

Article

Quasi-Static In Situ Deep Learning for Forward-Looking Sonar Target Detection in Complex Underwater Environments

Yixuan Chen ^{1,2,†}, Zhenqing Ding ^{1,2,†}, Yu Feng ^{1,2,3,†}, Jiale He ⁴, Ziqin Xie ^{1,5}, Tinggang Xiong ^{1,2}, Kai Chen ^{3,*} and Qi Gao ^{1,2,*}

¹ Wuhan Lingjiu Microelectronics Co., Ltd., No.1 Baihe Road, East Lake High-Tech Development Zone, Wuhan 430074, China; etch-yixuan.chen@foxmail.com (Y.C.); dingzq@ljmicro.cn (Z.D.); yufeng@smail.nju.edu.cn (Y.F.); ziqin.xie@foxmail.com (Z.X.); xtg_hb@aliyun.com (T.X.)

² Wuhan Digital Engineering Institute, Wuhan 430205, China

³ Key Laboratory of Modern Acoustics, Institute of Acoustics, Nanjing University, Nanjing 210093, China

⁴ Department of Electrical and Electronic Engineering, Guangdong Polytechnic College, Zhaoqin 526100, China; peanut-jiale.he@foxmail.com

⁵ Department of Computer Science, Hubei University of Technology, Wuhan 430068, China

* Correspondence: chen kai@nju.edu.cn (K.C.); gaoq@ljmicro.cn (Q.G.)

† These authors contributed equally to this work.

Abstract

Forward-looking sonar (FLS) target detection is essential for autonomous underwater vehicles (AUVs), yet its effectiveness is severely hindered by complex acoustic distortions, environmental volatility and the scarcity of fine-annotated data, which limit the success of standard deep learning approaches. To address these challenges, this study proposes a novel quasi-static in situ learning paradigm for underwater acoustic target detection (UATD). The hybrid methodology integrates scene priors into a lightweight deep learning detector by incorporating explicit probability weighting based on echo-intensity statistics and acoustic attenuation compensation. By using these models for pixel-wise image enhancement and fusing statistical descriptors with deep learning predictions at the score level, the framework dynamically adapts to in situ environmental contexts during quasi-static operational tranches. Experimental evaluations on the UATD dataset demonstrate that this in situ adaptation significantly enhances overall detection performance, achieving an F1-score of 0.865 for our approach, an 8.1% improvement over the baseline YOLOv12n, with only a 2.1 G increase in FLOPs, while outperforming YOLOv12x (F1 = 0.844) with 95% fewer FLOPs. Ultimately, this paradigm overcomes the limitations of purely deep learning-based methods, offering a robust and interpretable solution tailored for practical AUV deployment.

Keywords: underwater acoustic target detection; forward-looking sonar; sonar image enhancement; autonomous underwater vehicles; information fusion



Academic Editors: Dídac Diego
Tortosa, Umberta Tinivella and María
Campo Valera

Received: 6 April 2026

Revised: 13 May 2026

Accepted: 14 May 2026

Published: 16 May 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\) license](#).

1. Introduction

Forward-looking sonar (FLS) target detection has become increasingly important in underwater research due to the growing use of autonomous underwater vehicles (AUVs) that rely on FLS imagery for environmental perception and obstacle avoidance [1–6]. FLS has advantages over other active sensors in its ability to provide high-resolution real-time imagery in underwater environments. It outperforms optical sensors in terms of robustness under turbid conditions and at longer ranges [1,7], and outpaces side-scan sonar by offering a real-time wide-angle view [2,7]. Meanwhile, recent advances in deep

learning have substantially improved imagery-based detection tasks [8,9]. Deep learning appears to be a natural solution for FLS-imagery-based underwater acoustic target detection (UATD). Consequently, deep learning has been increasingly investigated for UATD tasks in recent years [10–18]. Zhang et al. presented a comprehensive survey that synthesizes recent advancements in UATD across multiple sonar modalities, thereby providing a contextualization of these contributions [19]. However, deep learning models have not achieved the same level of success in FLS-imagery-based UATD as in photographic image detection tasks. A key reason is that complex and time-varying underwater environments affect data acquisition and image formation, whereas deep learning models remain highly data-dependent.

Complex underwater environments pose significant challenges to data acquisition in FLS systems by distorting acoustic propagation. These environments are highly variable, with pronounced spatial and temporal fluctuations that make acoustic signals sensitive to local conditions. Consequently, FLS images often suffer from strong noise, low contrast, and poor consistency of target features, which constitute major bottlenecks for UATD tasks and the fine-grained annotation of FLS image datasets. Unlike the robust scalability of features in conventional photographic images—where semantics remain stable under scaling or rotation—object features in FLS images fail to maintain consistent semantics, impeding the training of effective deep learning models. Annotation efforts are further hindered by the difficulty in recognizing targets amid intense noise, and unlike photographic datasets, scaling techniques offer no remedial benefit here.

This difficulty is further exacerbated by the frequency-dependent nature of acoustic imaging in FLS systems: different operating frequencies can produce substantially different image patterns and target appearances, which may interfere with stable feature representation and therefore often require single-frequency operation. In contrast, multi-frequency light waves in the air yield varied colors while preserving consistent patterns in photographic images, enhancing object contrast and distinguishability through the simultaneous collection of optical signals. To illustrate these challenges, representative samples from the photographic Open Images V7 dataset [20] and the FLS UATD dataset [21] are presented in Figure 1. The comparison highlights typical difficulties in FLS imagery, including speckle noise, acoustic shadows, limited contrast, and ambiguous or complex target patterns. These arise from the intrinsic characteristics of underwater environments and FLS sensors, compounded by the influence of environmental factors and sonar frequency. Given the absence of viable alternative sensors, there is a pressing need for specialized approaches tailored to the unique features of underwater acoustics and FLS imaging.

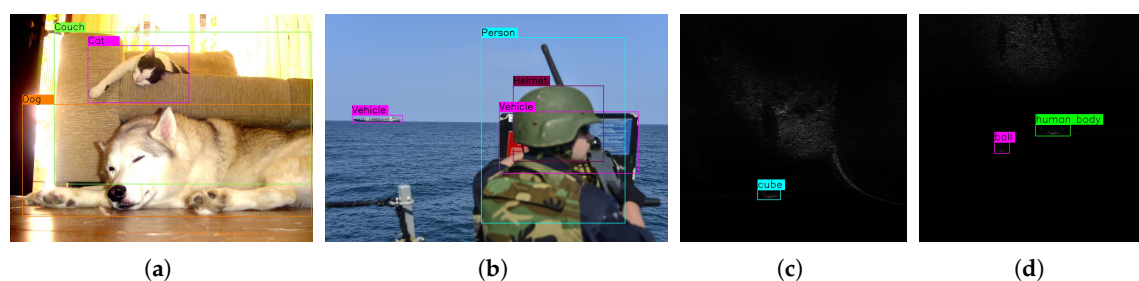


Figure 1. Examples illustrating the stark contrast between common photographic images (a,b) and FLS images (c,d).

Recent studies on UATD-specific methods can be broadly divided into two categories: customized deep learning models and statistical descriptors that characterize underwater environments and embedded objects. Regarding deep learning customization, existing studies have improved UATD accuracy by introducing specialized modules to handle noisy

and low-contrast FLS images [10–14]. Other studies have employed transfer learning and data augmentation to improve model generalization in UATD tasks [15–18]. Although these model-customization attempts have achieved some progress on certain datasets, the improvement remains limited because deep learning models are data-driven and fine-annotated FLS images are scarce [21]. Another limitation of deep learning models is their limited interpretability: examining their outputs or architectures often provides little insight into the factors that restrict detection performance. By contrast, statistical descriptors are interpretable and computationally inexpensive. Santos et al. [22,23] developed a descriptor for FLS images using a Gaussian probability density function graph that indicates information about marine environments and object features, and then conducted an object classification based on the descriptor. Avi et al. [24] performed UATD by leveraging statistical features of shadow regions in FLS images. Su et al. [25] noticed the distortions in FLS images caused by acoustic rolling shutter (ARS) effect, which hinders UATD tasks, and proposed a deep learning method to compensate for the ARS effect. However, statistical descriptors depend heavily on handcrafted features, which often generalize poorly across environmental variations and may fail to capture the deep abstract representations required for complex target recognition. Thus, existing studies face a fundamental trade-off: deep learning provides strong feature extraction but suffers from limited interpretability and dependence on scarce annotations, whereas statistical methods are transparent but lack sufficient expressive power under severe acoustic distortion and noise. A unified framework that combines the complementary advantages of both approaches is therefore still needed for FLS-based UATD.

To address the aforementioned limitations, we propose a dynamic hybrid approach that combines deep learning with dynamic physics-informed statistical modeling, drawing inspiration from prior studies on statistical information fusion and acoustic measurement theory [26–28]. In mature optical image classification, such hybridization is often valued primarily for computational efficiency rather than consistent accuracy gains [26]. We argue, however, that UATD benefits more substantially: explicit statistical descriptors can capture latent physical structure in sonar echoes, improving both detection accuracy and recall in cluttered, heterogeneous underwater scenes while keeping inference cost low. However, existing hybrid pipelines for optical imagery typically learn the statistical component once from historical data and keep it fixed during inference. This static parameterization is poorly suited to complex and volatile underwater acoustic environments. To address this issue, we introduce a dynamic in situ learning paradigm that models echo-intensity distributions in real time using a lognormal probability density function (PDF), together with an acoustic attenuation compensation model. The choice of lognormal PDF is theoretically and empirically supported by prior work [27,28]. Because our fused model remains lightweight, we can operate in a quasi-static in situ mode: when a specific task phase or underwater environment is relatively stable (i.e., quasi-static), we perform in situ data sampling to dynamically extract local statistical parameters and fine-tune a pre-trained deep learning model on the fly. Furthermore, we demonstrate that our statistical descriptions can be utilized to visually enhance FLS images, effectively highlighting object features and mitigating the difficulties associated with manual data annotation. The key contributions of this research are summarized as follows:

- We propose a quasi-static in situ paradigm for UATD that combines lightweight deep learning with statistical information fusion to adapt to local underwater environments. Unlike conventional approaches that rely on fixed parameters learned during training, our method injects contextual environmental cues online, improving detection robustness under varying acoustic conditions.

- We introduce statistical descriptors—specifically, an echo intensity distribution model and an acoustic attenuation compensation method—that explicitly capture object and environmental features to enhance deep learning model predictions. Compared to traditional methods, these descriptors simplify the fitting forms and application procedures, enabling more effective dynamic adjustments.
- We demonstrate that the proposed statistical descriptions can be applied to FLS image enhancement, making targets more visually salient and thereby facilitating manual annotation calibration and improving annotation reliability.

The remainder of this manuscript is organized as follows: Section 2 details the proposed methodology, introducing the novel statistical descriptors, the image enhancement technique, and the framework of the quasi-static in situ paradigm. Section 3 outlines the experimental setup, including the dataset, comparative baselines, ablation studies and evaluation metrics. Section 4 presents a comprehensive evaluation of the results, including visual comparisons, performance comparisons and ablation studies. Finally, Section 5 concludes the paper and discusses potential avenues for future research.

