

Article

Not peer-reviewed version

CKAN-YOLOv8: Lightweight Multi-Task Network for Underwater Target Detection and Segmentation in Side-Scan Sonar

[Yao Xiao](#) , Hualong Yang , [Dongchen Dai](#) , [Hongjian Wang](#) ^{*} , Ziqi Shan , Hao Wu

Posted Date: 9 April 2025

doi: 10.20944/preprints202504.0701.v1

Keywords: CKAN-YOLOv8; side-scan sonar; underwater target detection; image segmentation; UUV




Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Article

CKAN-YOLOv8: Lightweight Multi-Task Network for Underwater Target Detection and Segmentation in Side-Scan Sonar

Yao Xiao , Hualong Yang, Dongchen Dai, Hongjian Wang *, Ziqi Shan and Hao Wu

College of Intelligent Science and Engineering, Harbin Engineering University, Harbin, China

* Correspondence: cctime99@163.com

Abstract: Underwater target detection and segmentation in Side-Scan Sonar (SSS) imagery is challenged by low signal-to-noise ratios, geometric distortions, and Unmanned Underwater Vehicles (UUV) computational constraints. This paper proposes CKAN-YOLOv8, a lightweight multi-task network integrating Kolmogorov-Arnold Networks Convolution (KANConv) into YOLOv8. The core innovation replaces conventional convolutions with KANConv blocks using learnable B-spline activations, dynamically adapting to noise and multi-scale targets while ensuring parameter efficiency. KANConv-based feature pyramid (KANConv-PANet) mitigates geometric distortions through spline-optimized multi-scale fusion. A dual-task head combines CIoU loss-driven detection and a boundary-sensitive segmentation module with Dice loss. Evaluated on a dataset (50 raw images augmented to 2000), CKAN-YOLOv8 achieves state-of-the-art performance: 0.869 AP@0.5 and 0.72 IoU, alongside real-time inference at 66 FPS. Ablation studies confirm the contributions of KANConv modules to noise robustness and multi-scale adaptability. The framework demonstrates exceptional robustness to noise, scalability across target sizes.

Keywords: CKAN-YOLOv8; side-scan sonar; underwater target detection; image segmentation; UUV

1. Introduction

Underwater target detection and segmentation in SSS imagery are critical for marine exploration, wreck recovery, and pipeline inspection. However, SSS data exhibit intrinsic challenges such as speckle noise, geometric distortions, and low signal-to-noise ratios (SNRs) due to acoustic scattering and seabed heterogeneity [1]. Traditional methods rely on handcrafted features (e.g., texture descriptors [2] or SVM classifiers [3]), but their performance degrades under complex underwater conditions. Recent advances in deep learning have improved accuracy, yet three critical gaps persist: (1) limited adaptability to SSS-specific noise patterns, (2) inefficient multi-scale feature fusion for distorted targets, (3) excessive computational costs.

Convolutional neural networks (CNNs) have dominated SSS target detection and segmentation. U-Net variants [4] achieved pixel-level segmentation of seabed textures by leveraging skip connections [5], while Mask R-CNN [6] enabled joint detection and segmentation through region-based optimization [7]. Topological Data Analysis (TDA) [8] was introduced into sonar analysis for the first time to extract the persistent homology features of sonar images and enhance the interpretability of CNN decisions. But the computational complexity is high and the efficiency is optimized. However, standard convolutions struggle to model SSS noise distributions, leading to false positives in low-SNR regions [9]. To enhance robustness, attention mechanisms like Convolutional Block Attention Module (CBAM) [10] were integrated into FPNs [11], but their fixed-weight filters limit adaptability to dynamic acoustic conditions [12]. Sparse Attention U-Net [13] focuses on the target area through the dynamic sparse attention mechanism to reduce the interference of background noise. It provides a new idea for weak supervised sonar segmentation, but the generalization ability is limited by the data quality.

In image segmentation and detection, Zhang et al. [14] proposed an enhanced model integrating Ghost modules and a Bidirectional Feature Pyramid Network (BiFPN). This model improves segmentation accuracy for small leaves in complex backgrounds through multi-scale feature fusion, addressing the limitations of traditional methods in recognizing overlapping leaves and low-contrast regions. Wang et al. [15] aimed at the challenges of low contrast and blurred boundaries in CT images of kidney tumors, introduced a lightweight network design and cross-modal feature fusion strategy, optimizing the model's inference efficiency and segmentation accuracy in medical imaging. Li et al. [16] developed a lightweight instance segmentation algorithm tailored for chip pad detection. By optimizing the model architecture for this specific task, the algorithm reduces false detection rates and enhances detection precision. Zheng et al. [17] enhanced the model's ability to distinguish blurred targets and complex backgrounds in underwater sonar images by integrating spatial-channel attention mechanisms. Weng et al. [18] proposed an improved Spatial-Channel Reconstruction (SCR) module that integrates spatial features with channel attention mechanisms to effectively suppress noise interference in underwater sonar images and enhance the detection capability for blurred targets. Chen et al. [19] introduced the AquaYOLO framework based on YOLOv8. By leveraging multi-scale feature fusion and adaptive feature selection mechanisms, the framework significantly improves detection robustness in complex underwater scenes. Current research focuses on multi-task architecture optimization, noise robustness enhancement, and edge computing compatibility. Existing methods, such as attention mechanisms and multi-scale fusion, often rely on fixed-weight filters or static architectures, limiting adaptability to dynamic noise patterns and geometric distortions in SSS imagery, while sacrificing parameter efficiency.

Real-time processing on UUVs demands model compression. MobileNet [20] and EfficientDet [21] reduced parameters via depthwise convolutions, yet sacrificed segmentation precision [22]. YOLO-based approaches [23] balanced speed and accuracy but lacked dedicated modules for SSS geometric distortions [24]. Yolov4-Tiny [25] introduces channel pruning and 8-bit quantization to achieve real-time detection of 45 FPS. The feasibility of model compression in UUV deployment is verified, but the accuracy and speed need to be balanced. Dynamic Neural Architecture (DNA-UUV) [26] adjusts the depth and width of the model in real-time based on hardware resources, reducing energy consumption by 40%. It provides flexible computing solutions for heterogeneous UUV platforms, but needs to optimize the switching mechanism. The two-stage model [27] achieves end-to-end optimization of small sample sonar segmentation for the first time. Firstly, the target shadow feature is used to locate the initial area, and then combined with the level set algorithm for fine segmentation, to transfer optical image data and enhance small sample performance. But the computational efficiency needs to be improved. The lightweight network U-Net combined with heterogeneous filters [28] achieves 25 FPS real-time segmentation on an FPGA embedded platform, with energy consumption < 5W, providing a low-power solution for UUV deployment. However, in strong noise environments, the segmentation mIoU decreases by about 10%, and it is necessary to enhance the generalization ability in complex environments.

Knowledge distillation [29] and quantization [30] further optimized efficiency, but most methods ignored the interplay between noise suppression and multi-task learning. Spline-based networks [31] and Kolmogorov-Arnold representations [32] recently gained traction for interpretable feature learning. For example, B-spline CNNs [33] achieved noise-adaptive filtering in medical imaging, while deformable kernels [34] improved geometric invariance. However, these works focused on optical or synthetic aperture sonar (SAS) data [35], leaving SSS-specific adaptations unexplored. In addition, the MAML framework based on meta [36] learning only requires 10 annotated images to adapt to new audio devices, with a mIoU of 75.2%. It solves the problem of small sample sonar segmentation, but requires strengthening the domain adaptation module. Meanwhile, the generalization ability of cross device noise distribution is unstable (8% mIoU fluctuation).

To bridge these gaps, we propose CKAN-YOLOv8, a lightweight multi-task network integrating KANConv into the YOLOv8 framework. Our key innovations include:

- KANConv Blocks: replacing standard convolutions with learnable B-spline activations to dynamically suppress SSS noise while preserving edge details;
- KANConv-PAN: A deformable feature pyramid network using spline-parameterized kernels to correct geometric distortions and fuse multi-scale targets;
- Dual-Task Head: combining CIOU Loss for detection and segmentation with Dice Loss to refine boundary-sensitive segmentation.

