开销的同时取得了最佳的检测精度，能够很好地满足现代工业应用的需求。

为验证改进模型在真实检测场景中的性能表现，本文从测试集中选取部分 SSS 图像进行检测实验，结果如图 15 所示。从左至右依次为原图、YOLO 各系列模型

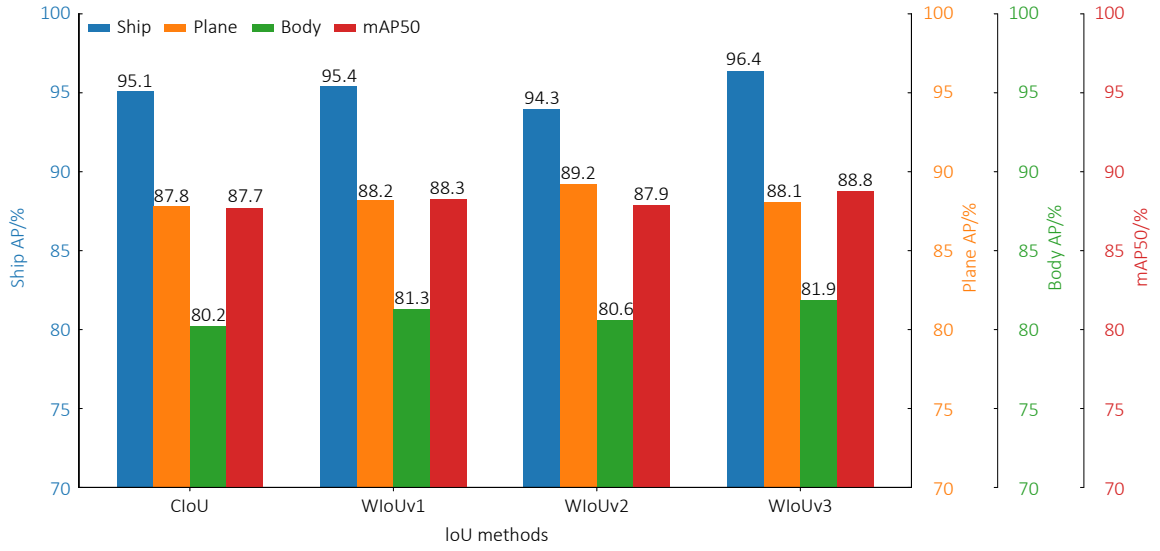


图 13 不同边界框损失函数对检测精度的影响

Fig. 13 Impact of different bounding box loss functions on detection accuracy

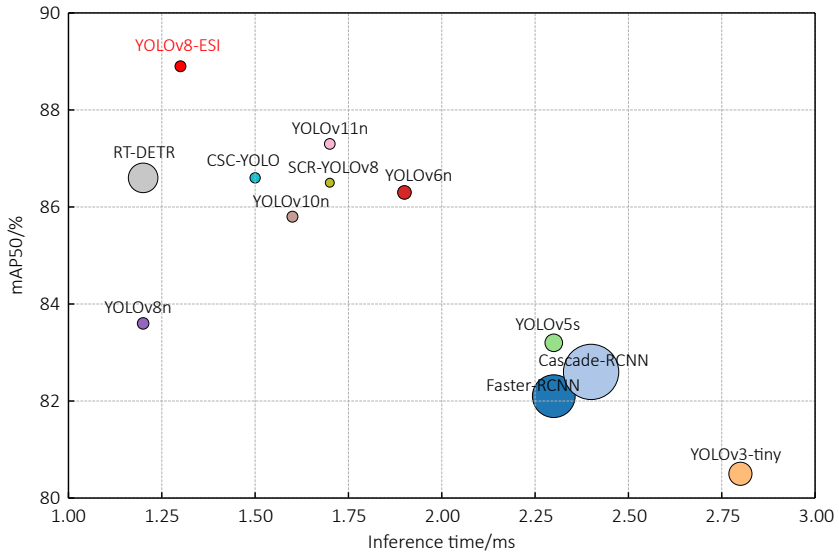


图 14 YOLOv8-ESI 与其他主流网络的视觉比较 (*x* 轴表示推理时间 (ms)，*y* 轴表示 mAP50 (%)，圆半径表示参数数量。YOLOv8-ESI 以红色气泡和字体突出显示)