2. Proposed Methodology

This section describes the proposed methodology and its underlying rationale. We begin by modeling the statistical distribution of echo intensities in FLS images using log-normal PDFs, followed by the selection of an appropriate attenuation model to characterize underwater acoustic wave propagation and its compensation. Building on this statistical framework, we integrate statistical information into the UATD workflow by enhancing FLS images through pixel-wise statistical transformations and improving classification accuracy via a lightweight logistic-regression-based prediction ensemble that fuses statistical and deep learning outputs at the score level. Finally, we outline a quasi-static in situ workflow that integrates these components for practical deployment.

2.1. Statistical Modeling of FLS Images

2.1.1. Lognormal Modeling for Echo Intensity

In this subsection, we model the statistical behavior of echo intensities in FLS images. Let the observed echo intensity be represented by the image pixel value, which is monotonically related to the actual intensity. Previous studies demonstrate that under specific operating frequencies and environmental conditions, echoes from targets and their surrounding backgrounds exhibit distinct intensity distributions that separate them from clutter noise [22–24,27]. We classify echoes from targets as target-dependent and those from the backgrounds containing the targets as environment-dependent. Although the signal-to-noise ratio (SNR) is traditionally applied to waveform analysis rather than images, we use an image-domain SNR proxy here, defined as the ratio of target echo intensities to background echo intensities, and regard it as process-dependent. This designation reflects the acoustic propagation process: under ideal conditions with minimal interference, targets are easily distinguished from the background, whereas severe interference results in signals being aliased with noise.

Traditionally, underwater acoustic echo intensities are modeled using the Rayleigh distribution [27], assuming additive superposition of echoes from numerous small scatterers within a resolution cell, leading to a Rayleigh-distributed envelope via the central limit theorem (CLT). This holds for FLS background reverberation but not for high-resolution target echoes, where a single dominant scatterer deviates from the additive assumption.

For targets, we adopt the lognormal distribution, which captures multiplicative process effects and heavy tails, making it better suited to the high variability and peak structures of target echoes. Target reflections are influenced by cascading factors such as geomet-

ric spreading factors (g_i), surface properties (s_i), multipath interference (m_i), and environmental modulations (e_i). Received power P_r is thus P_t multiplied by these factors: $P_r = P_t \prod g_i \prod s_i \prod m_i \prod e_i \cdots$. Taking logarithms yields $\log P_r = \log P_t + \sum \log f_i$; assuming independent factors f_i , the CLT implies $\log P_r \sim \mathcal{N}(\mu, \sigma^2)$, so $P_r \sim \text{LogNormal}(\mu, \sigma^2)$. This holds for both power and envelope descriptions due to its intrinsic multiplicative feature. Empirical evidence supports lognormal's superiority for shallow-water, short-range FLS target modeling [27,28]. If the echo intensities of targets and their backgrounds are each modeled as lognormal distributed random variables and are assumed to be statistically independent, a supposition that is typically reasonable, then the resulting ratio, SNR, is itself lognormal distributed. This follows directly from the multiplicative closure property of the lognormal distribution, providing simplicity for analysis.

Empirical statistics from the UATD dataset provide further support for this framework. As shown in Figure 2, pixel values represent echo intensities for targets and backgrounds, with SNRs computed as their ratios. The dataset includes images collected under two different FLS operating frequencies, 720 kHz and 1200 kHz, which are analyzed separately in Figure 2a and Figure 2b respectively. In each case, the curves represent smoothed histograms of the samples, obtained by partitioning the data range into 64 equal-width bins and connecting middle points of bin-tops with antialiased interpolation. Given that raw FLS pixel values are inherently quantized into 256 discrete levels representing continuous echo magnitudes, this 64-bin resolution has a reasonable granularity and can effectively preserve the underlying distribution features without severe distortion. These empirical distributions reveal that the three random variables (RVs)—the target-dependent intensity values, the environment-dependent background values, and the process-dependent SNR values—cluster around distinct means and are clearly separable. This justifies our use of these RVs to formulate a statistical model for echo intensities in FLS images.

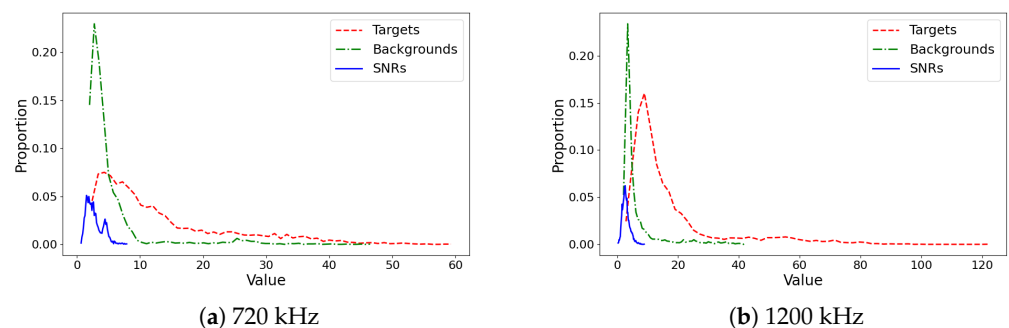


Figure 2. Smoothed histogram curves of echo intensities (pixel values) for targets and backgrounds, along with their SNR ratios, derived from images in the UATD dataset collected at different FLS operating frequencies: (a) 720 kHz and (b) 1200 kHz. These distributions highlight distinct central tendencies.

Having selected the relevant RVs and their distributions, we next specify the parameter-estimation procedure for them. Lognormal distribution has several advantageous properties: closed-form parameter estimation and closure under multiplication. Closed-form estimation enables real-time updates, which is essential for the in situ training workflow. Let $V > 0$ denote a generic intensity-related RV (target value, background value, or their ratio). Assuming:

$$V \sim \text{LogNormal}(\mu, \sigma^2) \iff Y = \ln V \sim \mathcal{N}(\mu, \sigma^2). \quad (1)$$

Accordingly, (μ, σ^2) can be updated online by maintaining the running mean and variance of $Y = \ln V$. When a new observation v_{i+1} arrives, define $y_{i+1} = \ln v_{i+1}$. Given current estimates (μ_i, σ_i^2) from i samples, the updated estimates are:

$$\mu_{i+1} = \frac{i\mu_i + y_{i+1}}{i+1} \quad (2)$$

$$\sigma_{i+1}^2 = \frac{i\sigma_i^2 + i(\mu_i - \mu_{i+1})^2 + (y_{i+1} - \mu_{i+1})^2}{i+1} \quad (3)$$

This update incurs $O(1)$ time and space overhead, thereby ensuring computational efficiency and providing the foundational basis for our in situ training workflow. Furthermore, closure under multiplication preserves the lognormal form and maintains these updating benefits. For example, if target and background intensities follow lognormal distributions, their ratio (SNR) also adheres to a lognormal distribution. This property is particularly useful for weighting operations via multiplication, as employed in our image enhancement approach detailed in Section 2.2.

To justify this selection, we evaluated other representative alternatives, specifically the Rayleigh and skew-normal distributions. While the Rayleigh distribution is classically used to model echo intensities and supports closed-form estimation, it lacks closure under addition and multiplication. Operations between Rayleigh-distributed RVs yield non-Rayleigh results, eliminating the benefits of closed-form estimation and complicating subsequent analysis. Conversely, the skew-normal distribution offers closure under addition but lacks a closed-form estimation method. Parameter estimation techniques, such as expectation maximization, scale poorly; our experiments demonstrated that for large datasets, estimating skew-normal parameters becomes computationally intensive, sometimes even unachievable. If the skewness is forced to zero, the skew-normal degrades into a standard Gaussian (normal) distribution, which provides closed-form estimation and additive closure—advantages exploited in prior works [22,23]. However, as evident in Figure 2, the empirical distributions exhibit clear skewness, and the symmetric Gaussian introduces substantial fitting errors. Empirical evaluations confirm that the lognormal distribution offers superior performance. To quantitatively validate the choice of lognormal distribution for modeling these RVs, we evaluated the goodness-of-fit against alternatives using Kullback–Leibler (KL) divergence. As presented in Table 1, the lognormal distribution consistently exhibits overall low KL divergence values across both frequencies, confirming its superior fit to the empirical data and supporting our initial rationale.

Table 1. KL divergence values assessing the goodness-of-fit for lognormal, Rayleigh, normal, and skew-normal distributions against the empirical distributions of targets, backgrounds and SNRs of samples in the UATD dataset. Lower values indicate better fit.

Random Variable	Lognormal	Rayleigh	Normal	Skew-Norm
720 kHz				
Targets	0.0280	0.9880	1.2243	0.3206
Backgrounds	0.3408	0.8512	1.1292	0.8554
SNRs	0.0526	0.0822	0.1722	0.0425
1200 kHz				
Targets	0.1774	1.3242	1.8528	1.0698
Backgrounds	1.0543	1.4403	2.2807	1.7222
SNRs	0.0473	0.0800	0.1190	0.0192

With the RVs, their distributions, and the parameter-estimation procedure specified, the statistical model for FLS echo intensities is complete. The parameter-estimation results

for the UATD dataset are presented in Section 4.1, while their subsequent applications to image enhancement and fusion with the deep learning model are discussed in Section 2.2.

2.1.2. Modeling of Acoustic Wave Attenuation

Accurate characterization of underwater acoustic wave attenuation is essential for establishing our model and has been extensively researched [29–37]. Prior studies have focused on characterizing the attenuation coefficient A , which relates the received power P_r to the transmitted power P_t as follows:

$$P_r = \frac{P_t}{A} \quad (4)$$

Traditionally, existing studies decompose A into two components: the absorption coefficient a , which accounts for energy dissipation, and the geometric spreading coefficient k , which captures spreading losses. The conventional formulation is

$$10 \log A = k \cdot 10 \log l_m + l \cdot 10 \log a \quad (5)$$

where l is the one-way propagation distance in kilometers, and $l_m = 1000l$.

Spreading loss constitutes the most fundamental category of acoustic attenuation and represents the geometric effect by which the intensity of a sound signal progressively decreases as it propagates away from its source, and the associated parameter k is correspondingly straightforward to specify [37]. For an unbounded medium in which acoustic energy radiates approximately uniformly in three dimensions, intensity decays as $\frac{1}{r^2}$, corresponding to spherical spreading and $k = 2$. In contrast, when propagation is strongly constrained by parallel boundaries, e.g., efficient trapping between surface and bottom over many wavelengths, intensity decays approximately as $\frac{1}{r}$, corresponding to cylindrical spreading and $k = 1$. In FLS operations from an AUV, ranges are typically short (sub-kilometer) and frequencies are high; over such distances, the acoustic field is often not strongly waveguided, and boundary-induced trapping is not fully developed. We therefore adopt spherical spreading as the default baseline model and set $k = 2$ for one-way attenuation, while noting that site-specific bathymetry, sound-speed structure, and multipath can yield deviations from this idealized value.

