J Intel Connect Veh

Research Article

AUV target perception in forward-looking sonar images: A low-pixel dataset and multi-frame segmentation network

Jingfeng Yu[1,2,✉], Chen Liang[3], Kailai Sun[4,5,✉], Jiugen Lin[1], Zhongju Sun[1], Ruiming Wang[3], Qianchuan Zhao[2]

Cite this article: https://doi.org/10.26599/JICV.2026.9210108

ABSTRACT: This study investigates low-pixel segmentation of a moving autonomous underwater vehicle (AUV) observed by a fixed forward-looking sonar (FLS) and formulates the task as a pixel-level segmentation problem under low-pixel and noise-interference conditions. We release the Underwater Sonar Moving Target (USMT) dataset, which contains 21 temporally ordered sequences and 7,136 annotated FLS frames collected using a fixed FLS while observing a single AUV in a controlled lake environment. To exploit motion cues across consecutive frames, we propose LSNet, a multi-frame segmentation network adapted to the characteristics of sonar imagery. LSNet incorporates three improvements: (1) a spatio-temporal scale fusion block that combines multi-scale spatial features with variance-based temporal motion cues; (2) a decoupled 3D spatio-temporal attention module that separates local perception from temporal verification; and (3) a Muon-Adam hybrid optimizer with a joint loss function. Technical validation shows that the rectangular B-scan representation is more suitable for segmentation than the sector-shaped representation. On the USMT dataset, LSNet achieves an IoU of 68.21%, outperforming the retrained LVNet reference by 19.57 percentage points and the strongest additional comparator, LMAFormer, by 5.45 percentage points, while maintaining an inference speed of 91.23 FPS. These results demonstrate the effectiveness of multi-frame modeling for low-pixel AUV target segmentation in external FLS-based monitoring.

KEYWORDS: autonomous underwater vehicle; forward-looking sonar; low-pixel target; segmentation; LSNet

1 Introduction

Low-pixel, long-range, small-target visual perception is of great engineering value. In addition to conventional optical images, physical-field images from specific domains play an irreplaceable role in the detection of small targets. Examples include infrared cameras used for imaging and tracking moving targets to provide early warning of intruders such as drones, vehicles, and living organisms, as well as surface or underwater unmanned platforms equipped with forward-looking sonars (FLSs) for surveying underwater moving targets to identify categories including autonomous underwater vehicles (AUVs), fish, and divers. With the rapid advancement of artificial intelligence, researchers increasingly employ deep learning methods to extract information from sparse sonar images, including target existence determination, localization, and classification.

Existing studies have focused on FLS imagery and low-pixel target detection. In 2022, Peng Cheng Laboratory released the UATD sonar image dataset for object detection (Xie et al., 2022), as shown in Fig. 1 (a). They collected close-range data of ten categories of preset underwater targets, forming a dataset comprising approximately 9,000 FLS images. Nankai University released the NKSID sonar image dataset (Jiao et al., 2024), which comprises 2,617 images of deep-sea objects across eight categories, as shown in Fig. 1 (b). Dahn et al. (2024) generated a multimodal image dataset of three different types of unexploded ordnance, consisting of approximately 74,000 frames, using an ARIS 3000 imaging sonar. Gomes et al. (2025) provided a comprehensive review of internationally open-source sonar datasets.

In addition, given the diversity of research problems and constraints on experimental costs, substantial efforts are concurrently devoted to generating simulated imagery. HoloOcean, widely adopted by the research community (Potokar et al., 2022a), is an open-source UE4-based simulator that generates sonar images in real-time via octree representations (Potokar et al., 2022b), as shown in Fig. 1 (c). UUV Simulator (Manhães et al., 2016) and UWSim (Prats et al., 2012) also support the simulation and generation of sonar images. UDRSim (Yu et al., 2026), recently introduced, as shown in Fig. 1 (d), is primarily designed to address the generation of self-annotated FLS image data in underwater dynamic environments. Furthermore, attempts have been made to augment existing sonar image data using diffusion models (Yang et al., 2023), and tests have demonstrated that such approaches can improve target detection performance.

In 2025, Shen et al. from Sun Yat-sen University proposed LVNet based on traditional Vision Transformers for multi-frame spatio-temporal context modeling, effectively capturing target motion features. Their method outperformed existing state-of-the-art approaches on the

Just Accepted

[1] Ganjiang Innovation Academy, Chinese Academy of Sciences, Ganzhou 341119, China. [2] Department of Automation, Tsinghua University, Beijing 100084, China. [3] Department of Automation, University of Science and Technology of China, Hefei 230026, China. [4] Singapore-MIT Alliance for Research and Technology Centre (SMART), Singapore 138602, Singapore. [5] Urban Mobility Lab, Massachusetts Institute of Technology, Cambridge MA 02139, USA.

✉ Corresponding author. E-mail: J. Yu, jfy@gia.cas.cn; K. Sun, skl24@mit.edu

Received: August 10, 2026; Revised: September 1, 2026; Accepted: October 15, 2026

© The Author(s) 2026. This is an open access article under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0, http://creativecommons.org/licenses/by/4.0/).

public IRDST datasets (Sun et al., 2023). To address the inherent accuracy loss in the binarization process of infrared small targets, researchers from Northwestern Polytechnical University and other institutions proposed BiisNet (Xin et al., 2026), which integrates full-precision features into the network information flow while retaining the core binarized convolution operations. Experimental results show that this network not only significantly surpasses other binarized architectures but also demonstrates strong competitiveness against state-of-the-art full-precision models. The above research has advanced the detection and recognition capabilities of small targets in the image domain, providing valuable references and insights for subsequent studies.

However, research on AUV target perception in forward-looking sonar images remains limited, particularly under low-pixel and complex underwater conditions. Existing studies have mainly focused on targets with relatively clear spatial features, while the challenges associated with sparse target representation, limited annotated data, and insufficient temporal information utilization have not been adequately addressed. (1) Due to the high risks associated with at-sea testing, substantial experimental costs, and the scarcity of high-quality annotated data, the issue of limited sample sizes in underwater sonar image datasets remains prominent, significantly impeding research progress in underwater applications. (2) At greater observation ranges, target echoes generally tend to occupy fewer effective pixels and contain less distinct spatial information. In some cases, the target may appear as a compact bright response, making perception difficult. (3) FLS imagery is affected by sonar-specific acoustic interference, including reverberation, multipath returns, sidelobes, and speckle-like intensity fluctuations. These responses can produce locally bright or structured patterns that resemble low-pixel target echoes, making target-background discrimination difficult from a single frame.

Although recent infrared small-target segmentation methods provide valuable references, directly applying these methods to sonar imagery remains challenging due to differences in imaging mechanisms, noise characteristics, and target appearance. Therefore, LVNet was selected as the development baseline because it is a representative multi-frame small-target segmentation framework with a similar problem formulation. To broaden the comparison, a standard single-frame U-Net and four representative multi-frame methods, namely DeepPro-Plus, RFR, DTUM, and LMAFormer, were also included.

The main contributions of this study are summarized as follows:

(1) We construct and publicly release the USMT dataset, which contains 21 temporally ordered sequences and 7,136 annotated FLS frames collected using a fixed FLS while observing a moving AUV. The released dataset contains rectangular B-scan images with pixel-level annotations.

(2) We propose LSNet, a multi-frame segmentation network that combines multi-scale spatial representation, variance-based motion perception, and decoupled spatio-temporal attention for low-pixel AUV targets in FLS imagery.