Fig. 14 Visual comparison between YOLOv8-ESI and other mainstream networks (the *x*-axis represents the inference time (ms), the *y*-axis represents mAP50 (%), and the radius of the circle indicates the number of parameters. YOLOv8-ESI is highlighted in red bubbles and font)

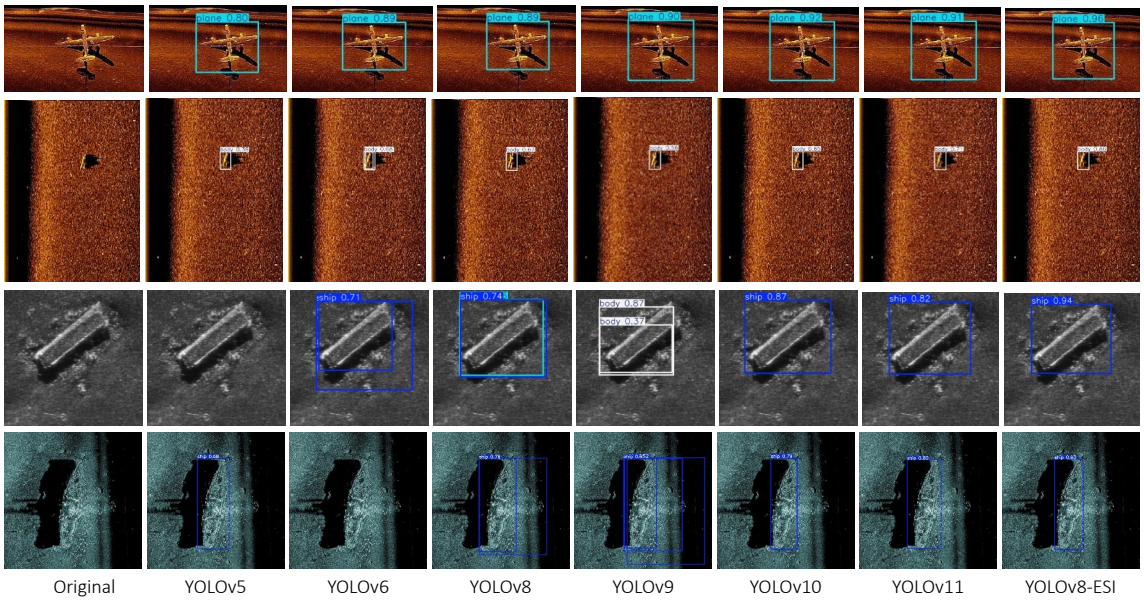


图 15 YOLO 系列和 YOLOv8-ESI 检测结果比较

Fig. 15 Comparison of detection results between YOLO series models and YOLOv8-ESI

的检测结果图以及改进模型 YOLOv8-ESI 的检测结果图。通过对比分析检测效果可以发现，尽管原始 YOLOv8n 网络能够在复杂的海床背景下成功检测出沉船、飞机残骸以及人类遗体等目标，但在检测过程中仍存在一定的误检情况。相比之下，本文提出的改进算法 YOLOv8-ESI 在检测过程中显著减少了误检现象的发生，并且实现了更高的目标边界检测精度。这一结果进一步验证了改进算法在复杂背景下的鲁棒性和检测精度提升的有效性，表明其在实际应用中具有更强的适应性和优越性。

5 结　论

水下环境复杂且能见度低，SSS 采集的图像通常存在噪声高、目标特征模糊及数据标注质量不一等问题。针对这些问题，本文提出了一种改进的轻量化检测模型 YOLOv8-ESI 模型。该模型通过三方面协同创新实现对 SSS 图像特性的精准适配。首先在骨干网络部分引入小波卷积层 WTConv，通过多尺度频域分解协同增强对图像低频结构信息与高频细节的提取能力，在抑制噪声的同时强化有效特征，从源头提升特征质量。其次，设计动态特征金字塔网络 (D-GFPN) 作为颈部结构，优化了跨尺度连接，融合动态上采样器 (DySample) 与轻量化特征融合模块 (CSP-DLAN)，实现了对声纳图像多尺度模糊特征的自适应、高效率融合。最后利用动态非单调聚焦 WIoU 的边界框损失函数替换 CIoU，减小了 SSS 图像数据集低质量示例产生的有害梯度，在加快网络收敛的同时，提高网络检测的精度。实验表明，与基准模型相比，本文所提出方法 mAP50 提升了 5.2%，参数量减少了 10%，计算量降低了 1.1GFLOPs。在大幅提高检测精度的同时还实现了模型的轻量化设计，适配水下设备的实时检测需求，为水下目标检测提供了更有效的技术支持。