Approaches to modeling the absorption coefficient a vary across studies and are often highly complex. Previous studies have developed comprehensive empirical formulas incorporating numerous environmental factors to maximize model generalization. A representative example is the Fisher & Simmons model [30], which formulates the absorption coefficient a as a function of wave frequency f (in kHz), water temperature t (in °C), and depth d (in m):

$$10 \log a(f, t, d) = A_1 P_1 \frac{f_1 f^2}{f_1^2 + f^2} + A_2 P_2 \frac{f_2 f^2}{f_2^2 + f^2} + A_3 P_3 f^2 \quad (6)$$

The output of Equation (6) is expressed in dB/km, and the constituent empirical parameters are detailed in Table 2. The Fisher & Simmons formulation and related extensions are representative of a broader trend in underwater-acoustic attenuation modeling: progressively more elaborate parameterizations have been introduced to incorporate additional dependencies such as salinity/conductivity, temperature, depth (measured by CTD sensors) and frequency-dependent solute interactions. While such formulations can improve generalization capability across operating conditions, their increasing complexity also reduces transparency and makes rapid adaptation to changing environments difficult. This evolution mirrors that of deep learning models, which grow larger with numerous pa-

rameters to achieve superior fitting and generalization capabilities over traditional methods, albeit at the expense of interpretability.

Table 2. Empirical formulas for constituent parameters of absorption coefficients $a(f, t, d)$ in the Fisher & Simmons model.

Empirical Parameters of $a(f, t, d)$
$A_1 = 1.03 \times 10^{-8} + 2.36 \times 10^{-10} - 5.22 \times 10^{-12}t^2$
$A_2 = 5.62 \times 10^{-8} + 7.52 \times 10^{-10}t$
$A_3 = (55.9 - 2.37t + 4.77 \cdot 10^{-2}t^2 - 3.48 \times 10^{-4}t^3) \times 10^{-15}$
$f_1 = 1.32 \times 10^3 \cdot (t + 273.1)e^{\frac{-1700}{t+273.1}} - 3.48$
$f_2 = 1.55 \times 10^7 \cdot (t + 273.1)e^{\frac{-3052}{t+273.1}}$
$P_1 = 1$
$P_2 = 1 - 10.3 \times 10^{-5}d + 3.7 \times 10^{-9}d^2$
$P_3 = 1 - 3.84 \times 10^{-5}d + 7.57 \times 10^{-10}d^2$

Our observation is that, for inherently complex and analytically intractable phenomena, delegating modeling to deep learning may be more effective than manually complicating static parameter expressions. Conversely, interpretable equations should prioritize simplicity, with dynamic parameter adjustments to complement the static, hard-to-tune nature of deep learning models. However, we argue that pursuing universally comprehensive static models is suboptimal for dynamic underwater environments. Motivated by this observation, we avoid further complicating the closed-form attenuation expression and instead adopt an in situ modeling paradigm in which only the interpretable physical structure is retained and a small set of coefficients is adjusted to the local environmental context. Accordingly, we preserve the fundamental relationships in Equations (4) and (5) and derive the proposed in situ attenuation formulation from the conventional logarithmic expression. Starting from Equation (5), applying the logarithmic identity gives

$$A = l_m^k a^l. \quad (7)$$

Equation (7) shows that the conventional attenuation coefficient consists of two multiplicative components: a range-dependent spreading term l_m^k and an absorption term a^l . Since $l_m = 1000l$, the constant unit-conversion factor can be absorbed into an empirical scale coefficient. Moreover, the absorption term can be expressed in exponential form as

$$a^l = e^{(\ln a)l}. \quad (8)$$

Therefore, the conventional attenuation model can be represented as a product of a power-law spreading term and an exponential absorption-related term. For near-range FLS measurements, directly using l^k may cause numerical instability or overcorrection near $l = 0$. We therefore introduce the shifted range term $(l + 1)^k$, while absorbing unit conversion and empirical scaling effects into λ . The resulting in situ attenuation formulation is

$$A(l) = \lambda(l + 1)^k e^{\alpha l} \quad (9)$$

where $\alpha \propto \ln a$ and $\ln \lambda \propto k$. This simplified multiplicative formulation provides a convenient basis for subsequent parameter estimation procedures.

When examining specific FLS images for parameter estimation, it is important to account for the fact that, in contrast to one-way propagation, FLS imaging involves a

two-way path (outgoing and returning echoes). Let $I(l)$ denote the measured pixel intensity at range l , and let $R(l)$ denote the effective reflection/backscatter coefficient of the insonified region. Under fixed sonar settings (working mode, P_t , and frequency), the image intensity can be modeled as

$$I(l) = \frac{C P_t R(l)}{[A(l)]^2} \quad (10)$$

where C is a positive scaling constant, P_t is the transmission power (assumed constant for a given working mode), and $R(l)$ is the reflection coefficient at distance l . The denominator $[A(l)]^2$ accounts for the two-way attenuation over distance l . For the FLS-based UATD scenario considered in this work, we average a large number of measurements at the same range. Consequently, the background reflection dominates. Under a quasi-static assumption, the average reflection coefficient $\bar{R}$ can be treated as a constant invariant to distance. By calculating the average pixel intensities $\bar{I}$ at two distinct distances l_1 and l_2 ($l_1 < l_2$), the relative attenuation between them can be directly estimated:

$$\frac{A(l_2)}{A(l_1)} = \sqrt{\frac{\bar{I}(l_1)}{\bar{I}(l_2)}} \quad (11)$$

Using a series of these empirical estimations across various distances, we fit Equation (9) to obtain the optimal environmental parameters $(\hat{\lambda}, \hat{k}, \hat{\alpha})$ via the Levenberg–Marquardt algorithm [38]. Once these parameters are derived, the resulting in situ attenuation model is subsequently applied to enhance the FLS images and is fused with the deep learning detector.

2.2. Statistical Information Fusion

In this work, incorporating the statistical information derived from the modeling procedures into the UATD workflow involves two primary components: FLS image enhancement and a classification prediction ensemble. FLS image enhancement utilizes statistical models to transform pixel values, thereby suppressing noise, highlighting targets, and improving both visual quality and subsequent deep learning model performance. The classification prediction ensemble employs logistic regression on statistical features to perform statistical prediction, incurring minimal overhead while combining with deep learning results at the score level to boost overall accuracy.

2.2.1. FLS Image Enhancement

As described in Section 2.1, we model the echo intensity statistics using lognormal PDFs, while underwater acoustic wave attenuation is modeled using Equation (9). Based on these models, we generate a probability-weighted map from the original image, which suppresses background or speckle-like returns while emphasizing statistically salient echoes associated with targets. Concurrently, the attenuation compensation adjusts for signal loss over distance to better reflect intrinsic target reflectivity. Finally, a channel-wise concatenation integrates these distinct representations: the raw data (channel 1), the attenuation-compensated data (channel 2), and the probability-weighted data (channel 3). This channel-wise concatenation fusion strategy produces pseudo-colored images that are visually more informative and improve downstream detector training and performance.

In the following, we first describe in detail the procedure used to derive the probability-weighted map. Let a fitted PDF be denoted by $f(x | \hat{\mu}, \hat{\sigma}^2)$. Although a digital FLS image discretizes intensities into 256 bins of fixed width Δx , the subsequent normalization back to $[0, 255]$ makes the multiplicative constant Δx irrelevant. Therefore, we directly use the PDF values as weights and denote them as probabilities. The selected random variables (RVs)

include target echo intensities, background intensities, and their ratio SNRs. We denote the corresponding probabilities as (p_1, p_2, p_3) :

- p_1 : Target-intensity confidence, obtained by evaluating the pixel intensity I under the fitted target-intensity PDF $f(x | \hat{\mu}_1, \hat{\sigma}_1^2)$. This typically suppresses weak diffuse noise as well as some strong localized clutter. However, a limited number of target returns whose intensities deviate substantially from the mean $\hat{\mu}_1$, as well as noise realizations whose intensities lie close to the mean, may be assigned inappropriate confidence levels.
- p_2 : Contextual confidence, obtained by evaluating the local background mean $\bar{I}_b$ under the fitted background-intensity PDF $f(x | \hat{\mu}_2, \hat{\sigma}_2^2)$. This can enhance true targets but may also raise noise in the same background.
- p_3 : Contrast-based confidence, obtained by evaluating $\text{SNR} = \frac{I}{\bar{I}_b}$ under the fitted SNR PDF $f(x | \hat{\mu}_3, \hat{\sigma}_3^2)$. This can suppress low-reflectivity noise and large-area interference (e.g., seabed returns) but may boost small-area noise peaks and harm weak targets.

We fuse these cues via a weighted geometric mean to preserve multiplicative relationships among PDFs and balance the complementary effects of the RVs:

$$I_{pw} = \text{normalize} \left(I \cdot (p_1^{w_1} p_2^{w_2} p_3^{w_3})^{\frac{1}{3}} \right), \quad w_1 + w_2 + w_3 = 3 \quad (12)$$

Optimal weights are determined via an exhaustive search over all possible combinations with a step size of 0.1, as finer steps yield negligible improvements.

For attenuation compensation, we estimate the range-dependent prior attenuation rate $\hat{A}(I | \hat{\lambda}, \hat{k}, \hat{\alpha})$ using fitted parameters and Equation (9), converting raw intensities to reflectivity-dependent values and normalizing to $[0, 255]$:

$$I_{ac}(I) = \text{normalize} \left(I(I) [\hat{A}(I)]^2 \right) \quad (13)$$

This attenuation compensation tends to strengthen distant targets that are attenuated by propagation while suppressing near-source speckle-like noise with low reflectivity.

The three-channel enhanced images are then constructed by concatenation $I_{\text{fused}} = [I_{\text{raw}}, I_{\text{ac}}, I_{\text{pw}}]$. The preservation of original information contained in the raw images ensures that when these fused images are fed into a deep learning detector, the resulting performance is at least not inferior to that obtained using the raw inputs.

2.2.2. Statistical Classification and Score-Level Ensemble

Deep learning object detectors, such as YOLO [39], consist of bounding box components $([p_c, b_x, b_y, b_h, b_w])$ and class scores $([c_1, c_2, \dots])$. Because the extracted statistical models do not contain spatial information suitable for bounding box regression, we utilize them exclusively to refine the classification outputs.

To achieve this, we construct a 12-dimensional statistical feature vector for each predicted target region:

$$\mathbf{V} = [\bar{I}_t, \sigma_t^2, \bar{I}_b, \text{SNR}, \hat{\mu}_1, \hat{\sigma}_1^2, \hat{\mu}_2, \hat{\sigma}_2^2, \hat{\mu}_3, \hat{\sigma}_3^2, \bar{I}_{pw}, \bar{I}_{ac}] \quad (14)$$

The components of this vector capture various regional characteristics: $\bar{I}_t$ and σ_t^2 denote the mean and variance of the target intensities, respectively, while $\bar{I}_b$ represents the mean intensity of the surrounding background. The terms $\hat{\mu}_i$ and $\hat{\sigma}_i^2$ ($i \in \{1, 2, 3\}$) represent the updated parameters of the available lognormal probability density functions. Finally, $\bar{I}_{pw}$ and $\bar{I}_{ac}$ denote the mean values of the target region extracted from the probability-weighted and attenuation-compensated maps, respectively.

A logistic regression classifier is then trained on the indicative vectors $\mathbf{V}$ via Broyden–Fletcher–Goldfarb–Shanno algorithm [40], producing a class-score vector aligned with the detector’s class outputs. Finally, we combine the statistical and deep learning class-score vectors by simple averaging, a fusion strategy reported to be both efficient and effective in [26].