(3) We validate the dataset and LSNet through image-representation comparison, comparative evaluation against LVNet, U-Net, and four representative multi-frame methods, and ablation studies. LSNet achieves an IoU of 68.21%, exceeding the strongest additional comparator, LMAFormer, by 5.45 percentage points while maintaining 91.23 FPS.

(4) The dataset and real-time segmentation framework provide a basis for external sonar-based AUV target monitoring and future investigations involving more diverse targets and marine environments.

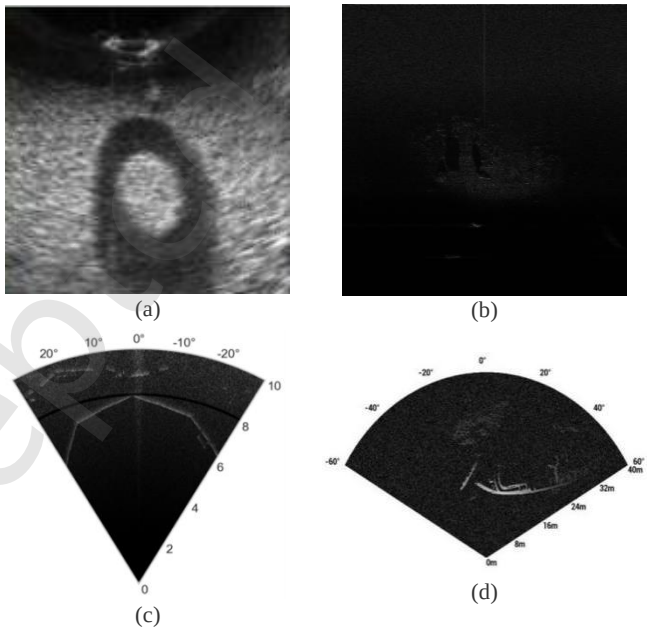


Fig. 1. Partial sonar dataset images. (a) NKSID sonar image; (b) UATD dataset; (c) HoloOcean; (d) UDRSim. (Jiao et al., 2024; Potokar et al., 2022a; Xie et al., 2022; Yu et al., 2026).

2 USMT dataset construction

The USMT dataset was constructed through field data collection, frame-level annotation, image-representation conversion, sequence organization, and technical validation. This section describes the acquisition setup, annotation workflow, data representations, and dataset characteristics.

2.1. Data collection

The experiment was conducted on an artificial lake in Qing County, Cangzhou, Hebei Province, China, with a length of approximately 380 m and a depth of 20–30 m (Fig. 2). The field trial was conducted in December 2025 under "Level 1" water conditions with low ambient noise; however, multipath propagation and reverberation remained pronounced due to environmental constraints. Data were collected using an FLS700D, with detailed sonar parameters listed in Table 1.

The sonar transducer was installed at a depth of 2.95 m below the water surface using a custom mounting fixture, oriented toward the open water directly ahead. The fixed installation maintained a stable transducer position and observation direction throughout the data acquisition process. The fixture was welded to the floating platform on

Just Accepted

the right side of Fig. 3, allowing the experimenters to control the sonar's host computer from the platform and adjust the underwater sonar parameters as needed. The sonar images could therefore be observed in real time, and the imaging range and other acquisition parameters could be adjusted according to the target position and echo conditions. In the present experimental configuration, the FLS remained fixed on the floating platform and the AUV was the moving object being observed. Therefore, the task investigated in this study represents external sonar-based monitoring of an AUV target rather than onboard AUV perception.

Table 1. Acoustic specifications of FLS700D.

Operating frequency	Range	Horizontal beamwidth	Vertical beamwidth	CHIRP	Speed of sound
400 kHz	150 m	180°	20°	support	1300-1600 m/s



Fig. 2. Satellite image and on-site photograph of the experimental site.



Fig. 3. FLS700D sonar and its mounting setup in the water experiment.

The target under test during the lake trial was a 533-mm-calibre AUV with a length of 6 m, as shown in Fig. 4. During the experiment, the AUV sailed at a speed of 0-2 kn and a navigation depth of 0-4 m.

In this study, the term "low-pixel" or "small target" refers to the image-space representation of the AUV rather than its physical dimensions. Owing to the limited spatial resolution of FLS imagery, the AUV echo occupies only a small region of the image and may appear as a compact bright response with limited shape information.

The collected sequences cover different target observation ranges relative to the fixed FLS. According to the acquisition conditions, 16 sequences comprising 6,646 frames were collected under relatively short-range conditions, while 5 sequences comprising 490 frames were collected under relatively long-range conditions, as summarized in Table 2.

Table 2. Distribution of USMT sequences under different observation-range conditions.

Observation condition	Sequences	Frames
Relatively short range	16	6,646
Relatively long range	5	490
Total	21	7,136

These categories describe the relative physical observation distance during data acquisition rather than distinct target-motion patterns. No predefined motion-direction categories were used during acquisition. In general, greater observation distances tend to reduce the image-space occupancy of the target. However, because the sonar imaging range was adjusted according to the target position and echo conditions during data collection, physical observation distance does not correspond strictly to a fixed target size in the resulting images. The maximum detection range of the target measured by the FLS700D sonar was 150 m.



Fig. 4. The target AUV under test.

2.2. Pixel-level annotation and image representation

For pixel-level annotation, a combined approach was adopted, leveraging large-model batch annotation followed by manual review and refinement. We deployed X-AnyLabeling locally and loaded the Segment Anything 2 (SAM 2) video model to initialize annotation on the imported video files. As the prompt for subsequent automatic annotation, the annotation of the first frame was particularly critical. The SAM 2 large model generated an initial mask for the box-selected target; after manual verification and confirmation, the annotation file for the first frame was finalized.

The annotation process is illustrated in Fig. 5. Once the first frame was annotated, the SAM 2 model automatically annotated the remaining frames of the imported video, and each frame was subsequently reviewed manually. This manual review was used to correct inaccurate boundaries and maintain annotation consistency throughout each sequence. Finally, empty annotation files and their corresponding images were identified and removed. The number of valid data remaining in the folder was counted, yielding 21 valid annotated sequences, comprising 7,136 FLS images. Each image has a corresponding annotation file in LabelMe JSON format.

JJSE accepted

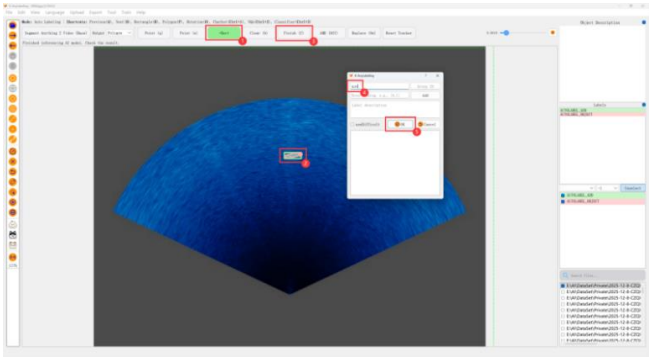
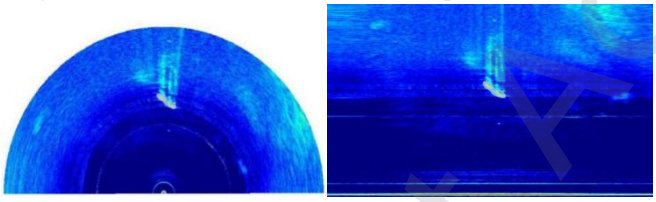


Fig. 5. Pixel-level annotation of sonar images.