作者贡献

谷剑峰：研究资料的收集、方法设计，分析，论文写作；刘锡祥，徐浩：指导论文选题、论文修改。

利益冲突

所有作者声明无利益冲突

参考文献

1. Williams D P. Fast target detection in synthetic aperture sonar imagery: a new algorithm and large-scale performance analysis[J]. *IEEE J Oceanic Eng*, 2015, **40**(1): 71−92.
2. Chen Q, Tian J, Huang H N, et al. Study on SAS image segmentation using SVM based on statistical and texture features[J]. *Chin J Sci Instrum*, 2013, **34**(6): 1413−1420.
陈强, 田杰, 黄海宁, 等. 基于统计和纹理特征的 SAS 图像 SVM 分割研究[J]. 仪器仪表学报, 2013, **34**(6): 1413−1420.
3. Yang F L, Du Z X, Wu Z Y, et al. Object recognizing on sonar image based on histogram and geometric feature[J]. *Mar Sci Bull*, 2006, **25**(5): 64−69.

阳凡林, 独知行, 吴自银, 等. 基于灰度直方图和几何特征的声纳图像目标识别[J]. 海洋通报, 2006, **25**(5): 64−69.

4. Luo J H, Jiang J P, Zhu P M. Automatic extraction of the side-scan sonar imagery outlines based on mathematical morphology[J]. *Haiyang Xuebao*, 2016, **38**(5): 150−157.
罗进华, 蒋锦朋, 朱培民. 基于数学形态学的侧扫声呐图像轮廓自动提取[J]. 海洋学报, 2016, **38**(5): 150−157.
5. Cui J, Hu C Q, Xu H D. Moving target detection for multi-beam forward-looking sonar based on frame-difference method[J]. *Chin J Sci Instrum*, 2018, **39**(2): 169−176.
崔杰, 胡长青, 徐海东. 基于帧差法的多波束前视声呐运动目标检测[J]. 仪器仪表学报, 2018, **39**(2): 169−176.
6. Cui J, Hu C Q, Xu H D, et al. Application of improved Mean Shift algorithm in moving target tracking of forward-looking sonar[J]. *Tech Acoust*, 2020, **39**(3): 279−283.
崔杰, 胡长青, 徐海东, 等. 改进的 Mean Shift 算法在前视声呐运动目标跟踪中的应用[J]. 声学技术, 2020, **39**(3): 279−283.
7. 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*, 2014: 580–587. https://doi.org/10.1109/CVPR.2014.81.
8. Girshick R. Fast R-CNN[C]//*2015 IEEE International Conference on Computer Vision (ICCV)*, 2015: 1440–1448. https://doi.org/10.1109/ICCV.2015.169.
9. Redmon J, Divvala S, Girshick R, et al. You only look once: unified, real-time object detection[C]//*Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 779–788. https://doi.org/10.1109/CVPR.2016.91.
10. Liu W, Anguelov D, Erhan D, et al. SSD: single shot MultiBox detector[C]//*Proceedings of 14th European Conference on Computer Vision – ECCV 2016*, 2016: 21–37. https://doi.org/10.1007/978-3-319-46448-0_2.
11. Zheng H W, Wang F, Gao J B. DES-YOLO: a more accurate object detection method[J]. *Opto-Electron Eng*, 2024, **51**(11): 240212.
郑华伟, 王飞, 高建邦. DES-YOLO: 一种更精确的目标检测方法[J]. 光电工程, 2024, **51**(11): 240212.
12. Zhao D D, Xie D H, Chen P, et al. Lightweight YOLOv5 sonar image object detection algorithm and implementation based on ZYNQ[J]. *Opto-Electron Eng*, 2024, **51**(1): 230284.
赵冬冬, 谢墩翰, 陈朋, 等. 基于 ZYNQ 的轻量化 YOLOv5 声呐图像目标检测算法及实现[J]. 光电工程, 2024, **51**(1): 230284.
13. Wang Y M, Liu J X, Yu S Q, et al. Underwater object detection based on YOLO-v3 network[C]//*2021 IEEE International Conference on Unmanned Systems (ICUS)*, 2021: 571–575. https://doi.org/10.1109/ICUS52573.2021.9641489.
14. Gao Y, Li Z Y, Zhang K Y, et al. GCP-YOLO: a lightweight underwater object detection model based on YOLOv7[J]. *J Real-Time Image Proc*, 2025, **22**(1): 3.
15. Cao J X, Chen Q, Guo J, et al. Attention-guided context feature pyramid network for object detection[EB/OL]. arXiv: 2005.11475, 2020. https://doi.org/10.48550/arXiv.2005.11475.
16. Wan D H, Lu R S, Hu B T, et al. YOLO-MIF: improved YOLOv8 with multi-information fusion for object detection in gray-scale images[J]. *Adv. Eng. Inform*, 2024, **62**: 102709.