2.3. The Quasi-Static In Situ Workflow

This workflow treats the UATD task as a sequence of quasi-static tranches, each typically lasting several hours, during which sonar frequency, operating range, depth, target types, and environmental conditions remain approximately stable. Rapid, drastic changes are rare and unaddressable, but for typical scenarios, the quasi-static assumption holds. Importantly, the tranche boundary is dynamically determined in a data-driven manner by an online stability test rather than by assuming that stability persists for some fixed time interval lasting several hours. Within a tranche, the fitted statistical parameters of lognormal PDFs $(\hat{\mu}_i, \hat{\sigma}_i^2)$ are updated online and used as a stability indicator. If, over the most recent 30 min, the interval of $(\hat{\mu}_i, \hat{\sigma}_i^2)$ persistently drifts from its initial (first 30 min) tranche interval such that their intersection-over-union (IoU) falls below 0.5, we terminate the current tranche and start a new one. A new tranche triggers re-initialization and in situ retraining of the deep learning detector from a pre-trained model, followed by redeployment.

A key requirement of the proposed workflow is that tranche initialization and in situ adaptation be completed within 30 min on a portable computing platform. This constraint reflects the operational context of UATD in maritime environments, where high-performance servers are typically unavailable. The 30 min window is also an empirical compromise: it is long enough to collect sufficient in situ samples for reliable estimation of the tranche-level statistics and to perform effective detector fine-tuning, yet short enough to avoid consuming a substantial fraction of a tranche. If training were to occupy a large portion of the tranche duration, the workflow would be impractical because little time would remain for producing usable detection outputs. Therefore, the detector must be lightweight and capable of rapid convergence during in situ training. Many SOTA methods achieve superior accuracy primarily through increased model capacity and computational cost, making them unsuitable for the above constraint. In our experiments, YOLOv12n [41] best satisfies the deployment requirements due to its compact design and transformer-based architecture, which can benefit from efficient attention implementations such as FlashAttention [42]. Although YOLOv12n alone is outperformed by larger models in offline comparisons, its fast adaptation makes it effective within the proposed quasi-static in situ training pipeline. When combined with the statistical monitoring and re-initialization mechanism, the overall system attains SOTA performance and improved generalization in the targeted operating conditions.

Under the above design choices, the proposed system operates in repeated tranche cycles, each comprising data collection, context characterization, and rapid model adaptation. Specifically, the operational procedure of each tranche unfolds as follows: we sample images, perform preliminary annotation using the streaming annotation tool provided in [21], and fit statistical models to obtain statistical descriptors for the tranche context. The resulting statistical information is then applied to enhance both sampled and subsequent FLS images, improving annotation efficiency and calibration quality. The calibration process requires prior knowledge of the targets’ physical dimensions. By examining visual patterns in the enhanced images and assessing whether the sizes inferred from bounding boxes substantially deviate from the targets’ physical dimensions, we determine the need to revise prior preliminary annotations and apply necessary corrections. The refined annotations are used to train an in situ detector specialized to the tranche context, while the raw

images and final annotations are deposited into the dataset for later incremental tuning of the pre-trained model to improve its generalization ability and, consequently, facilitate rapid convergence of the in situ detector initialized from this pre-trained model. The full pipeline is summarized in Figure 3.

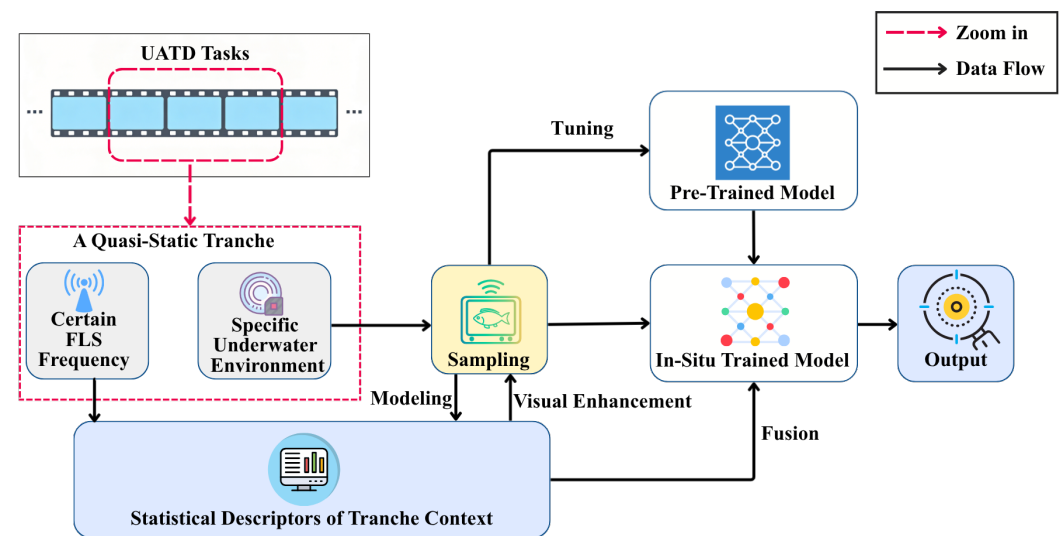


Figure 3. Overview of the quasi-static in situ paradigm. The framework highlights the adaptation loop, demonstrating the fusion of dynamically extracted statistical descriptions with deep learning features to overcome environmental volatility.

3. Experimental Setup

This section delineates the experimental setup designed to evaluate the efficacy of our proposed quasi-static in situ paradigm for FLS image enhancement and UATD tasks. To validate our methodology, we empirically assess the integration of pixel-wise statistical transformations, rooted in lognormal echo intensity distributions and acoustic wave attenuation compensation, into the deep learning pipeline. Furthermore, we design a series of comparative analyses and ablation studies to evaluate the effectiveness of each procedure introduced in our framework. The subsequent subsections detail the dataset, evaluation metrics, main comparative studies, and comprehensive ablation studies necessary to justify the specific statistical models employed in our workflow.

3.1. Dataset and Implementation Details

To evaluate our proposed quasi-static in situ paradigm, we utilize the publicly available UATD dataset [21], selected due to confidentiality constraints on proprietary data. This dataset comprises 9200 FLS images, each containing at least one annotated object, organized in temporal sequence. We treat the dataset as two quasi-static tranches based on the FLS operating frequencies: 2900 images at 720 kHz and 6300 images at 1200 kHz. As detailed in [21], other environmental conditions during data collection remain relatively stable, enabling us to simulate quasi-static scenarios by processing the images in chronological order.

The dataset includes annotations for objects across 10 categories, with their distribution summarized in Table 3. Notably, the object distributions are uneven across frequencies, with the 720 kHz tranche containing fewer samples overall and no instance of remotely operated vehicle (ROV). To maximize the challenge for our framework, where limited prior information is available before transitioning to a new tranche, we initiate the simulation with the 720 kHz tranche and proceed to the 1200 kHz tranche.

Table 3. Distribution of object categories in the UATD dataset across FLS operating frequencies (720 kHz and 1200 kHz). ROV stands for remotely operated vehicle.

Object Category	Total	720 kHz	1200 kHz
Cube	2987	1035	1952
Ball	3463	1072	2391
Cylinder	657	129	528
Human body	1434	501	933
Tyre	1368	561	807
Circle cage	860	160	700
square cage	1318	346	972
Metal bucket	487	346	141
Plane	1065	601	464
ROV	1000	0	1000

Within each tranche, we split images into training, validation and test sets with an 8:1:1 ratio, employing conventional shuffling and random sampling to align with evaluation protocols in prior works. This approach mitigates potential temporal leakage concerns arising from the sequential nature of the data. Annotation calibration, as described below, is performed only on the training split. The validation and test annotations are kept unchanged to ensure a unified evaluation standard and to maintain comparability with prior methods that do not modify labels. For subsequent training, previous methods do not incorporate temporal context, thus obviating the need for sequence restoration. In contrast, our proposed method utilizes temporal information, which can be restored using the order numbers in the annotations. Models are trained on the calibrated training split, tuned on the fixed validation split, and reported on the fixed test split.

In addition to bounding boxes and class labels, the dataset provides sufficient metadata, such as sonar range and azimuth, to estimate the distance (slant range) of each image row and an approximate physical size of the object's reflective region. These estimates are used to support our statistical modeling and to sanity-check label quality during calibration. An example annotation is presented in Table 4. Although the annotation includes an elevation angle, this value represents the overall elevation angle setting of sonar rather than per-pixel elevation information. Consequently, FLS images in the UATD dataset are not provided in a range–azimuth–elevation data cube. Instead, as is normal for most FLS images, the imagery is stored in a range–azimuth matrix. Furthermore, our empirical examination confirms that the three image channels are practically identical, yielding no additional spatial cues such as target elevation.

In the following, we provide a more detailed specification of the estimation procedure utilizing the provided metadata. Let R denote the maximum imaging range (meters) of an FLS frame, and let H denote the image height in pixels. Using 0-based pixel indexing along the vertical axis ($y \in \{0, \dots, H-1\}$), we map the distance l of row y as

$$l = R \cdot \frac{y+1}{H}. \quad (15)$$

For an object with bounding box corners $(x_1, y_1) - (x_2, y_2)$ ($x_1 \leq x_2$ and $y_1 \leq y_2$) on an image of width W and height H , and with sonar horizontal field of view (azimuth) Θ in degrees, the reflective surface dimensions, width W_r and height H_r , are estimated as:

$$W_r = R \cdot \frac{y_2 + y_1}{2H} \cdot \frac{\theta\pi}{180} \cdot \frac{x_2 - x_1 + 1}{W}, \quad (16)$$

$$H_r = R \cdot \frac{y_2 - y_1 + 1}{H} \quad (17)$$

These expressions follow a standard small-angle approximation that maps pixel spans to metric extents using the provided range and azimuth. These estimates are compared against the physical dimensions of objects, listed in Table 5, to validate and calibrate annotations.

Table 4. Example annotation instance from the UATD dataset.

<i>Metadata Category</i>	<i>Metadata Item</i>	<i>Instance Value</i>
Time Parameter	<i>Series Number</i>	707
	<i>Range (m)</i>	13.5033
Sonar Parameters	<i>Azimuth (°)</i>	120
	<i>Elevation (°)</i>	12
	<i>Sound speed (m/s)</i>	1467.2
	<i>Frequency (kHz)</i>	1200
	<i>Width (pixels)</i>	1024
Image Dimensions	<i>Height (pixels)</i>	1387
	<i>Number of Channels</i>	3
Object Information	<i>Object Type</i>	$obj_1 = ball$ $obj_2 = square\ cage$
	<i>Bounding Box</i>	$obj_1 = (504,912) - (574,967)$
		$obj_2 = (572,967) - (574,967)$

Table 5. Physical dimensions (in meters) of objects in the UATD dataset. Dimensions are reported using the notation L (length), W (width), H (height), and R (radius).