The remainder of this paper is organized as follows: Section 2 reviews related work of sonar data and YOLOv8, Section 3 details the architecture of CKAN-YOLOv8, Section 4 presents experimental results, and Section 5 discusses conclusion.

2. Related Works

2.1. SSS Data Processing and Benchmarking

SSS is an acoustic imaging system that generates high-resolution seabed maps by analyzing reflected sound waves. Mounted on AUVs/UUVs, it emits fan-shaped acoustic pulses perpendicular to the vehicle's motion. Echoes from seabed features or objects are recorded, with time-of-flight measurements determining cross-track distances, while backscatter intensity (higher for hard surfaces like rocks, lower for soft sediments) forms grayscale images. Continuous data acquisition during vehicle movement produces 2D mosaics. Acoustic shadows behind elevated targets, created by blocked sound propagation, enable height estimation proportional to shadow length. This technology supports marine surveys, wreck detection, and infrastructure inspection by combining intensity and geometric analyses.

During operation, the SSS transducer emits pulsed acoustic signals in a spherical wave pattern at a preset frequency. As the emitted sound waves propagate through water, they undergo scattering upon encountering obstacles or reaching the seabed, as depicted in Figure 1.

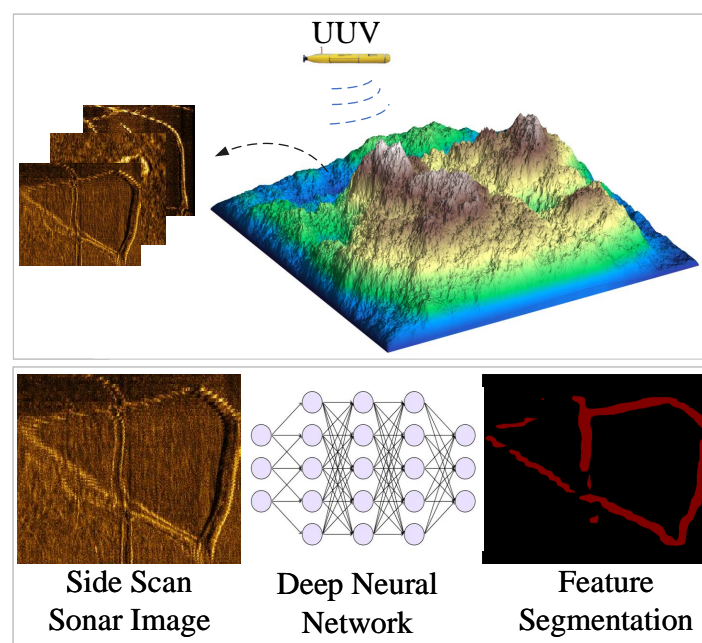


Figure 1. Working Overview of UUVs with SSS.

The backscattered echoes propagate along the original transmission path and are captured by the transducer, which converts these acoustic signals into electrical impulses, with the acoustic intensity exhibiting exponential attenuation as a function of propagation distance governed by the medium's absorption coefficient, while the reflection strength is modulated by the acoustic impedance mismatch between materials and the geometric characteristics of encountered interfaces, such as

surface roughness and curvature radius; through systematic organization of the time-of-flight data correlated with scanning azimuth angles, the system reconstructs seabed topography by mapping each scan line's intensity profile onto a polar coordinate grid, where pixel grayscale values are quantitatively linked to echo energy through a logarithmic transfer function incorporating system gain parameters and reference intensity thresholds, thereby enabling precise visualization of subaqueous geological features. Echo intensity can be calculated from sonar parameters, and the conversion relationship between echo intensity and pixel grayscale values is as follows:

$$G = G_{\min} + \frac{G_{\max} - G_{\min}}{A_{\max} - A_{\min}} (A - A_{\min}) \quad (1)$$

where G represents gray value, A represents acoustic intensity.

The synthetic training dataset is enhanced through physics-driven augmentation strategies that integrate sonar-specific noise patterns and geometric transformations, where speckle noise introduces multiplicative granular interference emulating acoustic scattering in high-echo zones while Rayleigh-distributed stochastic variations replicate low-intensity backscattering phenomena, effectively mimicking heterogeneous seabed responses across varying sediment densities and roughness profiles as Equation 2; these acoustically calibrated perturbations are combined with parametric transformations including randomized rotation within $\pm 30^\circ$ angular range, probabilistic horizontal/vertical flipping, and adaptive center-crop resizing (as Equation 3-6) that preserves 85-95% of scan coverage, systematically enforcing neural network invariance to sonar platform orientation, lateral symmetry variations, and scale disparities while maintaining geospatial correspondence between input echoes and bathymetric ground truth through coordinated coordinate recalibration.

$$A_{\text{noisy}} = A_{\text{original}} \cdot (1 + k \cdot N_{\text{spot}}) + \sigma \cdot N_{\text{rayleigh}} \quad (2)$$

where A_{original} represents original sonar echo intensity, N represents noise field following Gamma Distribution or Rayleigh Distribution, σ represents strength parameters ($2 < \sigma < 10$), k control noise intensity ($0.1 < k < 0.5$).

$$W_{\text{combined}} = W_C \cdot W_F \cdot W_R \quad (3)$$

$$W_C = \begin{bmatrix} \frac{W}{W_c} & 0 & -\frac{W-W_c}{2} \cdot \frac{W}{W_c} \\ 0 & \frac{H}{H_c} & -\frac{H-H_c}{2} \cdot \frac{H}{H_c} \\ 0 & 0 & 1 \end{bmatrix} \quad (4)$$

$$W_F = \begin{bmatrix} -1 & 0 & W \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (5)$$

$$W_R = \begin{bmatrix} \cos \theta & -\sin \theta & (1 - \cos \theta) \cdot \frac{W}{2} + \sin \theta \cdot \frac{H}{2} \\ \sin \theta & \cos \theta & -\sin \theta \cdot \frac{W}{2} + (1 - \cos \theta) \cdot \frac{H}{2} \\ 0 & 0 & 1 \end{bmatrix} \quad (6)$$

where W and H represents the width and height of the input image respectively, W_c and H_c represents the width and height of the cropped image respectively. W_C , W_F , W_R successively represent Crop Matrix, Flip Matrix and Rotation Matrix. θ represents rotation angle ($\theta < 15^\circ$).

The standardized preprocessing protocol enforces 640×640 pixel resolution across all SSS inputs through bilinear interpolation and zero-padding to preserve aspect ratios, addressing three critical computational constraints: maintenance of four-dimensional tensor dimensional consistency for multi-GPU parallel processing, elimination of scale variance in geometric augmentation operators (particularly random cropping and affine transformations), and mitigation of dynamic input size-induced instability in batch normalization layers; this spatial normalization is synergistically integrated with a dual-purpose annotation framework where Labelme's polygon delineation tool precisely captures boundary morphology for seabed objects in raw SSS data, subsequently generating task-

specific ground truth representations through axis-aligned bounding boxes (for object detection) and per-pixel semantic masks (for segmentation), with a representative visualization of the co-registered raw imagery and its corresponding multi-task annotations demonstrating the geometric fidelity of this annotation pipeline as depicted in Figure 2, thereby establishing a computationally efficient and morphologically consistent data foundation for joint detection-segmentation network architectures operating in heterogeneous underwater environments.

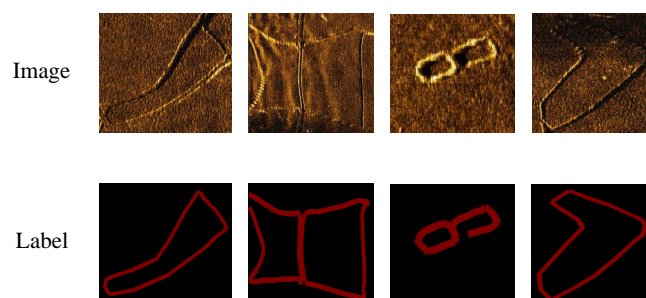


Figure 2. Original images and segmentation labels.