After processing the raw data and performing pixel-level segmentation annotation, sector-shaped sonar images were obtained, as shown in Fig. 6 (a). Meanwhile, considering that deep learning algorithms are typically better suited to rectangular images, the generated sector polar maps were converted into the commonly used B-scan rectangular format. The B-scan image is rectangular, with the horizontal axis representing angle (sampled at equally spaced linear intervals) and the vertical axis representing range, where range increases from top to bottom. This representation unfolds each range sampling point from a sector arc into a horizontal row of pixels. After generating the rectangular sonar images, the annotation process described above was repeated, finally yielding the rectangular pixel-level annotated images shown in Fig. 6 (b). The sector images have dimensions of *632 × 792* pixels, while the rectangular images have dimensions of *1,000 × 500* pixels.



(a) (b)

Fig. 6. Sonar images after annotation. (a) Sector-shaped sonar image; (b) Rectangular sonar image.

2.3. Data organization and preparation

For the image-representation comparison in Section 2.4, the annotated sector-shaped and rectangular B-scan images described above were evaluated. Based on the comparison results, the rectangular B-scan representation was adopted for all subsequent algorithmic experiments and is the only image representation included in the released USMT dataset.

After invalid images and empty annotation files were removed, the final USMT dataset contained 21 temporally ordered video sequences comprising 7,136 annotated rectangular B-scan FLS frames. Each image was paired with a LabelMe JSON annotation containing the corresponding pixel-level target mask. The data were organized according to their original video sequences so that consecutive frames could be used to construct multi-frame inputs while preserving the temporal continuity required by LVNet and LSNet.

Accordingly, USMT provides temporally ordered rectangular B-scan images and pixel-level annotations for multi-frame segmentation research.

2.4. Technical validation

To validate the utility and technical quality of the USMT dataset, we conducted preliminary segmentation experiments using the constructed dataset. Considering that target pixel occupancy varies across the dataset and can become particularly limited under relatively long-range observation conditions, we selected LVNet, a representative multi-frame neural network for infrared small-target detection, as the baseline network. Experiments were performed and evaluated on an NVIDIA GeForce RTX 4090 computing platform. Representative evaluation metrics are presented and analyzed in Table 3.

Table 3. Segmentation performance statistics for small targets on the USMT dataset obtained by LVNet.

Image Types	IoU(%)	nIoU(%)	Pd(%)	Fa(× 10^{-6})	FPS
Sector-shaped polar image	28.68	29.66	37.93	609.39	109.38
Rectangular B-scan image	48.64	50.72	41.66	1072.08	103.69

Let P_i and G_i denote the predicted target-pixel set and ground-truth target-pixel set of the i-th frame, respectively. The four evaluation metrics are formally defined as follows:

$$\mathrm{IoU} = \frac{\sum_{i=1}^{N} |P_i \cap G_i|}{\sum_{i=1}^{N} |P_i \cup G_i|}$$

$$\mathrm{nIoU} = \frac{1}{N} \sum_{i=1}^{N} \frac{|P_i \cap G_i|}{|P_i \cup G_i|}$$

$$P_{\mathrm{d}} = \frac{N_{\mathrm{det}}}{N_{\mathrm{gt}}}$$

$$F_{\mathrm{a}} = \frac{N_{\mathrm{FA}}}{N_{\mathrm{pixels}}}$$

Here, N is the number of evaluated frames; N_{det} is the number of correctly detected targets under the evaluation criterion; N_{gt} is the number of ground-truth targets; N_{FA} is the total number of pixels belonging to unmatched predicted connected components; and N_{pixels} is the total number of evaluated pixels. For P_d, a ground-truth target is considered correctly detected when the Euclidean distance between the centroid of a predicted connected component and that of the corresponding ground-truth component is less than 3 pixels. Prediction maps are binarized using a threshold of 0.5, and each predicted component can be matched to at most one ground-truth target. Pixels belonging to unmatched predicted components are counted as false-alarm pixels for the calculation of F_a.

IoU (Intersection over Union): The ratio of the intersection to the union of the predicted result and the ground truth. It better reflects the degree of overlap between the predicted region and the true region. nIoU (Normalized IoU): The average of the frame-wise IoU values over all evaluated frames. It measures the stability of the model at the sample level. P_d (Probability of Detection): The number of correctly detected targets divided by the total number of existing ground-truth targets. A higher P_d indicates fewer missed detections by the model. F_a (False Alarm Rate): The number

Just Accepted

of pixels belonging to unmatched predicted connected components divided by the total number of evaluated pixels. A lower F_a is better. FPS (Frames Per Second): The number of frames processed or transmitted per second in a video stream.

The results demonstrate that rectangular images are better suited to segmentation tasks than sector-shaped images. The LVNet model achieved an IoU of 48.64% on rectangular images compared to 28.68% on sector-shaped images, representing a 69.6% relative improvement. This performance difference can be attributed to the fact that deep learning models are typically optimized for rectangular grid structures, whereas sector-shaped polar coordinates introduce geometric distortions that complicate feature extraction.

3 Proposed method

3.1. Overview of LSNet

To address the three challenges identified in Section 1, we propose LSNet (Low-pixel Sonar Network), a multi-frame segmentation network adapted to underwater sonar images, as shown in Fig. 7. LSNet is developed from LVNet through three sonar-oriented improvements.

FLS images are affected by acoustic interference such as reverberation, multipath returns, sidelobes, and speckle-like intensity fluctuations, which can produce strong background responses similar to low-pixel target echoes. LSNet therefore exploits temporal information to distinguish moving-target responses from persistent or structured acoustic interference.

LVNet consists of two key components: a shallow Conv UNet and a deep Video Transformer UNet. The Conv UNet is used to enhance the local representation capability of patch embeddings, while the Video Transformer UNet focuses on spatio-temporal context modeling to effectively capture temporal dependencies and global representations. LVNet separates the shallow spatial stream from the deep temporal stream; features from the encoding stage of the UNet are directly concatenated into the decoding stage. By ignoring temporal information, this design allows sonar image noise that is not filtered out during the encoding stage to propagate directly into the shallow decoding stage. Moreover, the deep Transformer employs a 3D window self-attention mechanism, where numerous static background tokens receive considerable attention weights, thereby reducing the attention scores of true target tokens. To address these issues, we propose LSNet, which introduces three key improvements:

- Temporal information is introduced at the shallow encoding stage: variance is used to initially screen temporal information, extract motion information, and help filter out noise, thereby preventing useless information from being directly passed to the shallow decoding stage.
- The deep Video Transformer UNet adopts a decoupled 3D window self-attention module, which decomposes spatio-temporal modeling into two cascaded stages: local perception and pure temporal verification. Through pixel-level interactions along the pure temporal axis, spatial clutter interference is isolated.
- A hybrid optimizer and a joint BCE-Dice loss are introduced to improve optimization performance and mitigate the severe foreground-background imbalance.