Object	Dimensional Scale
Cube	$(L, W, H) = (0.5, 0.5, 0.5)$
Ball	$R = 0.25$
Cylinder	$(R, L) = (0.1, 0.5)$
Human body	$(L, W, H) = (0.5, 0.2, 1.5)$
Plane	$(L, W, H) = (1.0, 1.0, 0.5)$
Circle cage	$(R, L) = (0.2, 0.15)$
Square cage	$(L, W, H) = (0.4, 0.4, 0.15)$
Metal bucket	$(R, H) = (0.3, 1.2)$
Tyre	$(R, W) = (0.33, 0.23)$
ROV	$(L, W, H) = (0.45, 0.35, 0.25)$

Given (W_r, H_r) and the known object dimensions, we flag training annotations as implausible when the estimated reflective extent deviates substantially from any feasible view-dependent projection of the object. Flagged labels are then refined during the proposed visually enhanced annotation stage, where improved contrast and pseudo-color provide clearer cues for boundary adjustment. This procedure leverages existing annotations as the preliminary annotations, while avoiding any modification to validation/test labels.

3.2. Comparative Analysis and Ablation Studies

To comprehensively evaluate the performance of our proposed quasi-static in situ training workflow (referred to as QSIS in illustrations), we conduct comparative analyses against several SOTA object detection baselines, specifically YOLOv12 [41], Detection with Transformer (DETR) [43], and Fast Region-based Convolutional Network (FRCN) [44]. Since both YOLOv12 and DETR provide model variants with different capacities, we report results for multiple sizes to contextualize the accuracy-to-complexity trade-off. Specifically, we evaluate the nano (YOLOv12n), medium (YOLOv12m), and extra-large (YOLOv12x)

variants of YOLOv12 (collectively denoted as the YOLOv12 Series), as well as DETR with ResNet-50 and ResNet-101 backbones (denoted as the DETR Series).

To minimize confounding factors, each baseline is trained using the default hyperparameters and training recipes from its official implementation, without manual hyperparameter tuning for UATD. All methods use the same tranche-wise train/validation/test splits described in Section 3.1. Furthermore, to mitigate the impact of the stochasticity inherent in deep learning optimization, all reported quantitative results are averaged over five independent training-and-test runs. These protocols are also employed in subsequent ablation studies.

A series of ablation studies are further conducted to isolate and validate the individual contributions of the statistical components introduced in our framework, namely, probability weighting based on the echo-intensity distribution and acoustic attenuation compensation. Controlled ablations are conducted by removing individual sources of statistical information. Concretely, for a given component, we set the corresponding entries of the statistical indicative vector $\mathbf{V}$ to zero and replace the associated fused-image channel with the raw FLS intensity channel. Since the raw multi-beam FLS images in the dataset consist of three mostly identical sensor channels, originating from redundant sensing intended to reduce measurement noise, we initially compute their mean to collapse them into a single raw acoustic channel. In addition, we ablate the training-portion-only annotation calibration step to quantify whether the proposed visual enhancement provides additional cues that improve label quality and downstream detection.

The ablation variants are summarized as follows and are collectively referred to as the YOLOv12n-QSIS-A Series:

- YOLOv12n-QSIS-nc (no calibration of annotations): Omits the visually enhanced annotation calibration process to demonstrate its necessity in generating reliable ground truth.
- YOLOv12n-QSIS-nac (no attenuation compensation): Disables the acoustic wave attenuation compensation.
- YOLOv12n-QSIS-npw (no probability weighting): Disables the incorporation of echo intensity distribution.

A comprehensive schematic of the experimental workflow, illustrating both the comparative analysis and the ablation studies, is presented in Figure 4.

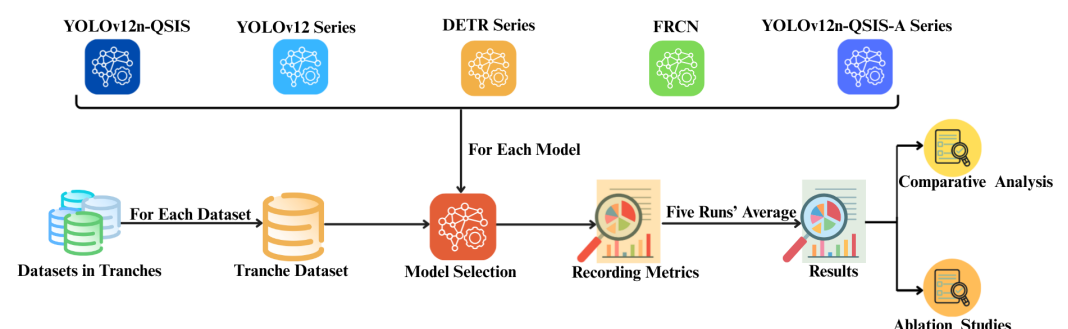


Figure 4. Schematic overview of the experimental workflow for the comparative analysis and ablation studies. The diagram delineates the evaluation pipeline comparing the proposed framework, YOLOv12n-QSIS, against established SOTA baselines (YOLOv12 Series, DETR Series, and FRCN), alongside the systematically ablated variants (YOLOv12n-QSIS-A Series) designed to isolate the contributions of acoustic compensation, probability weighting and annotation calibration.

3.3. Evaluation Metrics

Models from the comparative and ablation studies are evaluated using a comprehensive set of metrics tailored to object detection tasks, focusing on accuracy, efficiency, and

computational complexity. Detection accuracy is assessed via precision, recall, and F1-score, computed at a bounding box IoU confidence threshold of 0.5, as is standard in most practical applications. Although the UATD dataset is considered to comprise two quasi-static tranches associated with different sonar frequencies in our workflow, to maintain consistency with standard evaluation protocols and ensure a fair comparison with conventional state-of-the-art (SOTA) baselines, all quantitative metrics are computed and reported as overall performance aggregates across the entire test dataset, using a weighted averaging scheme in which the contribution of each tranche is proportional to its number of samples. For efficiency and scalability, we measure inference-time computational demands using floating-point operations (FLOPs) and model parameter counts, which indicate model size and potential deployment feasibility on resource-constrained underwater systems.

4. Results and Analysis

This section reports empirical evidence for the proposed quasi-static in situ workflow on the UATD dataset, focusing on whether the injected pixel-wise statistical modeling (echo-intensity probability weighting and acoustic attenuation compensation) yields measurably improved FLS imagery and more reliable supervision, and whether these gains translate into stronger and more robust UATD performance.

4.1. Statistical Enhancement and Calibration Results

The statistical parameters estimated from the training split according to Equations (1), (9) and (12) are summarized in Table 6. Overall, the probability weighting parameters fall within a plausible range, whereas the fitted attenuation parameters differ markedly from those of conventional sonar attenuation models. In particular, the fitted spreading loss coefficient k becomes negative, although it is typically constrained to be positive in traditional formulations where intensity is assumed to decrease monotonically with range.

Table 6. Estimated statistical parameters of the proposed probability weighting model and attenuation compensation model fitted on the training split.

Parameter Category	720 kHz	1200 kHz
Lognormal Distribution Parameters		
Target Intensity	$(\hat{\mu}_1, \hat{\sigma}_1^2) = (2.300, 0.587)$	$(\hat{\mu}_1, \hat{\sigma}_1^2) = (2.552, 0.521)$
Background Intensity	$(\hat{\mu}_2, \hat{\sigma}_2^2) = (1.392, 0.345)$	$(\hat{\mu}_2, \hat{\sigma}_2^2) = (1.587, 0.419)$
SNR (target-to-background ratio)	$(\hat{\mu}_3, \hat{\sigma}_3^2) = (0.903, 0.218)$	$(\hat{\mu}_3, \hat{\sigma}_3^2) = (0.964, 0.195)$
Probability Weights	$(w_1, w_2, w_3) = (0.2, 0.8, 2.0)$	$(w_1, w_2, w_3) = (0.2, 0.8, 2.0)$
Attenuation Model Parameters	$(\hat{\lambda}, \hat{k}, \hat{\alpha}) = (1.706, -1.758, 0.517)$	$(\hat{\lambda}, \hat{k}, \hat{\alpha}) = (1.252, -1.206, 0.517)$

This behavior is explained by the empirical range-intensity trend shown in Figure 5. In the context of the UATD dataset, near-distance echo intensities initially increase due to near-field effects and measurement artifacts (e.g., speckle and acoustic shadowing) before eventually decaying. Such non-monotonic behavior cannot be captured by traditional attenuation models that enforce a globally decreasing profile. If a conventional monotonic model were used for attenuation compensation, the near-source noise would be erroneously amplified rather than suppressed, which is detrimental for both visualization and downstream detection. Our model accurately captures this initial intensity surge, effectively mitigating near-source artifacts. Notably, the short-range intensity increase discussed herein is an observation restricted to the UATD dataset under its particular acquisition geometry and operating conditions, and therefore should not be construed as a universal property of FLS

imagery. Rather, it indicates that the proposed approach is capable of adapting to such specific scenarios.

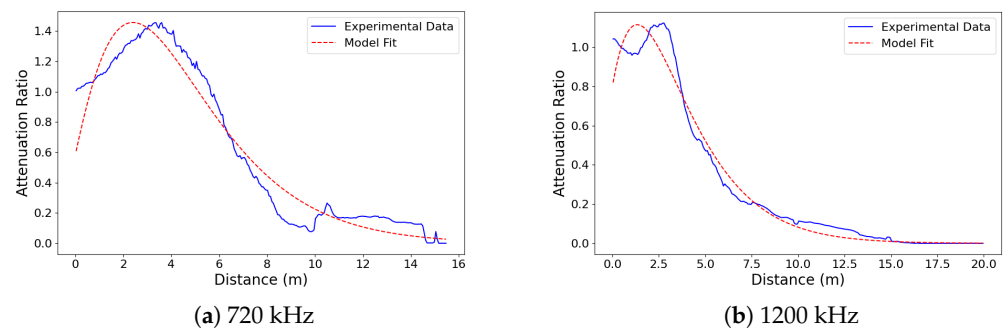


Figure 5. Curve fitting of the attenuation model over distance. The graph illustrates a short near-range intensity rise is observed before the expected decay, justifying the negative spreading cost parameters.

To provide an intuitive understanding of the enhancement process, Figure 6 presents a visual comparison between raw FLS images and the outputs of each enhancement component. The proposed enhancement improves perceptual target saliency while suppressing background clutter, which benefits both deep detectors and manual annotation during dataset construction. In Figure 6, the left and bottom axes show pixel height and pixel width, respectively, while the right and top axes indicate range (m) and azimuth (degree). For all subfigures, the horizontal axis spans pixel coordinates from 0 to 1020 (left to right), corresponding linearly to azimuth angles from -60 degrees to 60 degrees. For subfigures (a) (first row) and (c) (third row), the vertical axis spans pixel coordinates from 0 to 1020 (top to bottom), corresponding linearly to ranges from 0 m to 10 m. In contrast, subfigure (b) (second row) spans pixel coordinates from 0 to 1540 (top to bottom), corresponding linearly to ranges from 0 m to 15 m. The gridlines uniformly partition both axes into 10 equal intervals to facilitate position reading.

Applying attenuation compensation alone (Figure 6, column 2) improves target visibility by correcting range-related decay, but it can also amplify distant large-area high-intensity noise, revealing a limitation of using only the attenuation model.