17. Zheng L H, Hu T, Zhu J. Underwater sonar target detection based on improved ScEMA-YOLOv8[J]. *IEEE Geosci Remote Sens Lett*, 2024, **21**: 1503505.
18. Liu K Y, Sun Q, Sun D M, et al. Underwater target detection based on improved YOLOv7[J]. *J Mar Sci Eng*, 2023, **11**(3): 677.
19. Sun Y, Zheng W X, Du X, et al. Underwater small target detection based on YOLOX combined with MobileViT and double coordinate attention[J]. *J Mar Sci Eng*, 2023, **11**(6): 1178.
20. Daubechies I. *Ten Lectures on Wavelets*[M]. Philadelphia: Society for Industrial and Applied Mathematics, 1992. https://doi.org/10.1137/1.9781611970104.
21. Wang T, Lu C H, Sun Y N, et al. Automatic ECG classification using continuous wavelet transform and convolutional neural network[J]. *Entropy*, 2021, **23**(1): 119.
22. Gal R, Hochberg D C, Bermano A, et al. SWAGAN: a style-based wavelet-driven generative model[J]. *ACM Trans Graph (TOG)*, 2021, **40**(4): 134.
23. Finder S E, Zohav Y, Ashkenazi M, et al. Wavelet feature maps compression for image-to-image CNNs[C]//*Proceedings of the 36th International Conference on Neural Information Processing Systems*, 2022, **35**: 1497.
24. Finder S E, Amoyal R, Treister E, et al. Wavelet convolutions for large receptive fields[C]//*Proceedings of 18th European Conference on Computer Vision – ECCV 2024*, 2024: 363–380. https://doi.org/10.1007/978-3-031-72949-2_21.
25. Ghiasi G, Lin T Y, Le Q V. NAS-FPN: learning scalable feature pyramid architecture for object detection[C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019: 7029–7038. https://doi.org/10.1109/CVPR.2019.00720.
26. Wang K X, Liew J H, Zou Y T, et al. PANet: few-shot image semantic segmentation with prototype alignment[C]//*Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019: 9196–9205. https://doi.org/10.1109/ICCV.2019.00929.
27. Tan M X, Pang R M, Le Q V. EfficientDet: scalable and efficient object detection[C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 10778–10787. https://doi.org/10.1109/CVPR42600.2020.01079.
28. Jiang Y Q, Tan Z Y, Wang J Y, et al. GiraffeDet: a heavy-neck paradigm for object detection[C]//*Proceedings of the 10th International Conference on Learning Representations*, 2022.
29. Xu X Z, Jiang Y Q, Chen W H, et al. DAMO-YOLO: a report on real-time object detection design[EB/OL]. arXiv: 2211.15444, 2022. https://doi.org/10.48550/arXiv.2211.15444.
30. Liu W Z, Lu H, Fu H T, et al. Learning to upsample by learning to sample[C]//*Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023: 6004–6014. https://doi.org/10.1109/ICCV51070.2023.00554.
31. Zhong J C, Chen J Y, Mian A. DualConv: dual convolutional kernels for lightweight deep neural networks[J]. *IEEE Trans Neural Networks Learn Syst*, 2023, **34**(11): 9528−9535.
32. Viola P, Jones M. Rapid object detection using a boosted cascade of simple features[C]//*Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, 2001: I. https://doi.org/10.1109/CVPR.2001.990517.
33. 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*, 2019: 658–666. https://doi.org/10.1109/CVPR.2019.00075.
34. Zheng Z H, Wang P, Liu W, et al. Distance-IoU loss: faster and

better learning for bounding box regression[C]//*Proceedings of the 34th AAAI Conference on Artificial Intelligence*, 2020: 12993–13000. https://doi.org/10.1609/aaai.v34i07.6999.

