Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

UDR-Mamba: A Depth-Structured Reliability-Guided Mamba Framework for RGB-D Underwater Salient Object Detection

Ning Li
Changzhou University
Jiaxin Li
Changzhou University
Xiwen Yang
Changzhou University
Jianhang Shi
Changzhou University
Haojie Wu
Changzhou University
Gaochao Yang
00003463@cczu.edu.cn

Changzhou University

Research Article

Keywords: underwater salient object detection, RGB-D perception, depth structuralization, reliability-gated fusion, visual state-space model, Mamba

Posted Date: September 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10496350/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

UDR-Mamba: A Depth-Structured Reliability-Guided Mamba Framework for RGB-D Underwater Salient Object Detection

Ning Li[1], Jiaxin Li[1], Xiwen Yang[1], Jianhang Shi[1], Haojie Wu[1], Gaochao Yang[1*]

[1*]School of Computer Science and Artificial Intelligence, Changzhou University, No. 21 Gehu Road, Wujin District, Changzhou, 213164, Jiangsu, China.

*Corresponding author(s). E-mail(s): 00003463@cczu.edu.cn;
Contributing authors: lncczu@cczu.edu.cn; 1666078528@qq.com;
2389609105@qq.com; 1562533951@qq.com; 2235853651@qq.com;

Abstract

Underwater salient object detection (USOD) plays an important role in robotic navigation, marine perception, ecological monitoring, and vision-guided manipulation; however, degraded color appearance and unreliable depth measurements make accurate segmentation challenging. To exploit geometric cues from noisy depth maps while reducing the redundancy of conventional two-stream RGB–D models, we propose UDR-Mamba, a depth-structured and reliability-guided framework equipped with a shared VMamba encoder. Specifically, a Depth Structuralization Input Module transforms each raw depth map into normalized depth, Sobel depth edges, and local depth contrast, thereby providing richer structural representations. A Depth-Reliable Mamba Fusion Module then estimates scale-specific depth reliability based on RGB features, depth features, cross-modal discrepancies, and depth-edge cues, enabling the selective integration of trustworthy depth information. Subsequently, a Mamba pyramid decoder aggregates multiscale fused features to generate the final saliency prediction, which is supervised using binary cross-entropy and soft intersection-over-union losses. Experimental results on the USOD10K dataset show that UDR-Mamba achieves an $\boldsymbol{S_m}$ score of 0.9332, a maximum E-measure of 0.9713, a maximum F-measure of 0.9365, and an MAE of 0.0192. On the USOD dataset, it achieves the highest maximum E-measure of 0.9573, the second-highest maximum F-measure of 0.9314, and the lowest MAE of 0.0359. Extensive ablation studies further verify

the effectiveness of depth structuralization, reliability-gated fusion, and encoder weight sharing. These results demonstrate that explicit depth-structure modeling and reliability-guided Mamba fusion improve boundary-aware multimodal representations, while the shared encoder provides a favorable trade-off between accuracy and efficiency in complex underwater scenes.

Keywords: underwater salient object detection, RGB-D perception, depth structuralization, reliability-gated fusion, visual state-space model, Mamba

1 Introduction

Underwater salient object detection (USOD) aims to segment the most visually conspicuous object regions from complex underwater scenes. It provides an important foundation for tasks such as underwater robot navigation, marine object perception, ecological monitoring, and vision-guided manipulation. Early saliency studies were primarily inspired by human visual attention and localized objects using handcrafted cues, including local contrast, global contrast, color differences, and regional saliency [1, 4, 14]. With the development of deep learning, convolutional neural network (CNN)-based salient object detection methods have markedly improved detection accuracy through hierarchical semantic representations, contextual aggregation, and boundary refinement [5, 12, 19, 21, 22, 30, 34, 35, 39, 44]. However, underwater images are commonly affected by illumination attenuation, color shifts, scattering, low contrast, blurred textures, and weakened boundaries [18], which substantially reduce the discriminative power of RGB appearance features. Recent studies corroborate this difficulty: RG-YOLO identifies low contrast and limited illumination as major obstacles to underwater target detection, while a dual-branch restoration network explicitly recovers high-frequency details degraded in underwater imagery [7, 49]. When salient objects have colors, textures, or shapes similar to those of the background, models that rely solely on RGB information are prone to false detections and missed objects.