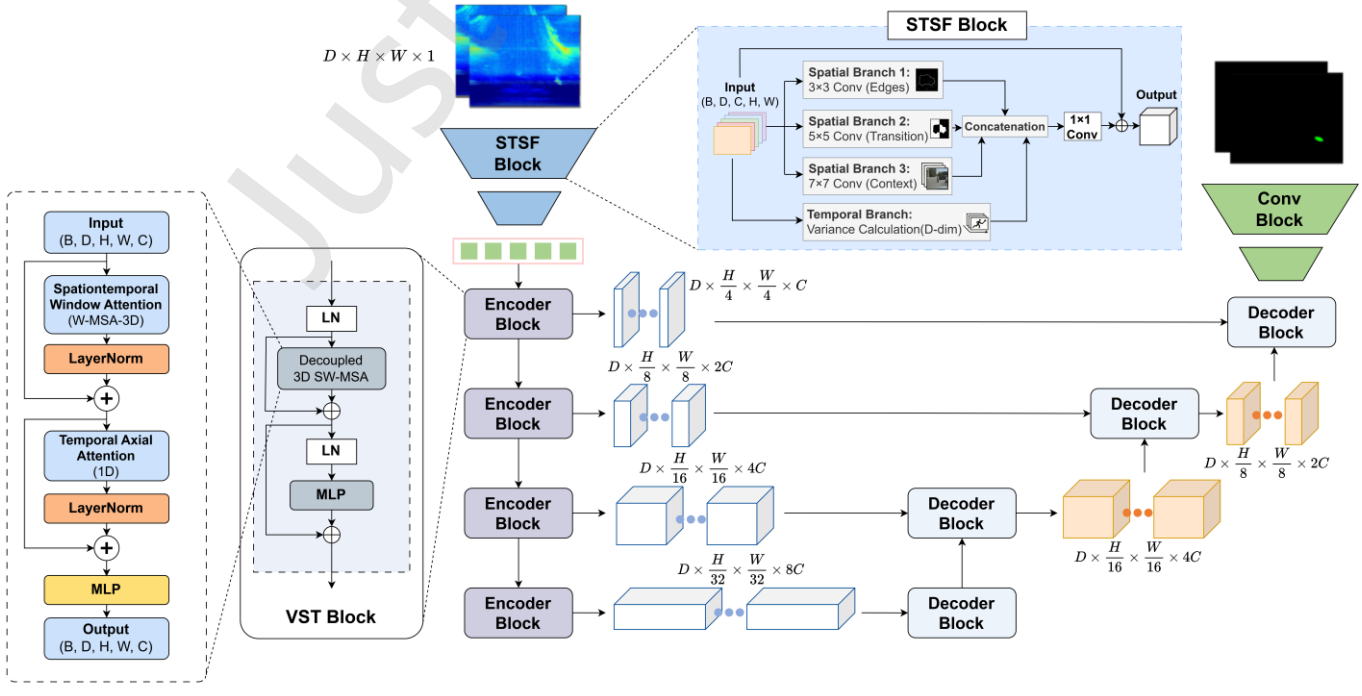


Fig. 7. Architecture diagram of LSNet.

Just Accepted

3.2. Spatio-temporal scale fusion block (STSF block)

The overall structure of the STSF block is shown in Fig. 8. The STSF block is designed based on the characteristics of sonar image changes across adjacent frames, primarily fusing spatial and temporal scale features. For the shallow stage, the Conv UNet structure is retained. The encoder consists of STSF blocks and downsampling layers, while the decoder consists of convolutional blocks and upsampling layers.

In the spatial dimension, multi-scale receptive fields are constructed using 3×3, 5×5, and 7×7 convolutional kernels to extract static spatial features. The 3×3 convolution extracts edge features, the 5×5 convolution emphasizes target-background transitions, and the 7×7 convolution models background context information.

In the temporal dimension, variance-based motion perception is adopted. Pixel-wise variance is computed along the temporal dimension D. The core principle is that persistent background structures and slowly varying acoustic interference generally exhibit smaller temporal variations than the echo responses produced by a moving target. Three spatial branches employ different convolutional kernels to extract multi-scale features. The input shape is transformed from (B, D, C, H, W) to (B × D, C, H, W), which involves changes in H and W and corresponds to static spatial features. On top of these three branches, an additional branch is added: variance is computed along the temporal dimension D from the original (B, D, C, H, W) input, which involves changes in D and corresponds to temporal motion features. The four branches operate in parallel and are then fused together. Residual connections are employed to maintain information flow and alleviate the gradient vanishing problem.

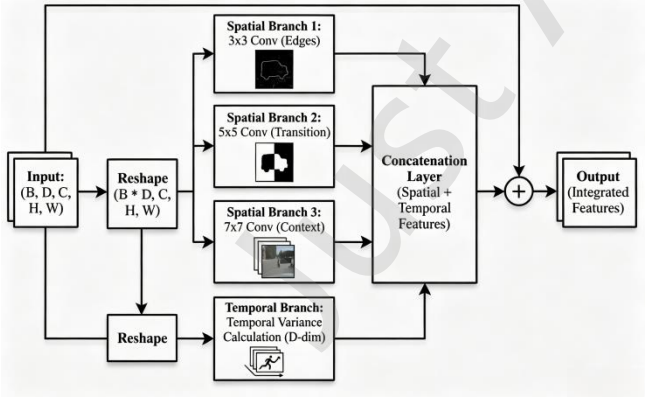


Fig. 8. STSF block.

3.3. Decoupled Swin block 3D (D3D)

The structure of the proposed D3D block is illustrated in Fig. 9. The 3D window self-attention mechanism in the deep Video Swin Transformer computes attention jointly within a local spatio-temporal cube composed of *(D, H, W)*. Because the softmax operation is globally normalized over all tokens within the 3D window, the overwhelmingly dominant static background tokens naturally consume the vast majority of the attention weights. This "background hegemony" is extremely detrimental to small target detection in low-pixel images—it severely reduces the attention scores of the true target tokens, essentially suppressing and diluting the weak motion signals at the feature interaction level.

This issue is particularly relevant to FLS imagery, where reverberation, sidelobes, multipath responses, and other structured acoustic clutter may occupy substantially more spatial locations than the low-pixel target itself.

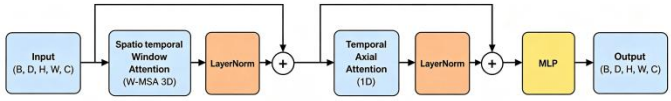


Fig. 9. D3D block.

Motivated by the above, we design the D3D module, which decouples the originally highly coupled 3D spatio-temporal attention into two cascaded independent processing stages.

Window Attention 3D: The original 3D window attention mechanism is retained to capture the micro-motion and shape contour of the target within a local neighborhood. Its structure is consistent with the original Swin block, ensuring the model's basic ability to perceive the local spatial topology and short-range cross-frame continuity of the target.

Temporal Axial Attention: After local perception, a lightweight temporal axial attention module is attached. This module reshapes the feature map into a sequence format of *(B × H × W, D, C)*, i.e., treats each spatial location as an independent one-dimensional sequence and performs self-attention only along the temporal axis *D*. This means that at this stage, each pixel only associates with itself across different frames, substantially reducing interference from other spatially highlighted backgrounds within the same frame.

The two stages are equipped with their own layer-norm and residual connections, executed serially, and share a multilayer perceptron for channel-wise feature reconfiguration. This decoupled design, with limited additional computational cost, endows the model with a dual perspective: it can see "what the target looks like" in the local spatial domain while verifying "whether the pixel is truly moving" along the pure temporal axis.

3.4. Muon-Adam optimizer and joint loss (OL)

(1) Muon-Adam Optimizer. Muon is a matrix-oriented optimizer that applies Newton-Schulz iterations to orthogonalize matrix-valued updates (Liu et al., 2025). In this study, we adopt a hybrid Muon-Adam strategy to accommodate different parameter structures within the network. Muon is applied to parameters with two or more dimensions, whereas Adam is retained for one-dimensional parameters. This hybrid strategy combines matrix-oriented optimization for convolutional and linear weights with conventional Adam updates for vector-valued parameters. Hybrid Muon-based optimization has also been adopted in recent real-time vision models, such as the MuSGD strategy used in YOLO26 (Sapkota et al., 2025).