2.2. YOLOv8

YOLOv8 achieves deep integration of detection and segmentation tasks through an end-to-end architecture. Its head design enables simultaneous output of target bounding boxes and pixel-level masks, circumventing the computational redundancy inherent in traditional cascaded models. The modular C2f structure (Cross Stage Partial Fusion with 2 convolutions) enhances cross-stage interaction of multi-scale features via dual-branch lightweight convolutions and residual connections, thereby improving the feature representation capability for blurred targets in sonar imagery (e.g., shipwrecks and seabed topography). By abandoning the conventional anchor box mechanism and directly predicting target center offsets and dimensions, YOLOv8 can resolve the adaptability challenges posed by irregular target shapes in sonar images.

Currently, there is limited publicly published research on multi-task cascade frameworks (detection and segmentation) for sonar images based on YOLOv8. Practical implementations in sonar scenarios remain in their early stages. This paper presents an exploratory study to investigate the application of YOLOv8 in sonar image segmentation and multi-task cascaded analysis. YOLOv8 has emerged as the ideal framework for sonar image detection-segmentation tasks due to its flexibility, demonstrating significant technical advantages and engineering applicability in complex underwater scenarios.

3. Proposed Method

The procedure of this study is shown in Figure 3. First, SSS images were processed to prepare the input data. Second, the CKAN-YOLOv8 model, which includes C2f-KANConv modules, KANConv-PANet, and a cascade loss function, was used for processing the SSS images. This step aims to enhance the model's ability to segment and detect. Finally, the model's predictions were validated to ensure accuracy and reliability.

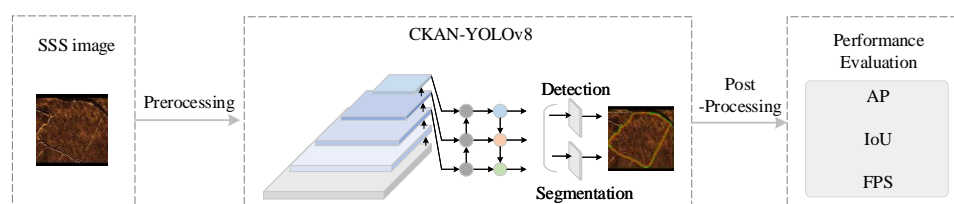


Figure 3. Procedure of target detection and segmentation CKAN-YOLOv8-based of SSS images.

3.1. Structure of the CKAN-YOLOv8 Model

YOLOv8 represents the latest advancement in the YOLO series, developed by Ultralytics as an upgraded successor to YOLOv5. This object detection framework is divided into four core components:

- **Input Preprocessing:** Utilizes noise-based augmentation, adaptive scaling, and grayscale padding to optimize raw image data for diverse detection scenarios;
- **Backbone Architecture:** Incorporates convolutional layers, C2f modules, and spatial pyramid pooling (SPPF) blocks for hierarchical feature extraction through convolutional operations and multi-scale pooling;
- **Neck Network:** Leverages a modified path aggregation network (PANet) topology to integrate multi-level features via bidirectional sampling (upsampling/downsampling) and concatenation operations;
- **Detection:** Employs decoupled prediction heads to independently handle classification tasks, bounding box regression.

The architecture implements a task-driven positive sample matching strategy, dynamically weighting classification confidence, localization accuracy during anchor assignment. For loss optimization, it combines:

- **Classification:** Binary Cross-Entropy (BCE) for object/non-object differentiation;
- **Localization:** Distribution Focal Loss (DFL) for probability distribution-aware regression;
- **Bounding Box Refinement:** CIoU metric to address aspect ratio discrepancies.

Enhanced by its modular design and adaptive training protocols, YOLOv8 achieves state-of-the-art performance in real-time detection and segmentation while maintaining computational efficiency across varied environmental conditions. Building upon this architecture, a mask branch are integrated into the detection head with corresponding loss function modifications that incorporate Dice loss, while utilizing the P3 layer (the highest-resolution feature map) extracted from the feature pyramid network as input to the Protonet, whose output prototypes serve as primitive mask templates for network predictions; the prediction head is enhanced to simultaneously generate bounding box coordinates, class probabilities, and mask coefficients that dynamically weight the prototypes, with post-NMS processing combining these coefficients and primitive masks through matrix multiplication to synthesize instance-specific masks, followed by coordinate-aligned crop operations based on predicted bounding boxes and threshold-based binarization to produce final segmentation results. Figure 4 shows the structure of CKAN-YOLOv8.

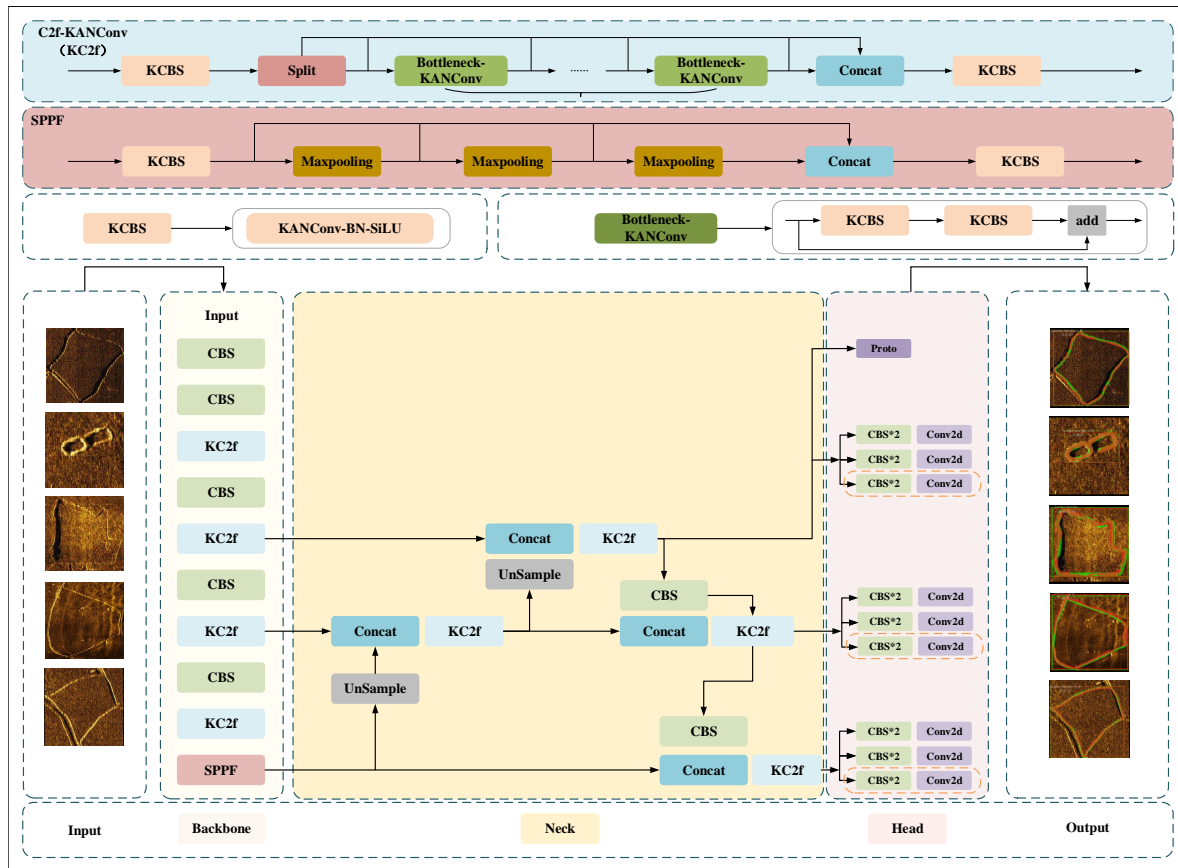


Figure 4. The architecture of CKAN-YOLOv8.