To illustrate the role of depth-structural cues in underwater salient object detection more intuitively, Fig. 1 presents several underwater examples and their corresponding depth-structure visualizations. Underwater RGB images often exhibit color degradation, low contrast, and blurred object boundaries, whereas normalized depth maps provide information about the overall spatial layout and depth edges highlight regions of structural discontinuity between objects and their backgrounds. In particular, the depth edges are highly consistent with the ground-truth (GT) boundaries in several examples and reveal potential object contours clearly when overlaid on the RGB images. This observation indicates that depth maps contain not only distance information but also structural cues that can support boundary localization and foreground-background separation. Meanwhile, some depth-edge maps contain noise and incomplete responses, further demonstrating that depth information should not be fused indiscriminately but instead requires structural modeling and reliability screening.

Depth maps provide geometric and structural information complementary to RGB appearance, including spatial layout, abrupt changes at object boundaries, and foreground-background distance differences. Existing studies on RGB-D salient object detection have demonstrated that the appropriate incorporation of depth cues can effectively improve salient object localization [2, 3, 8, 9, 15, 20, 23, 29, 31, 32, 37, 42, 43, 47]. Nevertheless, we observe two key challenges in underwater RGB-D salient object detection. First, raw depth maps are not always reliable. In complex underwater environments, they may contain noise, missing regions, inaccurate boundaries, or weak local structural responses. Directly concatenating, adding, or attentionally fusing depth features with RGB features may therefore introduce erroneous geometric information. Second, many methods treat a depth map merely as a single-channel auxiliary input and do not fully exploit its boundary discontinuities and local geometric contrast. We therefore argue that underwater RGB-D salient object detection requires not only the use of depth information but also the construction of a saliency-oriented depth-structure representation and the adaptive assessment of depth-cue reliability during cross-modal fusion.

In recent years, Transformers and visual state-space models have provided powerful long-range dependency modeling capabilities for dense prediction tasks. Methods such as ViT, Swin Transformer, and PVT have advanced global modeling in computer vision [6, 27, 38]. However, self-attention generally incurs substantial computational overhead in high-resolution dense prediction. Mamba performs linear-time sequence modeling through a selective state-space mechanism [10], while Vision Mamba and VMamba further extend state-space modeling to visual tasks [25, 50]. These models offer strong global-context modeling capabilities and favorable computational efficiency, making them well suited to tasks such as USOD that depend on multiscale structures and long-range foreground-background relationships. Nevertheless, simply replacing a CNN or Transformer backbone with VMamba does not fully resolve the challenges of underwater RGB-D scenes. The key questions remain how to construct reliable depth-structural cues and how to perform adaptive cross-modal fusion when depth quality is uncertain.

Based on these observations, we propose UDR-Mamba, a shared-VMamba framework guided by depth structuralization and reliability for RGB-D underwater salient object detection. We first design a Depth Structuralization Input Module (DSI) that transforms a raw depth map into a three-channel structural representation comprising normalized depth, depth edges, and local depth contrast. This design enables the depth branch to perceive spatial distributions, boundary discontinuities, and local geometric variations explicitly. We then adopt a two-stream encoding strategy with a shared VMamba backbone. The RGB image and depth-structure map undergo two modality-specific forward passes, while only one set of VMamba backbone parameters is optimized, reducing parameter redundancy while retaining the representational capacity of RGB-D modeling. Furthermore, we introduce a layer-wise Depth-Reliable Mamba Fusion Module (DRMF), which estimates a reliability gate from the RGB features, depth features, cross-modal discrepancies, and depth-edge cues. Depth information is thereby injected selectively into the RGB representation rather than fused indiscriminately.