(2) Joint Loss Function. To address the common challenges of extreme class imbalance and blurred object boundaries in sonar image segmentation, we propose to optimize the model using a joint loss function. This loss function consists

Just Accepted

of two complementary components: the Binary Cross-Entropy Loss (L_{BCE}), which provides smooth pixel-wise probability gradients, and the Dice Loss (L_{Dice}), which optimizes the global overlap between the predicted segmentation and the ground truth, thereby effectively overcoming the severe foreground-background class imbalance. Their combination considers both local pixel classification and the overall integrity of the target region.

In low-pixel sonar segmentation, background pixels substantially outnumber foreground target pixels, resulting in severe class imbalance. The traditional binary cross-entropy loss L_{BCE} evaluates the predicted probability of each pixel independently and is defined as follows:

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

where N is the total number of pixels in the image, $y_i \in \{0,1\}$ denotes the ground-truth label of pixel i, and $p_i \in (0,1)$ is the predicted probability that the pixel belongs to the foreground (obtained via the Sigmoid function). L_{BCE} effectively penalizes high-confidence incorrect predictions and provides stable and smooth gradient backpropagation for the network. However, in extremely imbalanced scenarios, a large number of easy background pixels dominate the gradient of L_{BCE}, causing the model to bias its predictions toward the background for all regions.

To compensate for the limitations of L_{BCE} in handling class imbalance, we introduce the Dice Loss. The Dice Loss originates from the Dice Similarity Coefficient. Its core idea is to directly optimize the overlap between the predicted mask and the ground-truth mask, and it is scale-invariant with respect to the size of the foreground region. Therefore, it naturally suppresses extreme pixel imbalance. The Dice Loss is defined as follows:

$$L_{Dice} = 1 - \frac{2\sum_{i=1}^{N} p_i y_i + \varepsilon}{\sum_{i=1}^{N} p_i + \sum_{i=1}^{N} y_i + \varepsilon} \quad (2)$$

where $p_i y_i$represents the intersection between the prediction and the ground truth. Since y_i takes values of 0 or 1, this product term effectively filters out background regions and retains only the overlapping part. The denominator accounts for the sum of pixels in both the prediction and the ground truth, effectively balancing the sizes of the foreground and background regions. ε is a smoothing coefficient (set to 1×10^{-5} in this paper), which primarily prevents the denominator from becoming zero when the target region is empty, and also helps smooth the gradient during the early stage of training.

Finally, we combine the pixel-wise L_{BCE} with the global L_{Dice} to construct the overall joint loss function as follows:

$$L_{Total} = \lambda_1 L_{BCE} + \lambda_2 L_{Dice} \quad (3)$$

where λ_1 and λ_2 are hyperparameters that balance the contributions of each term. In this paper, both are set to 0.5, giving equal weight to pixel-wise BCE supervision and region-overlap Dice supervision. This combined strategy enables the model to maintain pixel-wise prediction stability through L_{BCE} while attending to the global integrity of rare objects via L_{Dice}, thereby providing complementary supervision for low-pixel target segmentation.

4 Experimental evaluation

4.1. Experimental setup

Experiments were conducted on an RTX 4090 (24 GB) platform using PyTorch 2.8.0, Python 3.12, Ubuntu 22.04, and CUDA 12.8.

The 21 temporally ordered sequences were partitioned at the sequence level into 15 training sequences (5,097 frames), 3 validation sequences (1,018 frames), and 3 test sequences (1,021 frames), with no sequence shared between subsets. The training set comprised sequences 110354, 111430, 111443, 132644, 132808, 132939, 133404, 133618, 135251, 135628, 140449, 141050, 151007, 152113, and 160559; the validation set comprised sequences 111003, 153051, and 160334; and the test set comprised sequences 113200, 132412, and 152939. This sequence-disjoint partition prevents temporally adjacent frames from the same acquisition sequence from appearing in different subsets.

For LSNet training, comparative experiments involving LSNet, and ablation experiments, each multi-frame input consisted of four consecutive frames ($D = 4$) sampled at a one-frame interval, and non-overlapping clips were constructed with a stride of four frames. The original rectangular B-scan images had a resolution of 1,000 × 500 pixels (width × height) and were zero-padded to 1,024 × 512 pixels to satisfy the multiple-of-32 input requirement, without geometric resizing. For these experiments, the batch size was 1, the initial learning rate was 1×10^{-4}, and training was conducted for 50 epochs without data augmentation. A random seed of 2026 was used to provide reproducible and consistent stochastic conditions across the controlled experiments. All reported results correspond to single controlled runs under this fixed seed. For the comparative experiments, all methods were evaluated under the same dataset partition, preprocessing pipeline, input resolution, and evaluation protocol. For the comparative experiments, all methods were evaluated under the same dataset partition, preprocessing pipeline, input resolution, and evaluation protocol. For each method, the optimizer and loss function were selected according to the characteristics of the corresponding framework and the requirements of the binary segmentation task. Specifically, LVNet was reproduced using the Adam optimizer following its reported optimization strategy, while LSNet was trained using the Muon-Adam optimizer. The BCE-Dice loss was adopted for LSNet and LVNet reproduction to adapt the segmentation objective to the present binary segmentation task.

During training, the validation set was used for model selection and checkpoint determination. Specifically, the checkpoint achieving the highest IoU on the validation set during training was selected as the final model for subsequent test evaluation. The test set was not involved in checkpoint selection or hyperparameter tuning. The equal-weight BCE-Dice setting was fixed before the sensitivity analysis in Table 7. Table 7 was subsequently added to

Just Accepted

examine the sensitivity of the already determined loss configuration to different weighting coefficients.

4.2. Comparative results

LVNet (Shen et al., 2025) was selected as the development baseline because LSNet was derived from its multi-frame spatio-temporal architecture. In the comparative experiments, LVNet was reproduced based on its reported optimization strategy, while other comparison methods were reimplemented or retrained according to their original frameworks whenever applicable. All methods were evaluated under the unified experimental protocol described in Section 4.1.

Four representative multi-frame infrared small-target methods were further evaluated. DeepPro-Plus (Li et al., 2026) exploits temporal intensity profiles across consecutive frames to enhance weak moving-target responses. RFR (Ying et al., 2025), originally developed for infrared small-target detection in satellite videos, recurrently refines features across frames to aggregate temporal information. DTUM (Li et al., 2025) introduces a direction-coded temporal U-shaped module to model directional motion patterns. LMAFormer (Huang et al., 2024) employs a local motion-aware Transformer to capture local spatio-temporal dependencies of small moving targets. These methods were selected because they represent complementary temporal modeling strategies, including temporal-profile modeling, recurrent refinement, directional motion encoding, and Transformer-based local motion modeling. For evaluation consistency, all methods were evaluated using the same sequence-level train/validation/test partition, image preprocessing, and input resolution. The training schedule was set to 50 epochs with a batch size of 1, an initial learning rate of 1×10^{-4}, and no data augmentation. The multi-frame methods used four consecutive frames sampled at a one-frame interval, whereas U-Net used a single-frame input according to its original architecture.

Representative qualitative results from the test set are shown in Fig. 10. For each example, the left, middle, and right panels present the original sonar image, ground-truth mask, and predicted mask, respectively. Besides successful segmentation cases, Fig. 10 also includes representative failure cases, including missed detection and false alarm examples. These failure cases indicate that weak target responses and target-like background interference remain challenging for accurate segmentation in FLS images. As shown in Table 4, LSNet achieves the highest IoU, nIoU, and Pd among all compared methods, reaching 68.21%, 67.39%, and 50.46%, respectively, while maintaining an Fa of 777.55 $\times 10^{-6}$ and an inference speed of 91.23 FPS. Under the corresponding training configurations, LSNet improves IoU, nIoU, and Pd by 19.57, 16.67, and 8.80 percentage points, respectively, and reduces Fa by 294.53 $\times 10^{-6}$.