35. Tong Z J, Chen Y H, Xu Z W, et al. Wise-IoU: bounding box regression loss with dynamic focusing mechanism[EB/OL]. arXiv: 2301.10051, 2023. https://doi.org/10.48550/arXiv.2301.10051.
36. Weng Y L, Xiang X D, Ma L H. SCR-YOLOv8: an enhanced algorithm for target detection in sonar images[J]. *J Real-Time Image Proc*, 2025, **22**(2): 62.
37. Jiao S X, Xu F H, Guo H T. Side-scan sonar image detection of shipwrecks based on CSC-YOLO algorithm[J]. *Comput, Mater Continua*, 2025, **82**(2): 3019−3044.

作者简介



谷剑峰 (2001-)，男，硕士研究生，研究方向为机器视觉、图像处理。

E-mail：1620611336@qq.com



【通信作者】刘锡祥 (1981-)，男，教授，博士，研究方向为惯性导航、感知定位等。

E-mail：scliuseu@163.com

Underwater object detection in side-scan sonar images based on YOLOv8-ESI

Gu Jianfeng[1,2], Liu Xixiang[1,2*], Xu Hao[1,2]

[1] School of Instrument Science and Engineering, Southeast University, Nanjing Jiangsu 210096, China;

[2] Key Laboratory of Micro-Inertial Instrument and Advanced Navigation Technology, Ministry of Education, Nanjing Jiangsu 210096, China

Abstract

Objective Side-scan sonar (SSS) is an indispensable acoustic imaging technology for seabed exploration, underwater archeology, and search and rescue operations. It provides wide-range, high-resolution images of the seafloor by transmitting acoustic pulses and recording their backscattered echoes. However, due to the complex underwater environment and the inherent characteristics of acoustic imaging, SSS images are typically contaminated by severe speckle noise, geometric distortions, uneven resolution, and blurred target boundaries. These degradations significantly impair the performance of conventional object detection models that are originally designed for optical images. Furthermore, SSS datasets are often limited in scale and suffer from inconsistent annotation quality, as the low resolution and fuzzy features make it difficult to label targets precisely. Such low-quality training samples introduce harmful gradients during optimization, further compromising model robustness. To address these challenges, this paper proposes an improved detection model named YOLOv8-Efficient Sonar Image (YOLOv8-ESI), aiming to achieve accurate, robust, and lightweight underwater target detection that meets the real-time requirements of resource-constrained autonomous underwater vehicles.

Methods YOLOv8-ESI is developed based on the YOLOv8n baseline through three synergistic innovations tailored to SSS image characteristics. First, a wavelet convolutional layer (WTConv) is integrated into the backbone network. Unlike standard convolutions that operate solely in the spatial domain, WTConv performs cascaded two-dimensional Haar wavelet decomposition, transforming feature extraction into the frequency domain. The low-frequency sub-band (LL) captures global structural information such as the highlight-shadow patterns of sonar targets, while the high-frequency sub-bands (LH, HL, HH) preserve edge details and textures. By applying small-sized depthwise convolutions to these sub-bands, WTConv effectively enlarges the receptive field without introducing excessive parameters. Moreover, the multi-scale wavelet decomposition enables adaptive noise suppression through thresholding in the frequency domain, enhancing the model's ability to distinguish true targets from background clutter. Second, the neck network is systematically reconstructed into a dynamic feature pyramid network (D-GFPN). This architecture incorporates an efficient dynamic upsampler (DySample) that employs a multi-level point sampling strategy to preserve fine-grained details during feature resolution recovery, which is critical for small target detection in SSS images. A cross-stage partial dual-path lightweight aggregation network module (CSP-DLAN) is also designed, utilizing parallel 1×1 and 3×3 grouped depthwise convolutions along with channel shuffle mechanisms to achieve efficient multi-scale feature fusion with minimal computational overhead. This design balances the extraction of spatial details and channel-wise interactions while promoting cross-group information flow. Third, the original CIoU loss is replaced with WIoU, which employs a dynamic non-monotonic focusing mechanism based on outlier degree to assess anchor box quality. By assigning smaller gradient gains to low-

Foundation item: Jiangsu Provincial Key Research and Development Program of China (BF2024072)

*Correspondence: Liu X X, E-mail: scliuseu@163.com

quality samples and focusing on medium-quality anchors, WIoU effectively mitigates the harmful effects of annotation noise and bounding box inaccuracies prevalent in SSS datasets, thereby improving training stability and overall detection accuracy.