The probability weighting enhancement (column 3) effectively suppresses uniform background noise. However, it may tend to enhance the tails of small, high-intensity noise and weaken portions of the target signal, creating defective areas of unexpectedly low intensity within targets. Additionally, it sometimes amplifies weak background noise in rows containing targets, resulting in horizontal artifacts.

The fused result (column 4) effectively integrates the strengths of both channels. The signal degradation (defective areas) introduced by the probability distribution model is compensated by the signal preservation of the attenuation channel. Furthermore, distinct noise patterns generated by the individual filters manifest as unique color artifacts in the fused image, making them easily distinguishable from the consistent color patterns of actual targets.

The examples in Figure 6 illustrate the complementary effects between attenuation compensation and probability weighting. In Figure 6a (first row), attenuation compensation effectively enhances the target while avoiding noise amplification for this scene, whereas probability weighting substantially boosts background noise, particularly patchy artifacts. In Figure 6b (second row), the overall weak echo signals limit the target enhancement from attenuation compensation, but probability weighting prominently highlights the target. In Figure 6c (third row), attenuation compensation yields a sufficiently clear target but fails to suppress large-area noise, which is effectively mitigated by probability weighting. Channel-

wise fusion of these results enables effective complementarity, while pseudo-coloring provides targets with a more pronounced and consistent pattern.

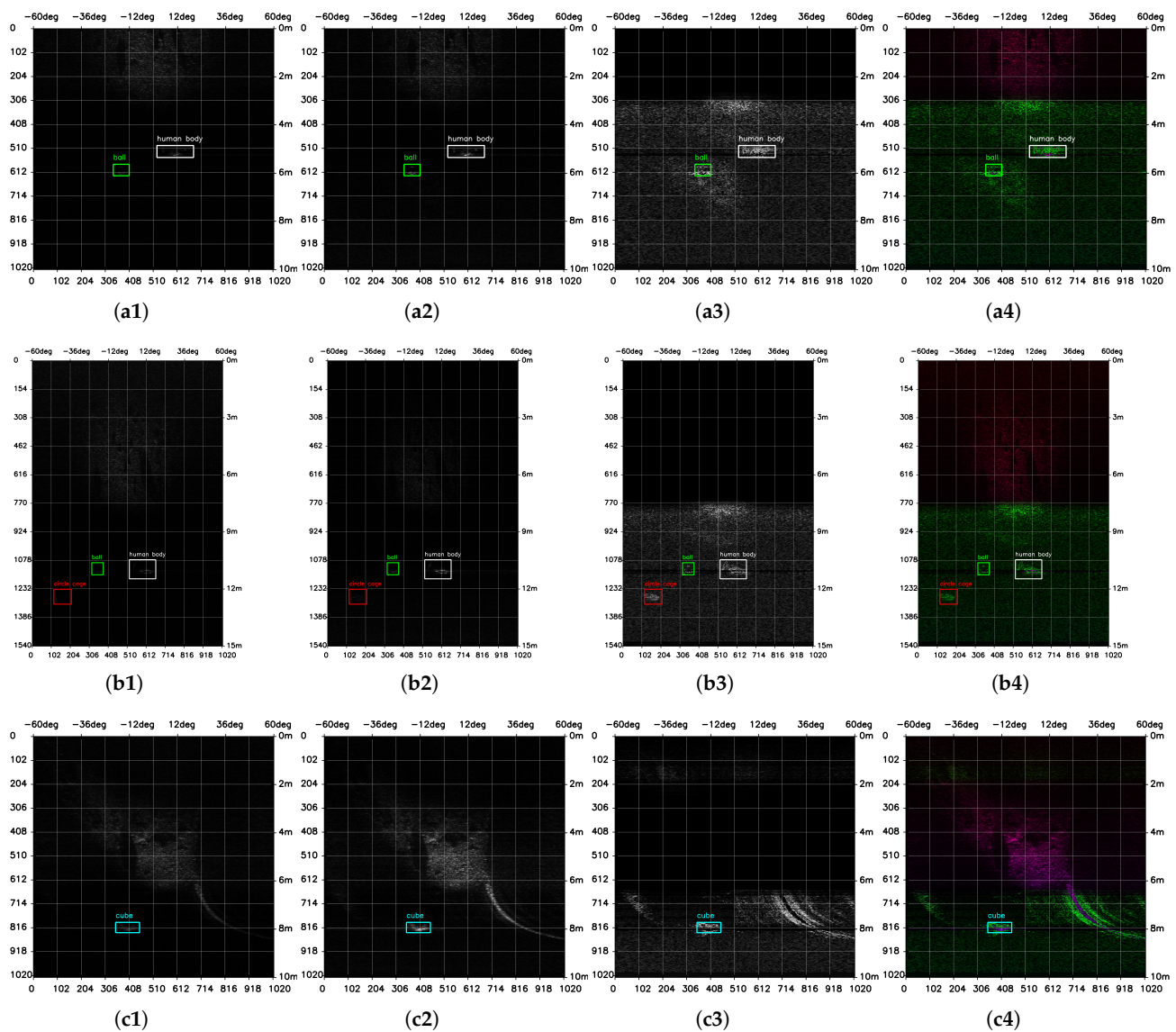


Figure 6. Visual comparison of enhancement components on representative FLS images (a,b,c). Columns from left to right: (1) Raw FLS images, (2) attenuation-compensated results, (3) probability weighting outputs, and (4) fused enhanced outcomes. Axes are labeled with azimuth (top), range (right), pixel width (bottom) and height (left) along the image frame, with a grid overlay. The fusion improves target saliency and pattern consistency.

The increased target saliency also enables more accurate dataset annotation and calibration. To demonstrate this, we analyze an instance containing a (spherical) ball and a square cage, summarized in Table 4. According to the original dataset documentation [21], targets were deliberately deployed prior to data collection, ensuring the ground-truth count and class of objects are correct, while bounding boxes may still suffer from human annotation uncertainty in low-visibility raw imagery.

The physical dimensions of the targets are listed in Table 5. For the ball with radius 0.25 m, the maximum reflective surface is bounded by a circle of diameter 0.5 m, and the maximum depth profile (front-to-back extent along range) is bounded by radius 0.25 m. Therefore, we expect its actual reflective width $W_r \leq 0.5$ and depth profile $H_r \leq 0.25$.

Based on the original annotations in Table 4, namely the range $R = 13.5033$ m, azimuth $\theta = 120^\circ$, and image dimensions $W = 1024$ and $H = 1387$, Equations (17) and (16) allows us to convert pixel-wise bounding box extents into metric estimates of the reflective width W_r and depth profile, i.e., the height of the reflective surface H_r . The initial bounding box of the ball yields a reflective width of 0.6238 m and a depth profile of 0.4381 m, which significantly exceeds the physical dimensions of the ball. Conversely, the original annotation for the square cage yields a width of 0.0167 m and a depth of 0.0084 m, which is implausibly small. These discrepancies occurred because the targets were indistinguishable in the raw data (Figure 7, upper left, white bounding boxes).

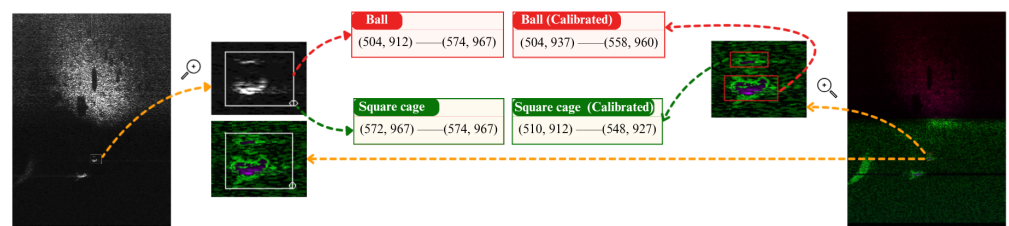


Figure 7. Comparison of target annotations in a zoomed view before (upper left, white bounding boxes) and after (right, red bounding boxes) calibration based on statistical enhancement. The location of the square cage in raw annotation is circled here due to its minimal pixel extent. A zoomed view of the original annotations overlaid on the enhanced image is also shown (lower left; white bounding boxes).

After applying our statistical enhancement, both targets become clearly delineated (Figure 7, right). We calibrated the bounding boxes (red boxes), yielding a corrected width of 0.4715 m and a depth of 0.2239 m for the ball, as well as a width of 0.3537 m and a depth of 0.1460 m for the cage. Both sets of calibrated dimensions closely align with their true physical sizes. Furthermore, a suspicious high-intensity pattern at pixel-wise coordinates (400, 1020) to (500, 1080) was revealed by the enhancement. Calculations indicate a width of 0.6710 m and a depth of 0.5841 m, which does not match any deployed target. This confirms it is a noise artifact, verifying that the original annotators were correct to exclude it.

Beyond numerical verification, the first and second columns of Figure 6 visually demonstrate that original FLS images, even after logarithmic transformation or attenuation compensation, exhibit insufficient contrast for effective target identification. To enhance readability, the left panel of Figure 7 has been processed using a contrast enhancement pipeline—contrast stretching with percentile clipping, contrast-limited adaptive histogram equalization, and a mild unsharp mask—making multiple shadows more apparent than in the unprocessed raw view. Nevertheless, this example underscores that our approach delineates the bounding boxes of targets more effectively. Although targets were deliberately deployed to ensure ground-truth accuracy, Xie et al. [21] struggled to track, identify, and distinguish the targets in some images, including this example, during the annotation process. This led to confusion, with the targets being annotated as a single object. As shown in the upper left of Figure 7 (white bounding boxes), only one prominent object is discernible to the naked eye in the raw image, which we believe contributed to the annotation error, and the contrast-enhanced illustration still does not provide sufficiently clear visual cues to enable accurate delineation of the bounding boxes. Consequently, the contrast enhancement and pseudo-coloring provided by our FLS image enhancement method effectively aid the annotation process.

Overall, this example demonstrates that statistical enhancement improves target interpretability and supports more accurate labeling through physically grounded consistency checks.

4.2. Performance Evaluation and Ablation Studies

To quantify the impact of injecting quasi-static in situ statistical information into the pipeline of UATD tasks, we evaluate the proposed enhanced detector (denoted as YOLOv12n-QSIS) against representative one-stage YOLOv12 variants, transformer-based DETR variants, and a two-stage FRCN baseline on the UATD dataset. We further compare our results with prior UATD-related methods reported on public datasets, including FLSD-Net [12], ATTMPConvNet [18], and WBF-ASFFNet [13]. Notably, ASFFNet [13] does not provide sufficient reproducible details, and its description of the UATD dataset substantially contradicts that of Xie et al. [21]. Although it reports markedly higher performance than other studies, the source of this discrepancy is unclear; we therefore include its numbers in our table only for reference. Table 7 reports Precision, Recall, F1 score, and computational cost in terms of FLOPs, together with model size.

Table 7. Overall detection performance and computational cost on the UATD dataset.