Among the five additional methods, LMAFormer achieves the strongest performance, whereas U-Net achieves the lowest false-alarm rate. Compared with LMAFormer, LSNet improves IoU, nIoU, and Pd by 5.45, 4.62, and 2.96 percentage points, respectively, while reducing Fa by 260.73 $\times 10^{-6}$. DTUM obtains an IoU of 59.31% and an Fa of 855.59 $\times 10^{-6}$, but its lower nIoU and Pd indicate less stable sample-level overlap and more missed targets. RFR yields moderate performance, while DeepPro-Plus shows relatively low IoU and high Fa on the present sonar dataset, suggesting that its temporal representation strategy may not fully match the characteristics of the present FLS segmentation task.

Although U-Net achieves the lowest Fa, its substantially lower IoU, nIoU, and Pd indicate a more conservative prediction tendency accompanied by more missed targets. It should be noted that the Pd value is calculated based on the predefined three-pixel centroid matching criterion. Due to the extremely limited pixel occupancy of AUV targets in FLS images, slight localization deviations may result in unmatched predictions, which partially limits the absolute Pd value. Nevertheless, under the same evaluation criterion, LSNet achieves the highest Pd among all compared methods, demonstrating improved target perception capability.

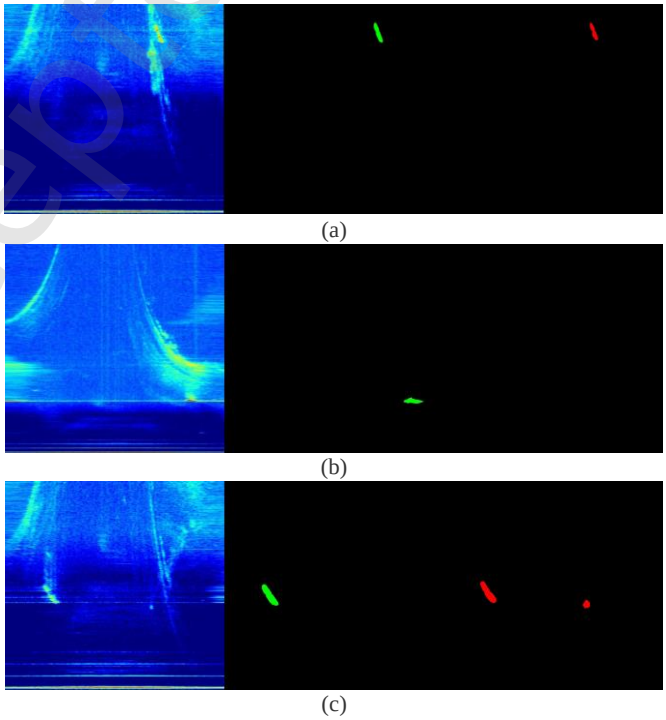


Fig. 10. Qualitative results of LSNet on the USMT dataset. The left, middle, and right panels represent the original sonar images, ground-truth masks, and predicted masks, respectively. (a) successful segmentation case; (b) missed detection case; (c) false alarm case.

Table 4. Comparative segmentation performance on the USMT dataset.

Algorithm	IoU(%)	nIoU(%)	Pd(%)	Fa($\times 10^{-6}$)	FPS
LVNet	48.64	50.72	41.66	1072.08	103.69
DeepPro-Plus	10.68	23.47	39.58	8645.90	26.35
U-Net	45.00	35.27	23.52	473.19	109.32
RFR	49.15	51.70	42.65	1290.36	76.65
DTUM	59.31	49.29	35.92	855.59	24.85
LMAFormer	62.76	62.77	47.50	1038.28	12.03
LSNet (Ours)	68.21	67.39	50.46	777.55	91.23

Just Accepted

4.3. Ablation study

To separately evaluate the contributions of the optimization strategy, loss function, and architectural components, we conducted two complementary ablation studies on USMT using the same 50-epoch training protocol. First, the original LVNet architecture was fixed, and four combinations of Adam or Muon-Adam with Dice or BCE-Dice loss were evaluated to isolate the effects of the optimizer and loss function (Table 5). The Adam+BCE-Dice setting corresponds to the retrained LVNet reference reported in Table 4. Second, STSF and D3D were evaluated independently under the Adam+Dice reference setting, while the progressive combinations with OL were retained to assess their contributions under the final training configuration (Table 6).

Compared with Adam+Dice, replacing Dice with BCE-Dice under the Adam optimizer improves IoU from 43.85% to 48.64%, nIoU from 45.00% to 50.72%, and Pd from 36.23% to 41.66%, while reducing Fa from 1110.84×10^{-6} to 1072.08×10^{-6}. Replacing Adam with Muon-Adam while retaining Dice produces a larger improvement in overlap performance, increasing IoU to 60.73% and nIoU to 59.33% and reducing Fa to 756.11×10^{-6}. When Muon-Adam and BCE-Dice are combined as the OL setting, IoU reaches 61.46% and Pd increases to 46.77%, while Fa is further reduced to 714.43×10^{-6}. These results indicate that the optimizer provides the larger improvement in IoU and nIoU, whereas the joint loss provides additional benefits in target detection and false-alarm suppression.

Table 5. Ablation study of the optimizer and loss function.

Optimizer	Loss	IoU(%)	nIoU(%)	Pd(%)	Fa ($\times 10^{-6}$)	FPS
Adam	Dice	43.85	45.00	36.23	1110.84	104.39
Adam	BCE-Dice	48.64	50.72	41.66	1072.08	103.69
Muon-Adam	Dice	60.73	59.33	40.09	756.11	101.75
Muon-Adam	BCE-Dice	61.46	58.57	46.77	714.43	105.46

We further evaluated the architectural components independently and in combination with OL, as summarized in Table 6. STSF-only and D3D-only were evaluated under the same Adam+Dice reference setting to determine whether the two architectural modifications remain effective independently of the proposed optimization and loss strategy. The progressive combinations with OL were retained to examine their contributions under the final training configuration.

Table 6. Ablation study of the proposed components and their combinations.

Setting	IoU(%)	nIoU(%)	Pd(%)	Fa($\times 10^{-6}$)	FPS
Reference (Adam+Dice)	43.85	45.00	36.23	1110.84	104.39
STSF	52.04	53.17	44.65	895.62	102.63
D3D	49.11	49.62	39.00	862.78	93.29
OL	61.46	58.57	46.77	714.43	105.46
OL+STSF	65.09	63.32	45.57	786.08	100.19
OL+D3D	66.33	65.72	44.11	812.17	94.55
OL+STSF+D3D	68.21	67.39	50.46	777.55	91.23

Under the same Adam+Dice reference setting, introducing STSF alone increases IoU from 43.85% to 52.04%, nIoU from 45.00% to 53.17%, and Pd from 36.23% to 44.65%, while reducing Fa from 1110.84×10^{-6} to 895.62×10^{-6}. D3D alone also improves IoU and nIoU to 49.11% and 49.62%, respectively, and reduces Fa to 862.78×10^{-6}. These independent results demonstrate that both architectural modifications remain effective without OL.