Results and Discussions Experiments are conducted on a self-constructed SSS dataset (SSSD) that integrates publicly available sources (Seabed Objects-KLSG and SONAR-2019) with additional images collected from underwater experiments and online repositories. The dataset contains 956 original images of three target categories: shipwrecks (602 images), aircraft wrecks (279 images), and human remains (75 images), with resolutions ranging from 300 pixel×300 pixel to 800 pixel×500 pixel. To enhance generalization and prevent overfitting, the training set is augmented through horizontal flipping, random rotation, random brightness adjustment, and Gaussian noise addition, resulting in 2926 training samples. The dataset is split into training, validation, and test sets with a ratio of 7:2:1. Ablation studies are performed to validate the contribution of each proposed module. Integrating WTConv alone improves mAP50 by 2.6% and mAP50:95 by 2.5% while reducing parameters, demonstrating its effectiveness in frequency-domain feature enhancement and noise suppression. The proposed D-GFPN outperforms mainstream neck structures including PANet, BiFPN, HS-FPN, and RepGFPN, achieving 87.3% mAP50 with only 2.98 M parameters and 7.8 GFLOPs, representing a 9.1% parameter reduction and 8.3% FLOPs reduction compared to RepGFPN while maintaining the same mAP50. Among different WIoU variants, WIoUv3 yields the best performance, particularly improving detection of challenging categories such as human remains by 1.7%, which validates its effectiveness in handling low-quality samples. The complete YOLOv8-ESI model achieves 88.8% mAP50, and 63.4% mAP50:95, surpassing baseline YOLOv8n by 5.2% and 3.8%, respectively, while reducing parameters by 10% (from 3.01 M to 2.69 M) and computational complexity by 1.1 GFLOPs. Comparative experiments against state-of-the-art detectors including YOLOv3-tiny, YOLOv5s, YOLOv6n, YOLOv9c, YOLOv10n, YOLOv11n, RT-DETR, SCR-YOLOv8, and CSC-YOLO demonstrate that YOLOv8-ESI achieves superior detection accuracy with competitive inference speed (1.3 ms per image) and model compactness. Heatmap visualizations confirm that the model focuses precisely on target regions with suppressed background activation, while detection examples illustrate reduced false positives and improved boundary localization compared to baseline models.

Conclusions This paper presents YOLOv8-ESI, a lightweight and accurate detection model specifically designed for side-scan sonar imagery. By integrating wavelet convolution for frequency-domain feature enhancement, a dynamic feature pyramid network for adaptive multi-scale fusion, and a wise IoU loss for robust training against annotation noise, the model effectively addresses the core challenges of noise interference, scale variability, feature blurring, and low-quality annotations inherent in SSS data. Experimental results demonstrate that YOLOv8-ESI significantly outperforms baseline and state-of-the-art models in both accuracy and efficiency, meeting the real-time detection requirements of underwater platforms. The ablation studies validate the synergistic effects of the proposed modules, and the comparative experiments confirm the model's superiority across multiple evaluation metrics. This work provides an effective technical solution for advancing underwater object detection in marine exploration, archaeology, and search and rescue operations. Future work may explore the integration of advanced denoising techniques and the extension of the model to other underwater imaging modalities such as synthetic aperture sonar and multi-beam echo sounders.

Keywords: side-scan sonar; YOLOv8n; wavelet convolution; dynamic up-sampler; loss function

Gu J F, Liu X X, Xu H. Underwater object detection in side-scan sonar images based on YOLOv8-ESI[J]. *Opto-Electron Eng*, 2026, **53**(6): 250370; DOI: 10.12086/oee.2026.250370

Opto-Electronic Journals Group
www.oejournal.org

扫描二维码，获取 PDF 全文