Model	Precision	Recall	F1	FLOPs (G)	Parameters (MB)
YOLOv12n-QSIS	0.896	0.836	0.865	6.3	2.6
YOLOv12n	0.805	0.795	0.800	4.2	2.6
YOLOv12m	0.855	0.800	0.827	44.5	20.2
YOLOv12x	0.894	0.799	0.844	131.3	59.1
DETR-R50	0.834	0.794	0.814	34.1	19.9
DETR-R101	0.85	0.795	0.822	47.6	29.9
FRCN	0.854	0.795	0.823	33.2	26.3
FLSD-Net [12]	0.861	0.811	0.835	5.4	1.8
ATTMPConvNet [18]	0.858	0.804	0.830	74.2	37.7
WBF-ASFFNet [13]	0.982	0.943	0.962	—	—

As shown in Table 7, YOLOv12n-QSIS achieves the best overall detection performance among all compared detectors while retaining a lightweight backbone and low computational cost. Compared with the vanilla YOLOv12n baseline, our proposed hybrid framework improves F1 score by 8.1%, from 0.800 to 0.865, with nearly the same parameter size and a modest increase in FLOPs (from 4.2 G to 6.3 G), indicating that the gain primarily comes from the proposed quasi-static in situ processing rather than increased model capacity. Moreover, our proposed method also outperforms substantially larger backbones, such as YOLOv12x (F1 score: 0.865 vs. 0.844, a 2.5% improvement), despite requiring 95% fewer FLOPs (6.3 G vs. 131.3 G) and 96% fewer parameters (2.6 MB vs. 59.1 MB), suggesting a more favorable accuracy–efficiency balance than scaling the network size alone. To visualize this trade-off, Figure 8 plots F1 score versus FLOPs for all evaluated models. The results show that, for conventional deep detectors, improved performance is largely coupled with increased computational cost, approximately following a monotonic trend. In contrast, the proposed hybrid approach delivers the strongest accuracy in the low-compute regime, avoiding the steep computational growth required by larger backbones.

To further isolate the contributions of individual components in the hybrid framework, we conducted an ablation study by systematically removing key modules from the workflow. The variants are defined as follows: QSIS-nc (no calibration), QSIS-nac (no attenuation compensation), and QSIS-npw (no probability weighting). The results, detailed in Table 8, confirm that the full quasi-static in situ learning pipeline yields the highest overall performance, highlighting the synergistic effect of these modules.

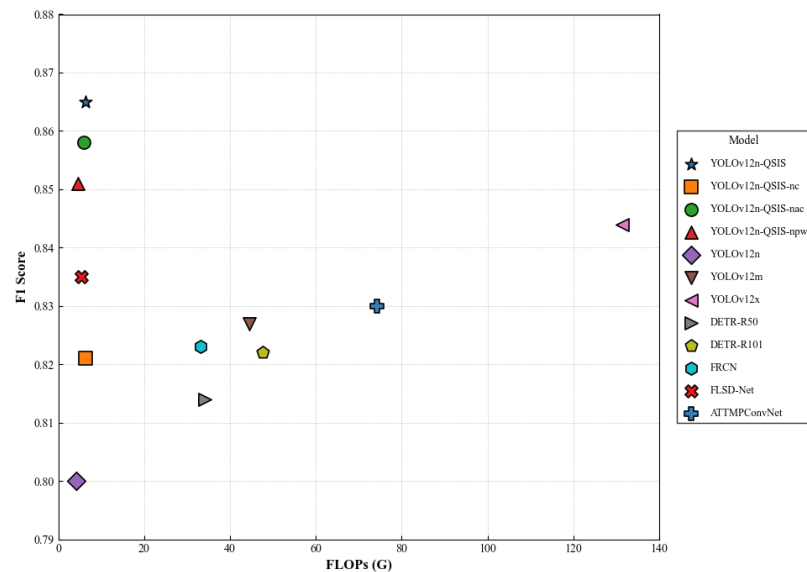


Figure 8. F1 score versus FLOPs for the evaluated models, illustrating the efficiency–performance trade-off and highlighting the superiority of the proposed quasi-static in situ learning framework.

Table 8. Ablation study results of components in the proposed quasi-static in situ (-QSIS) learning framework on the UATD dataset, evaluating the impact of removing calibration (-nc), attenuation compensation (-nac), or probability weighting (-npw).

Model	Precision	Recall	F1	FLOPs (G)
YOLOv12n-QSIS	0.896	0.836	0.865	6.3
YOLOv12n-QSIS-nc	0.849	0.795	0.821	6.3
YOLOv12n-QSIS-nac	0.882	0.836	0.858	5.9
YOLOv12n-QSIS-npw	0.867	0.835	0.851	4.6

Notably, removing the calibration module causes the largest degradation, particularly in recall. This observation is consistent with a key challenge in UATD: imperfect or inconsistent annotations can distort the target appearance distribution, making some instances difficult to retrieve. As discussed in Section 1, this challenge arises from inherent ambiguities in the FLS images, which substantially complicate the manual annotation process. While increasing model scale can reduce certain false positives, thus improving precision, it does not fundamentally resolve missed detections induced by unreliable pattern cues. The proposed calibration step explicitly mitigates this issue by correcting annotation-related inconsistencies, thereby improving recall. Although calibration introduces additional manual effort during the tranche context construction stage, this cost is incurred only offline and does not affect inference-time computation.

Attenuation compensation and probability weighting also yield consistent gains. Although probability weighting contributes more to the overall F1 score, attenuation compensation adds negligible computational cost compared to conventional model scaling, justifying its retention in the final architecture.

5. Conclusions

In this study, we proposed a novel quasi-static in situ learning paradigm for FLS-imagery-based UATD to address severe acoustic distortion, environmental complexity, and the scarcity of fine-annotated data. The proposed framework integrates lognormal modeling of echo intensities, acoustic attenuation compensation, and lightweight score-level fusion between statistical descriptors and deep learning predictions. By incorporating dy-

namically estimated environmental cues, the framework improves detection performance, interpretability, and adaptability, while also mitigating noise and facilitating annotation calibration. Experiments on the UATD dataset demonstrate the effectiveness of the proposed paradigm under simulated quasi-static tranches corresponding to different sonar frequencies. The method improves detection performance through dynamic model adaptation with limited computational overhead, indicating its potential for deployment on portable AUV platforms. These results suggest that combining physics-informed statistical modeling with lightweight deep learning can alleviate the limitations of both purely data-driven detectors and handcrafted statistical methods.

Despite these improvements, several limitations remain. First, the proposed workflow assumes that the environment remains approximately quasi-static within each tranche, which may not hold in highly dynamic scenarios. Second, the current stability indicators rely partly on target-related information, which may limit their applicability to open-ended exploration tasks involving unknown or changing targets. Third, the annotation calibration process still requires manual intervention, reducing scalability in more complex deployment conditions. Previous work has employed unsupervised learning for segmentation and classification [24], suggesting that automating calibration through such methods could enhance the workflow's adaptability.

Future research will focus on targeted extensions of the proposed framework to address these limitations. First, to improve robustness in highly dynamic environments where the quasi-static assumption may be violated, the framework could incorporate an expanded statistical toolbox with alternative modeling and fusion schemes, coupled with a reinforcement learning policy [45] for online selection and weighting based on environmental context, alongside a lighter, rapidly tunable detector. Second, automating the calibration module using the proposed statistical descriptors, potentially integrating zero-shot multi-modal architectures such as CLIP [46] and DenseFusion [47], would improve scalability and deployment efficiency. Further refinements in indicator design could also enhance compatibility with dynamic environments by partitioning variable environments into statistically quasi-static segments, thereby improving the generalization capability of the proposed framework.

Author Contributions: Conceptualization, Y.C., T.X., Z.D. and Q.G.; methodology, Y.C., Y.F., Q.G. and Z.D.; software, Y.C., Z.D. and Z.X.; validation, Y.C., Y.F., T.X. and Z.D.; formal analysis, Y.C., Y.F., K.C. and J.H.; investigation, Y.C., Y.F., J.H. and Z.X.; resources, Z.D., Y.F., K.C. and T.X.; data curation, Z.D., J.H. and Z.X.; writing—original draft preparation, Y.C., Y.F., Z.X., T.X. and Z.D.; writing—review and editing, Y.C., Y.F., J.H., Z.X., K.C. and Q.G.; visualization, Y.C., J.H. and Z.X.; supervision, T.X., K.C. and Q.G.; project administration, Y.F., T.X., K.C. and Q.G.; funding acquisition, Y.C., K.C., T.X. and Q.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The UATD dataset and streaming annotation tool used in this study were obtained from https://figshare.com/articles/dataset/UATD_Dataset/21331143/3 (accessed on 10 November 2025). The source code for the proposed scene prior-based data enhancement and fusion framework is publicly available at https://github.com/EtchChan/UATD_AttributeFusion (accessed on 5 April 2026). The official implementation of YOLOv12 is available at <https://github.com/sunsmarterjie/yolov12> (accessed on 21 December 2025), the official FlashAttention implementation is available at <https://github.com/Dao-AILab/flash-attention> (accessed on 28 December 2025), the official DETR implementation is available at <https://github.com/facebookresearch/detr> (accessed on 24 February 2026), and the official Faster R-CNN implementation is available at https://github.com/ShaoqingRen/faster_rcnn (accessed on 16 January 2026).

Acknowledgments: We would like to thank Yan Yan from the School of Journalism and Communication at Sun Yat-sen University for her substantial help in improving the visual effects of our paper's illustrations, as well as for her suggestions regarding grammar and phrases. The authors have reviewed and edited the content and take full responsibility for the content of this publication.

Conflicts of Interest: Yu Feng (intern), Ziqin Xie (intern), Yixuan Chen (intern), Zhenqing Ding, Tinggang Xiong and Qi Gao were employed by Wuhan Lingjiu Microelectronics Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ARS	Acoustic Rolling Shutter
AUV	Autonomous Underwater Vehicle
CLT	Central Limit Theorem
CTD	Conductivity, Temperature, Depth
DETR	Acoustic Rolling Shutter
FLOP	Floating-point Operation
FLS	Forward-Looking Sonar
FRCN	Fast Region-based Convolutional Network
IoU	Intersection over Union
PDF	Probability Density Function
ROV	Remotely Operated Vehicle
SOTA	State-Of-The-Art
UATD	Underwater Acoustic Target Detection
YOLO	You-Only-Look-Once detector