When introduced on top of OL, STSF increases IoU and nIoU from 61.46% and 58.57% to 65.09% and 63.32%, respectively, while D3D increases them to 66.33% and 65.72%. Although Pd decreases slightly and Fa increases in these two individual combinations, the full model combining OL, STSF, and D3D achieves the highest IoU, nIoU, and Pd, reaching 68.21%, 67.39%, and 50.46%, respectively, while maintaining an inference speed of 91.23 FPS. Overall, the results show that STSF and D3D provide independent improvements and further complement the OL strategy when combined in the complete LSNet.

To examine the sensitivity of the joint loss to its weighting coefficients, different combinations of λ_1 and λ_2 were evaluated using the complete LSNet while keeping all other settings unchanged. The results are summarized in Table 7.

Table 7. Sensitivity analysis of the joint-loss weights.

λ_1 (BCE)	λ_2 (Dice)	IoU (%)	nIoU (%)	Pd (%)	Fa ($\times 10^{-6}$)
0.25	0.75	66.92	63.76	49.81	660.88
0.50	0.50	68.21	67.39	50.46	777.55
0.75	0.25	66.49	65.53	46.82	790.15

The sensitivity results show that under the fixed-seed setting, the equal-weight setting ($\lambda_1 = \lambda_2 = 0.5$) achieves the highest observed IoU, nIoU, and Pd among the evaluated settings. Increasing the Dice weight to 0.75 reduces Fa to 660.88×10^{-6}, but decreases IoU and nIoU. Assigning a larger weight to BCE also decreases IoU, nIoU, and Pd and slightly increases Fa relative to the equal-weight setting. Because each configuration was evaluated in a single controlled run, the relatively small differences among these settings are interpreted as descriptive sensitivity trends rather than statistically significant differences. These results provide a descriptive sensitivity assessment of the equal-weight setting used in the final model.

4.4. Discussion

The results indicate that data representation and temporal modeling address complementary aspects of the task. Under the same retrained LVNet configuration, converting sector-shaped images to the rectangular B-scan representation increases the IoU from 28.68% to 48.64%. Using the rectangular representation, LSNet further raises the IoU to 68.21%. Its additional advantage over LMAFormer and the

Just Accepted

other multi-frame comparators shows that the gain is not limited to the choice of parent architecture; regular image geometry and sonar-oriented spatio-temporal modeling both contribute to performance.

The ablation results further clarify the contributions of the proposed optimization, loss, and architectural components. The controlled optimizer-loss ablation shows that replacing Adam with Muon-Adam produces the larger improvement in IoU and nIoU, while the BCE-Dice joint loss provides additional benefits in detection probability and false-alarm suppression. The independent architectural ablations further show that both STSF and D3D improve performance without OL. When introduced on top of OL, the two modules provide further gains in IoU and nIoU, and their full combination achieves the highest IoU, nIoU, and Pd while retaining real-time inference.

Across the additional comparison methods, LSNet provides strong overall performance in region overlap, sample-level stability, and detection probability while maintaining a relatively low false-alarm rate. Although U-Net achieves a lower Fa, its substantially lower IoU, nIoU, and Pd indicate a trade-off between conservative foreground prediction and target detection capability. The improvements over LMAFormer suggest that variance-guided shallow fusion helps preserve weak motion cues, while the decoupled deep attention stage verifies temporal consistency without allowing prominent static background structures to dominate feature interactions. With an inference speed of 91.23 FPS, LSNet may support online external AUV target monitoring in sonar-based testing and safety-observation scenarios.

The intended application evaluated in this study is external FLS-based monitoring of a moving AUV. Because the sonar remained fixed during data acquisition, the present experiments do not represent onboard AUV perception involving sensor motion, attitude variation, or platform vibration.

However, the present evaluation uses a single AUV, a single FLS device, and a controlled freshwater lake environment. Differences in acoustic propagation, reverberation, multipath effects, environmental clutter, and target aspect angles may introduce a domain gap when transferring the method to coastal, harbour, or open-sea environments. Therefore, the reported results should be interpreted within the current controlled monitoring scenario. In addition, all experiments were conducted with a fixed random seed to ensure controlled and reproducible comparisons. Although repeated trials with multiple random seeds could further characterize statistical variation, such analysis will be considered in future evaluations with larger and more diverse datasets.

5 Conclusion

This study addresses low-pixel AUV target perception in FLS images by constructing the publicly available USMT dataset and proposing LSNet. USMT contains 21 temporally ordered sequences and 7,136 annotated rectangular B-scan FLS frames. Technical validation compares sector-shaped and rectangular representations, while the released USMT dataset provides the rectangular B-scan representation used in the subsequent experiments.

LSNet extends the LVNet multi-frame infrared small-target segmentation architecture through the STSF block, D3D block, and the OL strategy comprising a Muon-Adam optimizer and a joint BCE-Dice loss. On USMT, LSNet achieves an IoU of 68.21% and an inference speed of 91.23 FPS. It outperforms the retrained LVNet reference by 19.57 percentage points in IoU and exceeds U-Net and four additional multi-frame comparison methods, including a 5.45-percentage-point IoU gain over the strongest of them, LMAFormer. The ablation results show that Muon-Adam and the BCE-Dice loss make distinct contributions to performance, while STSF and D3D remain effective independently and provide further complementary gains in the complete LSNet.

The released dataset and real-time network provide a basis for external sonar-based AUV target monitoring. Nevertheless, USMT was collected using a fixed FLS device while observing a single AUV in a controlled lake environment, which may introduce a domain gap in more complex marine conditions. Future work will extend the dataset to additional target types and sea-trial scenarios and further investigate cross-environment generalization.

Author contributions

Jingfeng Yu: Conceptualization, Methodology, Writing – review & editing, Validation. Chen Liang: Methodology, Software, Validation, Formal analysis, Data curation, Writing – original draft. Kailai Sun: Formal analysis, Supervision, Validation, Writing – review & editing. Jiugen Lin: Investigation, Data curation, Validation. Zhongju Sun: Investigation, Software, Data curation. Ruiming Wang: Software, Validation, Data curation. Qianchuan Zhao: Supervision, Writing – review & editing.

Replication and data sharing

The USMT dataset generated and analyzed during the current study is publicly available in the ScienceDB repository at https://doi.org/10.57760/sciencedb.40992.
The replication package associated with this study has been published in ETS-Data and is publicly available at https://doi.org/10.26599/ETSD.2026.9190036.

Acknowledgements

This research was supported by the Ganjiang Innovation Academy, Chinese Academy of Sciences.

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

Just Accepted

References

Dahn, N., Firvida, M.B., Sharma, P., Christensen, L., Geisle, O., Mohrmann, J., et al., 2024. An acoustic and optical dataset for the perception of underwater unexploded ordnance (UXO). In: Proceedings of OCEANS 2024-Halifax. IEEE, pp. 1–6.

Gomes, L.S., Almeida, G.P., Moreira, B.U., Quiroz, M., Xavier, B., Soares, L., et al., 2025. Sonar image datasets: A comprehensive survey of resources, challenges, and applications. In: Proceedings of the 38th SIBGRAPI Conference on Graphics, Patterns and Images. IEEE, pp. 1–6.

Huang, Y., Zhi, X., Hu, J., Yu, L., Han, Q., Chen, W., et al., 2024. LMAFormer: Local motion aware transformer for small moving infrared target detection. IEEE Trans Geosci Remote Sens 62, 5008117. doi: 10.1109/TGRS.2024.3502663.

Jiao, W., Zhang, J., Zhang, C., 2024. Open-set recognition with long-tail sonar images. Expert Syst Appl 249, 123495.