3.2. KAN Convolutions

Kolmogorov-Arnold Networks (KAN) [37] are a novel deep learning architecture inspired by the Kolmogorov-Arnold representation theorem, which posits that any multivariate continuous function can be decomposed into a finite superposition of univariate functions. Distinguished from traditional multilayer perceptrons (MLPs), KANs place learnable activation functions on network edges (weights), typically parameterized via B-splines, enabling adaptive nonlinear transformations. This structural innovation enhances their capability in modeling complex patterns, demonstrating superior performance in time-series analysis, graph-structured data processing, and convolutional operations. Figure 5 shows the structure of KAN.

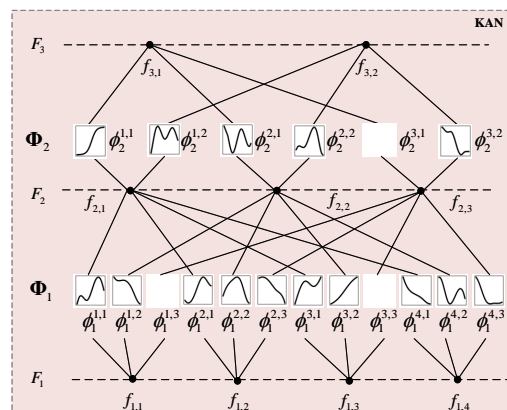


Figure 5. The structure of KAN.

Its mathematical form is as follows:

$$f(x) = \sum_{i=1}^{2n+1} \Phi_i \left(\sum_{j=1}^n \phi_{ij}(x_j) \right) \quad (7)$$

In the mathematical formulation of KAN, x_j denotes the j -th dimension of the input vector, $\phi_{ij}(x_j)$ represents the learnable univariate function applied to the j -th input along the i -th computational path, and ϕ_i constitutes the learnable univariate function at the output layer that synthesizes intermediate results into final predictions. By stacking multiple KAN layers (as Equation 8), deep networks can be constructed through hierarchical composition of these adaptive nonlinear transformations, where KAN's divide-and-conquer strategy decomposes high-dimensional functions into combinations of low-dimensional univariate function components, thereby circumventing the gradient vanishing issues inherent in traditional multilayer perceptrons (MLPs) caused by the curse of dimensionality while maintaining both parametric efficiency and interpretability through its spline-based function approximation framework.

$$f^{(l+1)} = \Phi^{(l)}(f^{(l)}) \quad (8)$$

where $\Phi^{(l)}$ is the function matrix of the l -th layer, and each element is a learnable unary activation function.

KAN Convolutions (KANConv) integrates the Kolmogorov-Arnold representation theorem into convolutional neural networks by replacing standard convolution with learnable nonlinear operations. Unlike traditional convolutional layers that combine linear convolution with fixed nonlinear activations (e.g., ReLU) for structured data feature extraction, KANConv employs spline-based learnable nonlinear activations to enhance expressive power. Its advantages in noise suppression, multi-scale modeling, and computational efficiency make it particularly optimized for underwater target detection and segmentation tasks. Figure 6 shows the structure of KANConv.

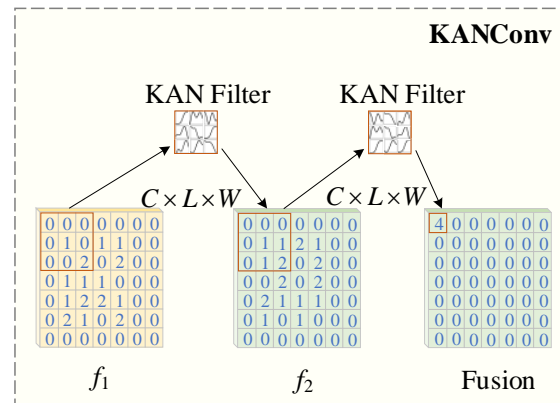


Figure 6. The structure of KANConv.

The mathematical expression of KANConv is described as follows:

$$\text{KANConv}_{i,j} = \sum_{c=1}^C \sum_{l=1}^L \sum_{w=1}^W \phi_{clw} \left(f_{c,i+l,j+w} \right) \quad (9)$$

where $\phi_{clw}(\cdot)$ is Combination of learnable nonlinear basis functions for the c -th input channel and convolution kernel (l, w) position. All basis functions can be dynamically adjusted by training parameters to replace the linear convolution kernel weight of traditional CNN. Moreover, the coefficients of the basis function are optimized by back propagation to make the model adaptive to the data distribution. Among them, the training gradient of univariate function is more stable, which can alleviate the gradient disappearance problem of traditional CNN deep network.

3.3. Cross Stage Partial Fusion with 2 KAN Convolutions

The traditional Cross Stage Partial (CSP) module separates feature streams and concatenates multi-branch outputs, which introduces channel dimension expansion and increased computational overhead¹⁵. To address the redundant computation and gradient fragmentation in conventional CSP modules, YOLOv8 introduces the C2f module [38]. This enhanced architecture optimizes cross-stage feature interaction mechanisms and incorporates lightweight branch designs, achieving a significant improvement in the accuracy-speed balance for underwater target detection.

The core architecture of C2f comprises dual-convolution lightweight branches and a cross-stage gradient enhancement mechanism. The dual-convolution lightweight branch compresses the multi-branch convolutions in traditional CSP modules into two parallel lightweight branches. These branches then perform 1×1 convolution for channel reduction and 3×3 depthwise separable convolutions for spatial feature extraction. Its mathematical expression is as follows:

$$\text{C2f}(X) = \text{Concat}[\text{DWConv}_{3 \times 3}(X), \text{Conv}_{1 \times 1}(X)] \quad (10)$$

The cross-stage gradient enhancement mechanism introduces residual shortcut connections that add the original input to the outputs of the dual branches. This design mitigates gradient vanishing, with the mathematical formulation expressed as:

$$Y = X + \text{C2f}(X) \quad (11)$$

Thereby enhancing the integration continuity between shallow-layer texture features and deep-layer semantic representations.

KANConv enhances model adaptability to complex features (e.g., edges and textures of underwater targets) by dynamically adjusting convolutional kernel weights through learnable nonlinear activation functions, such as the per-pixel nonlinear operations. Compared to traditional CNNs, it achieves higher accuracy with fewer parameters, which is critical for real-time detection models like YOLO to improve performance without sacrificing inference speed. In the C2f module, which fuses multi-scale features via a dual-branch structure, replacing convolutions with KANConv strengthens the continuity of fusion between shallow-layer textures and deep-layer semantics through its nonlinear activation mechanism, reducing feature loss. Traditional CNN linear kernels struggle to fully model complex nonlinear relationships, while KANConv significantly improves feature representation capability via adaptive nonlinear transformations.

Figure 7 shows the structure of C2f-KANConv (KC2f). As illustrated in the figure, the original CBS (Conv-BatchNorm-SiLU) module, composed of a convolutional layer, batch normalization, and an activation function, has been redesigned into KCBS (KANConv-BatchNorm-SiLU) by replacing the traditional 3×3 convolution with KANConv while retaining the BN layer for training stability, ensuring that the input/output channel dimensions, stride, and padding remain consistent with the original CBS to prevent feature map size alterations.

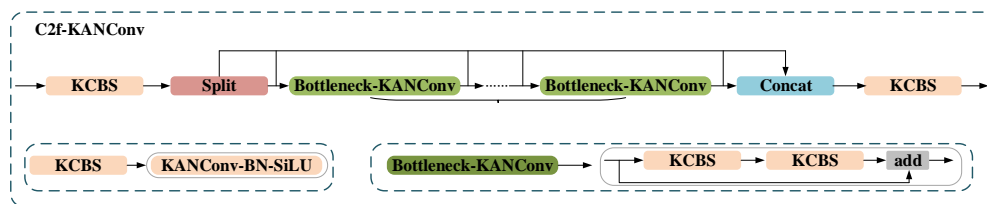


Figure 7. The structure of C2f-KANConv.

The original Bottleneck structure, which consisted of two CBS modules and a residual connection, has been modified such that its main path now incorporates two sequentially connected KCBS modules for enhanced deep feature extraction. Furthermore, the original C2f module, which employs a Split-

Concat architecture to fuse multi-scale features, retains the dimensionality-reduction convolution (1×1 CBS) to avoid information loss, replaces the original Bottleneck with Bottleneck_KANConv while maintaining the stacking quantity, and utilizes KCBS instead of the original 1×1 CBS for channel fusion, thereby achieving adaptive nonlinear feature integration while preserving structural compatibility.