References

1. Ferreira, F.; Djapic, V.; Micheli, M.; Caccia, M. Forward looking sonar mosaicing for Mine Countermeasures. *Annu. Rev. Control* **2015**, *40*, 212–226. [\[CrossRef\]](#)
2. Zheng, H.; Sun, Y.; Zhang, G.; Zhang, L.; Zhang, W. Research on Real Time Obstacle Avoidance Method for AUV Based on Combination of ViT-DPFN and MPC. *IEEE Trans. Instrum. Meas.* **2024**, *73*, 1–15. [\[CrossRef\]](#)
3. Vidal, E.; Palomeras, N.; Istenić, K.; Gracias, N.; Carreras, M. Multisensor online 3D view planning for autonomous underwater exploration. *J. Field Robot.* **2020**, *37*, 1123–1147. [\[CrossRef\]](#)
4. Hurtós, N.; Palomeras, N.; Nagappa, S.; Salvi, J. Automatic detection of underwater chain links using a forward-looking sonar. In Proceedings of the 2013 MTS/IEEE OCEANS—Bergen, Bergen, Norway, 10–14 June 2013; pp. 1–7. [\[CrossRef\]](#)
5. DeMarco, K.J.; West, M.E.; Howard, A.M. Sonar-Based Detection and Tracking of a Diver for Underwater Human-Robot Interaction Scenarios. In Proceedings of the 2013 IEEE International Conference on Systems, Man, and Cybernetics; IEEE: Piscataway, NJ, USA, 2013; pp. 2378–2383.
6. Zhang, T.; Liu, S.; He, X.; Huang, H.; Hao, K. Underwater Target Tracking Using Forward-Looking Sonar for Autonomous Underwater Vehicles. *Sensors* **2019**, *20*, 102. [\[CrossRef\]](#)
7. Hafiza, W.; Shukor, M.M.; Jasman, F.; Mutalip, Z.A.; Abdullah, M.S.; Idrus, S.M. Advancement of Underwater Surveying and Scanning Techniques: A Review. *J. Adv. Res. Appl. Sci. Eng. Technol.* **2024**, *41*, 256–281. [\[CrossRef\]](#)
8. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [\[CrossRef\]](#) [\[PubMed\]](#)
9. Pouyanfar, S.; Sadiq, S.; Yan, Y.; Tian, H.; Tao, Y.; Reyes, M.P.; Shyu, M.L.; Chen, S.C.; Iyengar, S.S. A Survey on Deep Learning: Algorithms, Techniques, and Applications. *Acm Comput. Surv.* **2019**, *51*, 1–36. [\[CrossRef\]](#)
10. Zhang, H.; Tian, M.; Shao, G.; Cheng, J.; Liu, J. Target Detection of Forward-Looking Sonar Image Based on Improved YOLOv5. *IEEE Access* **2022**, *10*, 18023–18034. [\[CrossRef\]](#)
11. Xie, B.; He, S.; Cao, X. Target Detection for Forward Looking Sonar Image based on Deep Learning. In Proceedings of the 2022 41st Chinese Control Conference (CCC), Hefei, China, 25–27 July 2022; pp. 7191–7196. [\[CrossRef\]](#)
12. Yang, H.; Zhou, T.; Jiang, H.; Yu, X.; Xu, S. A Lightweight Underwater Target Detection Network for Forward-Looking Sonar Images. *IEEE Trans. Instrum. Meas.* **2024**, *73*, 1–13. [\[CrossRef\]](#)
13. Lilan, Z.; Bo, L.; Xu, C.; Shufa, L.; Cong, L. Sonar Image Target Detection for Underwater Communication System Based on Deep Neural Network. *Comput. Model. Eng. Sci.* **2023**, *137*, 2641–2659. [\[CrossRef\]](#)

14. Long, H.; Shen, L.; Wang, Z.; Chen, J. Underwater Forward-Looking Sonar Images Target Detection via Speckle Reduction and Scene Prior. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–13. [\[CrossRef\]](#)
15. Sung, M.; Kim, J.; Lee, M.; Kim, B.; Kim, T.; Kim, J.; Yu, S.C. Realistic Sonar Image Simulation Using Deep Learning for Underwater Object Detection. *Int. J. Control. Autom. Syst.* **2020**, *18*, 523–534. [\[CrossRef\]](#)
16. Islam, M. Deep Learning-Based Sonar Image Object Detection System. *Int. J. Inform. Inf. Syst. Comput. Eng. (Injiiscom)* **2024**, *6*, 186–199. [\[CrossRef\]](#)
17. Wang, Y.; Liu, Z.; Li, G.; Lu, X.; Liu, X.; Zhang, H. Hybrid Modeling Based Semantic Segmentation of Forward-Looking Sonar Images. *IEEE J. Ocean. Eng.* **2025**, *50*, 380–393. [\[CrossRef\]](#)
18. Sadhu, L.G.S. Advanced deep learning framework for underwater object detection with multibeam forward-looking sonar. *Struct. Health Monit.* **2025**, *24*, 1991–2007. [\[CrossRef\]](#)
19. Zhang, Y.; Wang, C.; Zhang, Q.; Li, Q.; Da, L. A Review of Underwater Acoustic Target Detection and Recognition Technology Based on Information Fusion. *J. Signal Process.* **2023**, *39*, 1711–1727.
20. Kuznetsova, A.; Rom, H.; Alldrin, N.; Uijlings, J.; Krasin, I.; Pont-Tuset, J.; Kamali, S.; Popov, S.; Mallocci, M.; Kolesnikov, A.; et al. The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale. *Int. J. Comput. Vis.* **2020**, *128*, 1956–1981. [\[CrossRef\]](#)
21. Xie, K.; Yang, J.; Qiu, K. A Dataset with Multibeam Forward-Looking Sonar for Underwater Object Detection. *Sci. Data* **2022**, *9*, 739. [\[CrossRef\]](#)
22. Dos Santos, M.M.; Ballester, P.; Zaffari, G.B.; Drews, P.; Botelho, S. A Topological Descriptor of Acoustic Images for Navigation and Mapping. In Proceedings of the 2015 12th Latin American Robotics Symposium and 2015 3rd Brazilian Symposium on Robotics (LARS-SBR), Uberlândia, Brazil, 29–31 October 2015; pp. 289–294. [\[CrossRef\]](#)
23. Dos Santos, M.; Ribeiro, P.O.; Núñez, P.; Drews, P., Jr.; Botelho, S. Object Classification in Semi Structured Environment Using Forward-Looking Sonar. *Sensors* **2017**, *17*, 2235. [\[CrossRef\]](#)
24. Abu, A.; Diamant, R. A Statistically-Based Method for the Detection of Underwater Objects in Sonar Imagery. *IEEE Sens. J.* **2019**, *19*, 6858–6871. [\[CrossRef\]](#)
25. Su, J.; Qian, J.; Tu, X.; Qu, F.; Wei, Y. Analysis and Compensation of Acoustic Rolling Shutter Effect of Acoustic-Lens-Based Forward-Looking Sonar. *IEEE J. Ocean. Eng.* **2024**, *49*, 474–486. [\[CrossRef\]](#)
26. Huertas-Tato, J.; Martín, A.; Fierrez, J.; Camacho, D. Fusing CNNs and statistical indicators to improve image classification. *Inf. Fusion* **2022**, *79*, 174–187. [\[CrossRef\]](#)
27. Abraham, D.; Gelb, J.; Oldag, A. Rayleigh-non-Rayleigh mixtures for active sonar clutter. *J. Acoust. Soc. Am.* **2010**, *128*, 2380. [\[CrossRef\]](#)
28. Foote, K.G.; MacLennan, D.N. Comparison of copper and tungsten carbide calibration spheres. *J. Acoust. Soc. Am.* **1984**, *75*, 612–616. [\[CrossRef\]](#)
29. Thorp, W.H. Analytic Description of the Low-Frequency Attenuation Coefficient. *J. Acoust. Soc. Am.* **1967**, *42*, 270. [\[CrossRef\]](#)
30. Fisher, F.H.; Simmons, V.P. Sound Absorption in Sea Water. *J. Acoust. Soc. Am.* **1977**, *62*, 558–564. [\[CrossRef\]](#)
31. Sehgal, A.; Tumar, I.; Schonwalder, J. Variability of available capacity due to the effects of depth and temperature in the underwater acoustic communication channel. In Proceedings of the OCEANS 2009-EUROPE, Bremen, Germany, 11–14 May 2009; pp. 1–6. [\[CrossRef\]](#)
32. Etter, P. *Underwater Acoustic Modeling and Simulation*, 4th ed.; CRC Press-Taylor & Francis Group: Abingdon, UK, 2013; pp. 1–492. [\[CrossRef\]](#)
33. Anandalatchoumy, S.; Sivaradje, G. Comprehensive study of acoustic channel models for underwater wireless communication networks. *Int. J. Cybern. Inform.* **2015**, *4*, 227–240.
34. Meyer, V.; Audoly, C. A Comparison Between Experiments And Simulation For Shallow Water Short Range Acoustic Propagation. In Proceedings of the ICSV 24 24th International Congress on Sound and Vibration, London, UK, 23–27 July 2017.
35. Onur, T.O. Investigation of Parameters Affecting Underwater Communication Channel. *J. Eng. Sci.* **2020**, *7*, F39–F44. [\[CrossRef\]](#)
36. Zanj, E.; Gambi, E.; Zanj, B.; Disha, D.; Kola, N. Underwater Wireless Sensor Networks: Estimation of Acoustic Channel in Shallow Water. *Appl. Sci.* **2020**, *10*, 6393. [\[CrossRef\]](#)
37. Ribeiro, B.A.; Xavier, F.C.; Barroso, V.R.; Silva, V.F.d.; Netto, T.A.; Ferraz, C. Study and Feasibility of Underwater Acoustic Data Transmission. *J. Mar. Sci. Eng.* **2026**, *14*, 648. [\[CrossRef\]](#)
38. Fienen, M.N.; White, J.T.; Hayek, M. Parameter ESTimation With the Gauss–Levenberg–Marquardt Algorithm: An Intuitive Guide. *Ground Water* **2025**, *63*, 93–104. [\[CrossRef\]](#) [\[PubMed\]](#)
39. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2016; pp. 779–788.
40. Fletcher, R. *Practical Methods of Optimization*, 2nd ed.; A Wiley-Interscience Publication; John Wiley & Sons: Chichester, UK, 1987–2006; 13th Reprint.

41. Tian, Y.; Ye, Q.; Doermann, D. YOLOv12: Attention-Centric Real-Time Object Detectors. In *Proceedings of the Advances in Neural Information Processing Systems*; Belgrave, D., Zhang, C., Lin, H., Pascanu, R., Koniusz, P., Ghassemi, M., Chen, N., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2025; Volume 38, pp. 78433–78457.
42. Dao, T.; Fu, D.Y.; Ermon, S.; Rudra, A.; Ré, C. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, USA, 28 November–9 December 2022.
43. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. In *Proceedings of the Computer Vision—ECCV 2020*; Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M., Eds.; Springer: Cham, Switzerland, 2020; pp. 213–229.
44. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Proceedings of the Advances in Neural Information Processing Systems*; Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2015; Volume 28.
45. Mazyavkina, N.; Sviridov, S.; Ivanov, S.; Burnaev, E. Reinforcement learning for combinatorial optimization: A survey. *Comput. Oper. Res.* **2021**, *134*, 105400. [[CrossRef](#)]
46. Chen, H.Y.; Lai, Z.; Zhang, H.; Wang, X.; Eichner, M.; You, K.; Cao, M.; Zhang, B.; Yang, Y.; Gan, Z. Contrastive Localized Language-Image Pre-Training. In *Proceedings of the 42nd International Conference on Machine Learning*; Singh, A., Fazel, M., Hsu, D., Lacoste-Julien, S., Berkenkamp, F., Maharaj, T., Wagstaff, K., Zhu, J., Eds.; PMLR: San Diego, CA, USA, 2025; Volume 267, pp. 8386–8402.
47. Li, X.; Zhang, F.; Diao, H.; Wang, Y.; Wang, X.; Duan, L.Y. DenseFusion-1M: Merging Vision Experts for Comprehensive Multimodal Perception. In *Proceedings of the Advances in Neural Information Processing Systems*; Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2024; Volume 37, pp. 18535–18556. [[CrossRef](#)]

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.