Li, R., An, W., Wang, Y., Ying, X., Dai, Y., Wang, L., et al., 2026. Probing deep into temporal profile makes the infrared small target detector much better. IEEE Trans Pattern Anal Mach Intell 48(8), 10157–10175. doi: 10.1109/TPAMI.2026.3683258.

Li, R., An, W., Xiao, C., Li, B., Wang, Y., Li, M., et al., 2025. Direction-coded temporal U-shape module for multiframe infrared small target detection. IEEE Trans Neural Netw Learn Syst 36(1), 555–568. doi: 10.1109/TNNLS.2023.3331004.

Liu, J., Su, J., Yao, X., Jiang, Z., Lai, G., Du, Y., et al., 2025. Muon is scalable for LLM training. arXiv:2502.16982.

Manhães, M.M.M., Scherer, S.A., Voss, M., Douat, L.R., Rauschenbach, T., 2016. UUV simulator: A Gazebo-based package for underwater intervention and multi-robot simulation. In: Proceedings of OCEANS 2016 MTS/IEEE Monterey. IEEE, pp. 1–8.

Potokar, E., Ashford, S., Kaess, M., Mangelson, J.G., 2022a. HoloOcean: An underwater robotics simulator. In: Proceedings of the 2022 International Conference on Robotics and Automation (ICRA). IEEE, pp. 3040–3046.

Potokar, E., Lay, K., Norman, K., Benham, D., Neilsen, T.B., Kaess, M., et al., 2022b. HoloOcean: Realistic sonar simulation. In: Proceedings of the 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, pp. 8450–8456.

Prats, M., Pérez, J., Fernández, J.J., Sanz, P.J., 2012. An open source tool for simulation and supervision of underwater intervention missions. In: Proceedings of the 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, Vilamoura, Portugal, 7–12 October 2012, pp. 2577–2582.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Lecture Notes in Computer Science, vol. 9351. Springer, Cham, pp. 234–241. doi: 10.1007/978-3-319-24574-4_28.

Sapkota, R., Cheppally, R.H., Sharda, A., Karkee, M., 2025. YOLO26: Key architectural enhancements and performance benchmarking for real-time object detection. arXiv:2509.25164.

Shen, Z., Chen, S., Zhang, T., Wang, H., Zhang, Z., Ji, R., et al., 2025. Low-level matters: An efficient hybrid architecture for robust multi-frame infrared small target detection. IEEE J Sel Top Appl Earth Obs Remote Sens 18, 23757–23766. doi: 10.1109/JSTARS.2025.3607473.

Sun, H., Bai, J., Yang, F., Bai, X., 2023. Receptive-field and direction induced attention network for infrared dim small target detection with a large-scale dataset IRDST. IEEE Trans Geosci Remote Sens 61, 1–13.

Xie, K., Yang, J., Qiu, K., 2022. A dataset with multibeam forward-looking sonar for underwater object detection. Sci Data 9, 739.

Xin, B., Mao, Q., Wang, B., Zheng, J., Zhao, Y., Chen, C.L.P., et al., 2026. 10K is enough: An ultra-lightweight binarized network for infrared small-target detection. IEEE Trans Geosci Remote Sens. doi: 10.1109/TGRS.2026.3683449.

Yang, Z., Zhao, J., Zhang, H., Yu, Y., Huang, C., 2023. A side-scan sonar image synthesis method based on a diffusion model. J Mar Sci Eng 11, 1103.

Ying, X., Liu, L., Lin, Z., Shi, Y., Wang, Y., Li, R., et al., 2025. Infrared small target detection in satellite videos: A new dataset and a novel recurrent feature refinement framework. IEEE Trans Geosci Remote Sens 63, 5002818. doi: 10.1109/TGRS.2025.3542368.

Yu, J., Wan, H., Sun, Z., Ma, X., He, Y., Zhao, Q., 2026. UDRSim: An underwater dynamic real-time multimodal imaging simulator. In: Proceedings of the Eighteenth International Conference on Machine Vision (ICMV 2025), vol. 14114. SPIE, pp. 599–606.



Jingfeng Yu is currently pursuing the Doctor of Engineering degree in Innovative Leading Engineering at Tsinghua University, Beijing, China, majoring in electronic information. He conducts full-time scientific research at the Ganjiang Innovation Academy, Chinese Academy of Sciences. In addition, he serves as a master's supervisor with the University of Science and Technology of China. His research interests mainly include autonomous underwater vehicles (AUVs), underwater intelligent perception, cross-domain multimodal target recognition and underwater acoustic signal processing.

Just Accepted



Chen Liang received the B.E. degree in Automation from Anhui University, Hefei, China. He is currently pursuing the M.S. degree in Electronic Information at the University of Science and Technology of China, Hefei, China. His research interests include computer vision, underwater intelligent perception, autonomous underwater vehicles (AUVs), intelligent control, and deep learning.



Kailai Sun received the B.S. degree in automation from the Beijing University of Posts and Telecommunications, Beijing, China, in 2018, the Ph.D. degree in control science and engineering, Tsinghua University, Beijing, China, in 2023. He was a Research Fellow with the College of Design and Engineering, National University of Singapore, in 2024. He is currently a postdoc with the SMART, Massachusetts Institute of Technology. He has authored or co-authored more than 40 peer-reviewed journal and conference papers, including about 20 first-authored publications. His research interests include computer vision, world model, agentic AI, remote sensing, AI for science, smart buildings and cities.



Jiugen Lin is currently pursuing a doctoral degree in engineering at Harbin Engineering University in China, specializing in the field of electronic information. He is engaged in full-time research at the Ganjiang Innovation Academy, Chinese Academy of Sciences. His main research areas include autonomous underwater vehicles (AUVs), underwater intelligent unmanned systems, underwater combat simulation, and underwater acoustic signal processing.



Zhongju Sun received his M.S. degree in Pattern Recognition and Intelligent Systems from Bohai University in 2024. He is currently working as an Algorithm Engineer at the Ganjiang Innovation Academy, Chinese Academy of Sciences. His research interests include 3D point cloud perception, 2D data perception, and underwater environment simulation.



Ruiming Wang received his B.E. degree in Electrical Engineering from Jilin University, Changchun, China. He is currently pursuing a Master's degree in Control Engineering at the University of Science and Technology of China, Hefei, China. His research interests include computer vision, multimodal recognition, and machine learning.



Qianchuan Zhao received the B.E. degree in automatic control, the B.S. degree in applied mathematics, in 1992, and the M.S. and Ph.D. degrees in control theory and its applications from Tsinghua University, Beijing, China, in 1996. He was a Visiting Scholar with Carnegie Mellon University, Pittsburgh, PA, USA, in 2000; and Harvard University, Cambridge, MA, USA, in 2002. He was a Visiting Professor with Cornell University, Ithaca, NY, USA, in 2006. He is currently a Professor and the Director of the Center for Intelligent and Networked Systems (CFINS), Department of Automation, and the BNRist, Tsinghua University. He has published more than 100 research papers in peer reviewed journals and conferences. His current research interests include controlling and optimizing complex networked systems with applications in smart buildings, smart grid, and manufacturing automation. Prof. Zhao received the 2009 and 2018 China National Nature Science Awards and the 2014 National Science Foundation for Distinguished Young Scholars of China. He is currently the Editor-in-Chief of the Results in Control and Optimization (RICO), an Editor of the IEEE TRANSACTIONS ON AUTOMATION SCIENCE AND ENGINEERING (T-ASE), and an Associate Editor of the Journal of Optimization Theory and Applications.

Just Accepted