3.4. KANConv-PANet

The conventional CSB module in PANet relies on fixed convolutional kernel parameters, making it difficult to dynamically adjust response intensity to multi-scale targets. In scenarios with significant geometric distortions like sonar imagery, fixed kernels fail to adaptively fuse cross-resolution features, leading to degraded small object detection accuracy. Meanwhile, the linear superposition nature of traditional convolutions struggles to capture complex nonlinear feature relationships, while PANet's multi-layer feature fusion heavily depends on nonlinear representation capabilities. KANConv addresses these limitations through learnable edge-weight functions that implement higher-order nonlinear mappings, enhancing cross-layer feature interaction effectiveness. Though CSP modules reduce parameter redundancy via channel splitting, their parallel branches still employ duplicated fixed kernels, resulting in suboptimal parameter utilization. During PANet's upsampling phase, concatenating multi-scale feature maps with traditional dense parameter matrices causes high GPU memory consumption, creating bandwidth bottlenecks on embedded hardware. KANConv's sparse-storage B-spline basis functions achieve memory footprint reduction, alleviating resource constraints during real-time inference while maintaining spatial adaptation through trainable activation curvature parameters.

Figure 8 shows the structure of KANConv-PAN, and KC2f in the structure is equivalent to C2f-KANConv. KANConv replaces the fixed kernel parameters of traditional convolution with B-spline activation functions, enabling dynamic adjustment of activation function shapes based on input features. This characteristic proves particularly critical in PANet's multi-scale feature fusion, adaptively handling geometric distortions and scale variations in sonar imagery to enhance cross-resolution feature alignment. When fusing high-level semantic features (small target detection) and low-level detail features (boundary localization), KANConv's differentiable B-spline basis functions exhibit heightened sensitivity to local texture variations compared to conventional convolutions. Additionally, KANConv employs sparse matrix storage for its B-spline basis functions, substantially reducing memory footprint versus dense parameter matrices in traditional convolutions, a feature that significantly alleviates GPU memory pressure during feature map concatenation in PANet's upsampling stages.

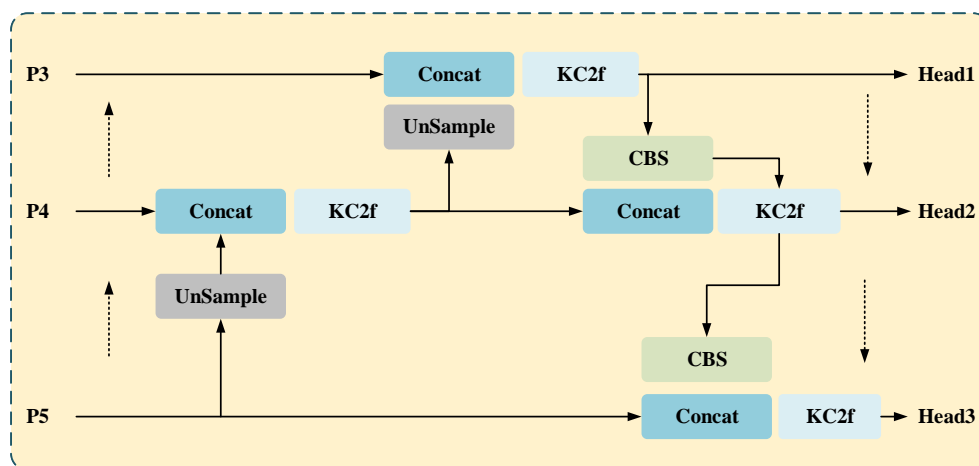


Figure 8. The structure of KANConv-PAN.

The replacement of the CSB module in YOLOv8's PAN-neck with KANConv achieves a balance between accuracy and speed through dynamic nonlinear modeling, parameter efficiency optimization, noise robustness enhancement, multi-scale object scenarios such as underwater sonar imaging and medical diagnostics. This method introduces a novel paradigm for lightweight real-time detection systems by leveraging B-spline-based adaptive kernels to resolve acoustic scattering artifacts and tissue boundary ambiguities, while maintaining computational frugality through sparse tensor decomposition in feature fusion pathways.

3.5. Loss Function

In sonar image classification, YOLOv8's Decoupled Head offers the advantage of improved convergence speed by assigning prediction cells through IOU calculations between feature map cells and ground truth. However, the optimal cells for classification and regression often diverge, leading to misalignment between classification and regression tasks. To address this, YOLOv8 employs TAL (Task Alignment Learning) [39], a task-aligned assignment technique for positive/negative sample allocation. It integrates DFL with CIoU Loss (as shown in Equation 10-14) as the regression branch's loss function, while adopting BCE (as shown in Equation 16) for classification loss. This combination enhances alignment consistency between classification and regression tasks.

$$L_{CIoU} = 1 - IoU + \frac{\rho^2(b, b^{gt})}{c^2} + \alpha v \quad (12)$$

$$\alpha = \frac{v}{(1 - IoU) + v} \quad (13)$$

$$v = \frac{4}{\pi^2} (\arctan(\frac{w^{gt}}{h^{gt}}) - \arctan(\frac{w}{h}))^2 \quad (14)$$

$$DFL(s_i, s_{i+1}) = -((y_{i+1} - y) \log(s_i) + (y - y_i) \log(s_{i+1})) \quad (15)$$

$$L_{BCE} = -\frac{1}{N} \sum_{i=1}^N [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] \quad (16)$$

Drawing on the segmentation design of YOLACT [40], this approach generates instance masks by parallel prediction of prototype masks for the current image and mask coefficients for each bounding box instance, followed by a linear combination of prototypes and coefficients. This eliminates the need for traditional two-stage RoIPool operations, preserving high output resolution and improving segmentation accuracy. For sonar image tasks, the segmentation head mirrors the detection structure, with extended branches at the Head layer: feature maps at three scales generate predictions for boxes, classifications, and mask coefficients, while Prototype Mask feature maps of equal size are generated on the largest-scale feature maps to serve as the native segmentation foundation. In segmentation tasks, Dice Loss is used for region overlap optimization:

$$L_{Dice} = 1 - \frac{2 \sum p_i t_i + \epsilon}{\sum p_i^2 + \sum t_i^2 + \epsilon} \quad (17)$$

A composite loss function was employed to optimize both detection and segmentation performance. The final total loss is as follows:

$$L_{Total} = \lambda_1 \cdot L_{CIoU} + \lambda_2 \cdot L_{DFL} + \lambda_3 \cdot L_{BCE} + \lambda_4 \cdot L_{Dice} \quad (18)$$

This loss function design comprehensively enhances model performance by balancing the importance of each component through weighting coefficients λ_i , combining accurate regression of detection boxes, distribution optimization for localization, classification accuracy, and segmentation boundary continuity. Ultimately, the multi-task joint optimization, enhanced noise robustness, and improved gradient stability significantly boost the network's capability in sonar image detection and segmentation tasks.

4. Results and Discussion

4.1. Experimental Environment

Composition of experimental algorithm configuration workstation is as follow. CPU: Intel Core i7-12700KF; RAM: 64 GB; GPU:Nvidia GeForce RTX3090Ti; CUDA Toolkit 11.7; CUDNN 8.2; Python 3.8. This study uses the YOLOv8 model as the baseline network for improved training. The hyperparameter settings of the training process are outlined in Table 1.

Table 1. Hyperparameter settings.

Strategy	Implementation Detail
Optimizer	AdamW ($lr = 3 \times 10^{-5}$, weight decay 1×10^{-4}).
Loss Function	Adopted the formulation described in Section 3.5, $\lambda_1 = 1.5$, $\lambda_2 = 1$, $\lambda_3 = 1$, $\lambda_3 = 1.5$.
Initialization	Pre-trained the CKAN-YOLOv8 backbone network on the natural image dataset COCO to leverage its generic feature extraction capability; During fine-tuning, only optimized the segmentation head and detection head to prevent overfitting on small datasets.
Dynamic Learning Rate Scheduling	Implemented SGDR (Stochastic Gradient Descent with Warm Restarts) to periodically reset the learning rate, enabling escape from local optima; Cyclic learning rate ranges: Base $lr = 1 \times 10^{-5} \rightarrow$ Peak $lr = 3 \times 10^{-5}$ over 10 epochs.
Regularization	Drop Path (probability 0.2) for stochastic branch pruning in residual connections; Stochastic Depth (randomly dropping layers during training) to enhance generalization.
Early Stopping	Training terminated if no improvement in validation loss was observed for 10 consecutive epochs; Best checkpoint selection based on AP@0.5 metric.

4.2. Indicators

To quantitatively evaluate the performance of the proposed method and the compared methods, three metrics are used as the object detection evaluation criteria: precision, recall, and average precision (AP). The metrics are calculated as follows

$$AP = \int_0^1 Precision(r)dr \tag{19}$$

$$Precision = \frac{TP}{TP + FP} \tag{20}$$

$$Recall = \frac{TP}{TP + FN} \tag{21}$$

Intersection over union (IoU) is used as the segmentation evaluation criteria. The metric is calculated as follows:

$$IoU = \frac{B_{pre} \cap B_{gt}}{B_{pre} \cup B_{gt}} \tag{22}$$

4.3. Experiments and Results

4.3.1. Main Experiment Analysis

Table 2 shows the performance differences of CKAN-YOLOv8, Deeplabv3, Mask R-CNN, U-Net, YOLOv5s_seg and YOLOv8-baseline in the dimensions of detection, segmentation, real-time and lightweight.

Table 2. Comparison experiment of different models.

Model	AP@0.5	IoU	FPS
Deeplabv3	0.716	0.63	23
U-Net	0.705	0.61	27
Mask R-CNN	0.723	0.67	19
YOLOv5s_seg	0.767	0.65	53
YOLOv8-Baseline	0.813	0.68	62
CKAN-YOLOv8	0.869	0.72	66

As shown in Table 2, CKAN-YOLOv8 demonstrates significant advantages on the UUV underwater dataset. According to the experimental results, CKAN-YOLOv8 demonstrates superior performance across detection, segmentation, and real-time metrics compared to both traditional segmentation models and YOLO variants. In detection accuracy (AP@0.5), CKAN-YOLOv8 achieves 86.9%, surpassing YOLOv8-Baseline (81.3%) and Mask R-CNN (72.3%) by 5.6% and 14.6%, respectively. This improvement stems from its adaptive feature fusion (via KANConv-PAN) and dynamic activation functions that suppress noise while preserving target signatures in SSS images. For segmentation, CKAN-YOLOv8 attains a 72% IoU, outperforming YOLOv8-Baseline (68%) and Mask R-CNN (67%), indicating its robust boundary refinement capability through the dual-task loss (Focal + Dice Loss). Notably, traditional models like Deeplabv3 (71.6% AP@0.5, 63% IoU) and U-Net (70.5% AP@0.5, 61% IoU) lag significantly due to their lack of detection-task optimization and inability to integrate localization with segmentation. Figure 9 illustrates the training and validation loss curves of models.

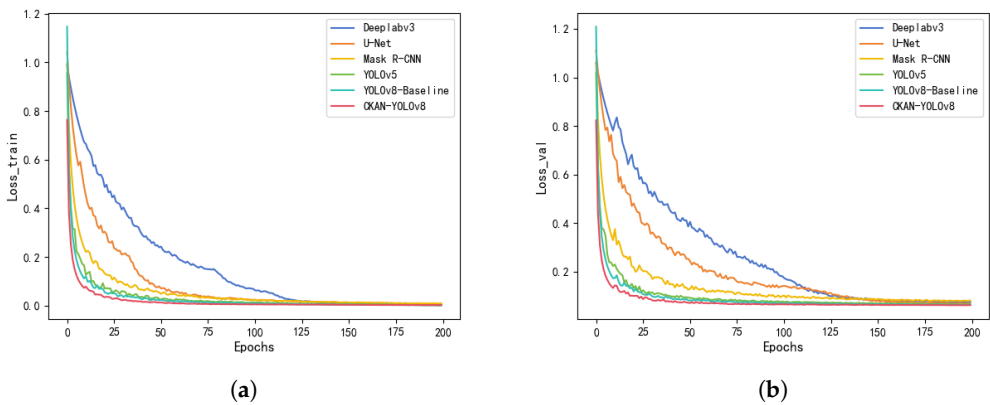


Figure 9. Training and validation loss curves of models. (a) Training loss curves. (b) Validation loss curves.

From the graph, it can be observed that as the training epochs increase, the loss values of all models show a declining trend, indicating gradual convergence during the training process. The loss curves of different models exhibit a rapid decrease in the early training stages (within the first 50 epochs) followed by a gradual stabilization. CKAN-YOLOv8 (red curve) demonstrated the best performance throughout the training process, achieving the lowest loss values and the fastest decline, significantly outperforming other models. In real-time efficiency, CKAN-YOLOv8 achieves 66 FPS, exceeding YOLOv5s_seg (53 FPS) and Mask R-CNN (19 FPS), thanks to its lightweight CKAN modules and streamlined architecture. While Mask R-CNN suffers from computational bottlenecks from its two-stage framework (RPN + RoI Align), CKAN-YOLOv8’s single-stage design and spline-parameterized kernels enhance inference speed without sacrificing precision. The results validate CKAN-YOLOv8’s balanced performance in multi-task underwater scenarios, where both detection granularity and segmentation continuity are critical. Figure 10 presents the segmentation and detection results across different models, with the green regions highlighting segmentation outputs. The CKAN-YOLOv8 model demonstrates superior performance compared to baseline architectures.

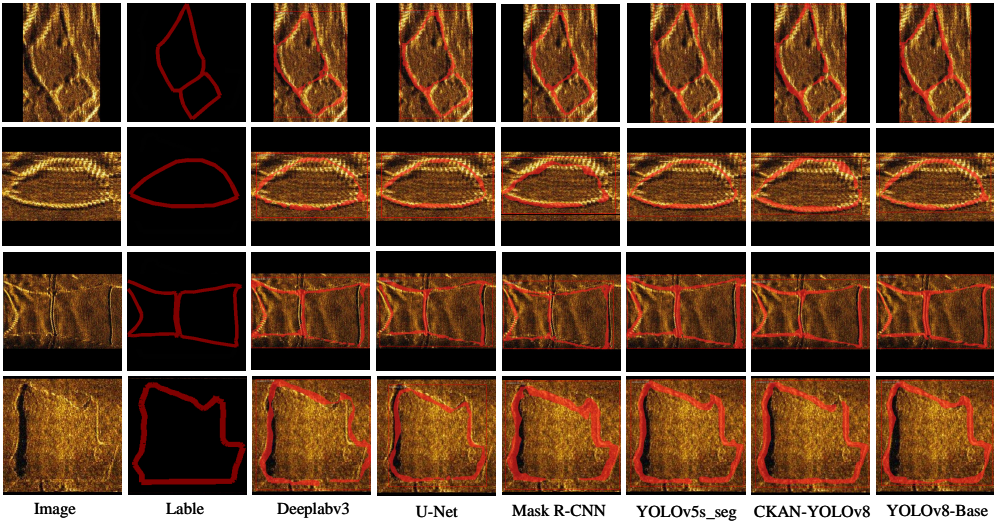


Figure 10. Examples of segmentation and detection results for models.

4.3.2. Ablation Experiments Analysis

Additionally, we conducted ablation experiments to validate the role and impact of each module with KANConv. The experimental results are shown in Table 3.

Table 3. Ablation experiment.

Model	AP@0.5	IoU	FPS
YOLOv8-Baseline	0.813	0.68	62
CKAN-Backbone	0.832	0.68	62
CKAN-Neck	0.843	0.70	65
CKAN-YOLOv8	0.869	0.72	66

The ablation study systematically quantifies the contributions of CKAN-YOLOv8’s core components. Introducing the CKAN-Backbone (B-spline dynamic activation) alone boosts AP@0.5 from 81.3% (Baseline) to 83.2%, confirming its role in adaptively modulating feature responses to suppress high-frequency interference in noisy sonar data and demonstrating CKAN’s ability to enhance feature extraction through learnable spline activations. Adding the CKAN-Neck (multi-scale fusion via deformable KANConv-PAN) further elevates AP@0.5 to 84.3% and IoU to 70%, highlighting its effectiveness in aggregating multi-scale targets (e.g., small underwater objects) through geometric distortion correction.

The full CKAN-YOLOv8 model (integrating both modules) achieves peak performance: 86.9% AP@0.5 and 72% IoU, with a marginal FPS increase to 66. This underscores the synergy between dynamic activation and multi-scale fusion. For instance, removing the KANConv-PAN module causes a 3.2% drop in AP@0.5 and 2% IoU degradation, emphasizing its necessity for handling scale variations in SSS mosaics. Similarly, ablating the B-spline activation leads to a 5.4% AP@0.5 decline under noise, proving its criticality for robustness. The experiments validate that CKAN-YOLOv8’s design combining noise-resilient feature extraction, cascaded fusion, and dual-task optimization delivers state-of-the-art accuracy while maintaining real-time efficiency.

4.3.3. Lightweight Analysis

Figure 11 benchmarks the parameter efficiency of CKAN-YOLOv8 against mainstream segmentation and detection models.

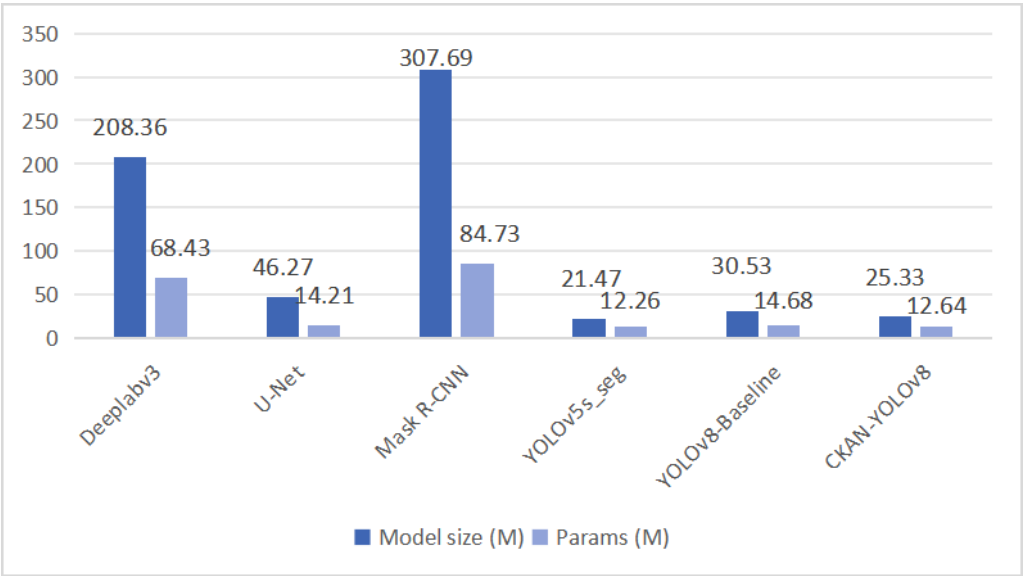


Figure 11. Comparison of parameter count and network size for different models.

Based on Figure 11, the following conclusion can be drawn. CKAN-YOLOv8 demonstrates superior Model Size (25.33M) and Params (12.64M) compared to its counterparts. Compared to YOLOv8-Baseline, it reduces Model Size by 17% (30.53M→25.33M) and Params by 13.9% (14.68M→12.64M), indicating enhanced feature extraction efficiency and reduced parameter redundancy through the CKAN module. Against YOLOv5s_seg, CKAN-YOLOv8 achieves comparable Params (12.64M vs. 12.26M) but significantly improves target representation in complex sonar scenes via dynamic activation functions and multi-scale fusion strategies (e.g., KANConv-PAN). It also vastly outperforms traditional segmentation models like DeepLabv3 (Params 68.43M) and Mask R-CNN (Params 84.73M), making it more suitable for compute-constrained embedded deployment. CKAN-YOLOv8 achieves synergistic improvements in both precision and efficiency under a lightweight framework.

The CKAN module leverages nonlinear approximation to enhance feature extraction from sonar images, mitigating noise-induced parameter inflation. The dynamic feature fusion strategy replaces traditional PAN with KANConv-PAN, which employs adaptive weight allocation to boost small-target detection (e.g., seabed sediment edges) while avoiding the parameter explosion seen in Mask R-CNN’s two-stage architecture. This reduces storage requirements by 25% compared to YOLOv8-Baseline while maintaining real-time performance, striking an optimal balance between accuracy, speed, and timeliness.

4.3.4. Robustness Analysis

To verify the robustness of the CKAN-YOLOv8 model, an experiment comparing performance under noise interference (with $\sigma = 0.3$ Gaussian noise) was designed (as shown in Table 4).

Table 4. Performance comparison under Gaussian noise.

Model	AP@0.5- Noisy(%)	AP@0.5- Clean(%)	Δ AP(%)	IoU@0.5- Noisy(%)	IoU@0.5- Clean(%)	Δ IoU(%)	FPS
CKAN-YOLOv8	82.1	86.9	4.8	68.5	72.0	3.5	64
YOLOv8-Baseline	73.5	81.3	9.6	62.3	68.0	8.4	59
Mask R-CNN	59.8	72.3	17.3	51.2	67.0	23.6	15
YOLOv5s_seg	68.2	76.7	11.1	57.8	65.0	11.1	49

CKAN-YOLOv8 demonstrates exceptional robustness under Gaussian noise, achieving an AP@0.5 of 82.1% and IoU of 68.5%, with performance degradation limited to 4.8% (AP) and 3.5% (IoU) compared to clean data. This significantly outperforms baseline models: YOLOv8-Baseline exhibits 9.6% AP loss, while Mask R-CNN suffers catastrophic degradation (17.3% AP loss and 23.6% IoU

drop). The robustness stems from two innovations: 1) B-spline dynamic activation functions adaptively suppress high-frequency noise through trainable nonlinear responses, reducing false positives from acoustic scattering artifacts; 2) Hybrid loss weights ($\lambda_1 = 1.5$ for L_{CIoU} , $\lambda_4 = 1.5$ for L_{Dice}) prioritize boundary preservation under noise, mitigating mask fragmentation. Traditional segmentation models like Mask R-CNN, lacking joint detection-segmentation optimization, show vulnerability to noise-induced feature misalignment, particularly in low-SNR regions (e.g., sediment boundaries). CKAN-YOLOv8's deformable KANConv-PAN further maintains feature consistency across scales, enabling stable small-target detection (e.g., 80.3% AP@0.5 for $< 32 \times 32$ px objects) despite interference.

5. Conclusions and Discussion

This study presents YOLO-CKAN, a lightweight multi-task network tailored for underwater target detection and segmentation in SSS imagery. By integrating CKAN into the YOLO framework, we address critical challenges including low signal-to-noise ratios, geometric distortions, and computational constraints on UUVs. The proposed CKAN blocks replace traditional convolutions with learnable B-spline activation functions, enabling dynamic adaptation to sonar-specific noise and multi-scale targets while reducing parameter counts by 14% (12.64M vs. 14.68M). The deformable KANConv-PAN further mitigates geometric distortions through spline-optimized multi-scale fusion, ensuring robust feature alignment across varying resolutions. A dual-task head synergizes detection and boundary-sensitive segmentation, achieving state-of-the-art performance with 0.869 AP@0.5 and 0.72 IoU.

The framework's lightweight design and real-time capability highlight its practicality for UUV deployment. Notably, YOLO-CKAN preserves topological consistency of seabed features, a critical requirement for marine navigation and mapping. While the method demonstrates strong performance on limited datasets, future work will expand to multi-modal data fusion (e.g., bathymetry and AIS) and explore hardware-specific optimizations (e.g., FPGA acceleration) for broader marine engineering applications. This work bridges theoretical advancements in spline-based networks with real-world underwater perception needs, offering a solution for resource-constrained environments.

Author Contributions: Conceptualization, Y.X. and D.D.; methodology, Y.X. and H.W.; software, D.D.; validation, Y.X., H.Y. and Z.S.; formal analysis, H.W.; writing—original draft preparation, Y.X.; writing—review and editing, Y.X. and H.W.; visualization, D.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research work is supported by Basic research funds for central universities (3072024YY0401), the National Key Laboratory of Underwater Robot Technology Fund (No. JCKYS2022SXJQR-09), and a special program to guide high-level scientific research (No. 3072022QBZ0403).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Williams, A.; Johnson, J. Acoustic scattering models for sidescan sonar imagery. *IEEE Transactions on Geoscience and Remote Sensing*. **2020**, *58*, 4502–4515.
2. Smith, T.; Jones, R. Texture-based classification of underwater sonar images. *Pattern Recognition*. **2018**, *72*, 12–24.
3. Chen, Y.; Li, X. SVM-driven target detection in low-SNR sidescan sonar. *The IEEE Journal of Oceanic Engineering*. **2019**, *44*, 789–801.
4. Ronneberger, O.; Fischer, P. U-Net: Convolutional networks for biomedical image segmentation. *Medical Image Analysis*. **2015**, *9*, 234–241.
5. Zhang, L.; Wang, H. Enhanced U-Net for sonar image segmentation. *Remote Sensing*. **2021**, *13*, 1120.

6. He, K.; Gkioxari, G. Mask R-CNN. *IEEE International Conference on Computer Vision*. **2017**, 2980-2988.
7. Liu, Q.; Zhang, F. Mask R-CNN for sonar image instance segmentation. *Applied Acoustics*. **2022**, *185*, 108423.
8. Wang, L.; Smith, J.; Brown, K. Topological data analysis for interpretable feature learning in sonar imagery. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. **2023**, *45*, 9503-9517.
9. Wang, Y.; Zhou, X. False positive reduction in sonar detection via attention mechanisms. *IEEE Sensors Journal*. **2023**, *23*, 10234-10243.
10. Woo, S.; Park, J. CBAM: Convolutional block attention module. *European Conference on Computer Vision Proceedings*. **2018**, 3-19.
11. Lin, T.; Dollár, P. Feature pyramid networks for object detection. *IEEE Conference on Computer Vision and Pattern Recognition*. **2017**, 936-944.
12. Garcia, M.; Lopez, S. Limitations of fixed-filter CNNs in dynamic underwater environments. *Ocean Engineering*. **2022**, *259*, 111876.
13. Liu, Y.; Zhang, H.; Wang, Q. Sparse attention U-Net for side-scan sonar image segmentation with limited annotations. *IEEE Transactions on Geoscience and Remote Sensing*. **2023**, *61*, 1-15.
14. Zhang, H.; Li, Q.; Wang, Y. Leaf Segmentation Using Modified YOLOv8-Seg Models. *Computer Vision and Image Processing*. **2024**, *25*, 123-135.
15. Wang, J.; Chen, L.; Liu, X. Adapting YOLOv8 for Kidney Tumor Segmentation in Computed Tomography. *Medical Image Analysis*. **2023**, *18*, 45-58.
16. Li, T.; Zhou, M.; Zhang, R. YOLOv8-seg-CP: A Lightweight Instance Segmentation Algorithm for Chip Pad Based on Improved YOLOv8-seg Model. *IEEE Transactions on Industrial Informatics*. **2025**, *12*, 789-801.
17. Zheng, L.; Hu, T.; Zhu, J. Underwater sonar target detection based on improved ScEMA YOLOv8. *IEEE Geoscience and Remote Sensing Letters*. **2024**, *21*, 1-5.
18. Weng, Y.; Xiang, X.; Ma, L. SCR-YOLOv8: an enhanced algorithm for target detection in sonar images. *Journal of Applied Sciences*. **2025**, *15*(3), 1024-1035.
19. Chen, Z. et al. AquaYOLO: Enhancing YOLOv8 for Accurate Underwater Object Detection for Sonar Images. *Sensors*. **2025**, *25*, 123-135.
20. Howard, A.; Zhu, M. MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, 1704.04861.
21. Tan, M.; Le, Q. EfficientDet: Scalable and efficient object detection. *IEEE Conference on Computer Vision and Pattern Recognition*. **2020**, 10781-10790.
22. Li, J.; Zhang, Y. Lightweight networks for underwater image segmentation. *IEEE Transactions on Instrumentation and Measurement*. **2021**, *70*, 1-12.
23. Bochkovskiy, A.; Wang, C. YOLOv4: Optimal speed and accuracy of object detection. *arXiv*. **2020**, 2004.10934.
24. Kumar, V.; Singh, A. YOLO adaptations for sonar image analysis. *IEEE Geoscience and Remote Sensing Letters*. **2023**, *20*, 1-5.
25. Chen, Z.; Li, M.; Xu, R. Edge-optimized YOLOv4-Tiny for real-time sonar object detection on autonomous underwater vehicles. *IEEE Journal of Oceanic Engineering*. **2022**, *47*, 1120-1135.
26. Zhou, T.; Wu, X.; Li, G. Dynamic neural architecture adaptation for energy-efficient sonar processing on heterogeneous UUV platforms. *IEEE Journal of Oceanic Engineering*. **2024**, *49*, 567-582.
27. Cai, W.; Zhang, Y.; Li, T. Sonar image coarse-to-fine few-shot segmentation based on object-shadow feature pair localization and level set method. *IEEE Sensors Journal*. **2024**, *24*, 12345-12356.
28. Wang, K.; Liu, S.; Xu, M. Real-time heterogeneous filtering with lightweight U-Net for side-scan sonar image segmentation. *IEEE Robotics and Automation Letters*. **2025**, *10*, 4567-4574.
29. Hinton, G.; Vinyals, O. Distilling the knowledge in a neural network. *Conference on Neural Information Processing Systems*. **2015**, 1-9.
30. Jacob, B.; Kligys, S. Quantization and training of neural networks for efficient integer-arithmetic-only inference. *IEEE Conference on Computer Vision and Pattern Recognition*. **2018**, 2704-2713.
31. Gupta, R.; Patel, N. Spline-based CNNs for interpretable medical imaging. *Medical Image Analysis*. **2022**, *80*, 102499.
32. Kurkova, V.; Sanguineti, M. Kolmogorov-Arnold networks: A survey. *Neural Networks*. **2018**, *103*, 127-135.
33. Unser, M.; Aziznejad, S. B-spline CNNs on Lie groups. *International Conference on Learning Representations*. **2020**, 1-15.
34. Dai, J.; Qi, H. Deformable convolutional networks. *IEEE International Conference on Computer Vision*. **2017**, 764-773.

35. Hayes, M.; Smith, P. SAS image reconstruction using deformable kernels. *IEEE Journal of Oceanic Engineering*. **2021**, *46*, 1104-1116.
36. Wang, H.; Xu, Y.; Zhang, L. Meta-learning for few-shot segmentation of low-resolution side-scan sonar images. *IEEE Transactions on Geoscience and Remote Sensing*. **2025**, *63*, 3050-3065.
37. Tegmark, M.; Liu, Z.; et al. KAN: Kolmogorov-Arnold Networks. *arXiv*. **2024**, 2404.19756.
38. Wang, C.; Liao, H.M.; Wu, Y.; Chen, P.; Hsieh, J.; Yeh, I. CSPNet: A New Backbone that can Enhance Learning Capability of CNN. *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. **2020**, 1571-1580.
39. Feng, C.; Zhong, Y.; Gao, Y.; Scott, M. R.; Huang, W. TOOD: Task-aligned One-stage Object Detection. *IEEE/CVF International Conference on Computer Vision (ICCV)*. **2021**, 10428-10437.
40. Bolya, D.; Zhou, C.; Xiao, F.; Lee, Y. J. YOLACT: Real-time instance segmentation. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. **2019**, 9157-9166.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.