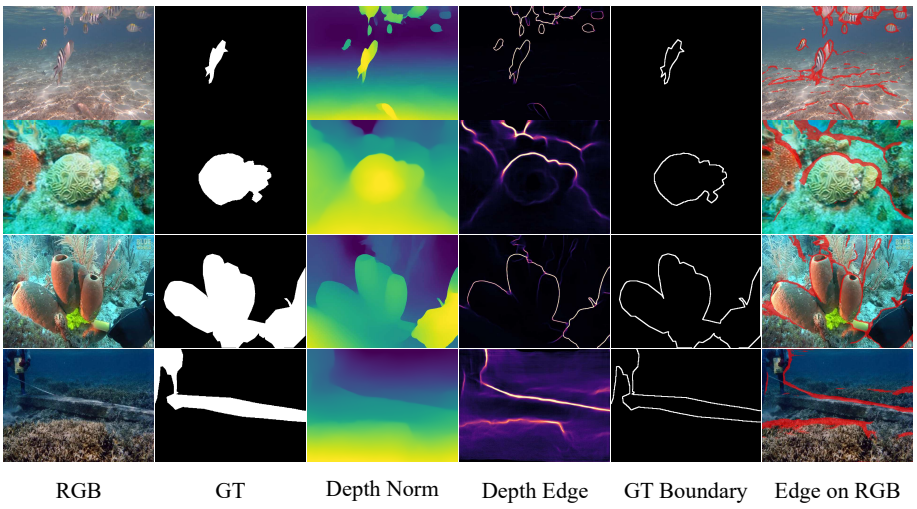


Fig. 1 Visualization of depth-structural cues. From left to right, the columns show the RGB image, ground-truth mask, normalized depth, depth edges, ground-truth boundary, and depth edges overlaid on RGB

Finally, we construct a single-head Mamba Pyramid Decoder that aggregates the multiscale fused features in a top-down manner and incorporates depth-edge cues during detail refinement. Unlike methods that rely on multiple auxiliary prediction heads and manually weighted losses, our approach uses only the final saliency prediction and a simple training objective comprising BCEWithLogits and soft IoU. Our main contributions are summarized as follows:

1. We propose a two-stream VMamba framework for RGB-D underwater salient object detection that exploits visual state-space modeling to enhance long-range contextual reasoning in complex underwater scenes.
2. We design a Depth Structuralization Input Module (DSI) that explicitly transforms a raw depth map into normalized depth, depth edges, and local depth contrast, thereby more fully exploiting its geometric and boundary cues.
3. We introduce a layer-wise Depth-Reliable Mamba Fusion Module (DRMF) that adaptively estimates depth reliability from cross-modal discrepancies and depth-edge information, enabling the selective injection of trustworthy depth information.

2 Related Work

2.1 Salient Object Detection and Underwater Salient Object Detection

Salient object detection has long been an important dense prediction task in computer vision. Early methods primarily relied on visual-attention mechanisms and hand-crafted features to detect salient regions through local contrast, frequency-domain analysis, and global regional contrast [1, 4, 14]. Deep learning methods subsequently

drove further progress in this field. DHS employs multiscale deep features for saliency prediction [19]; DSS strengthens hierarchical feature representations through deep supervision and short connections [12]; and Amulet aggregates multilevel convolutional features to balance semantic information and spatial details [44]. Subsequently, R3Net progressively refines predictions through recurrent residual refinement [5], PiCANet introduces pixel-wise contextual attention to model global relationships [22], and PoolNet, BASNet, F^3Net, U^2-Net, and MINet further improve saliency-map quality through pooling-based aggregation, boundary awareness, feedback-driven focusing, nested architectures, and multiscale interactions, respectively [21, 30, 34, 35, 39]. Zhang et al. further weight multi-level feature maps with Gram-matrix statistics and recursively fuse the predictions to preserve spatial details in complex scenes [46].

Although these methods perform well on natural images, they primarily rely on RGB appearance information and therefore struggle to cope consistently with underwater image degradation. The introduction of USOD10K established a large-scale benchmark for underwater salient object detection. Its data provide pixel-level saliency annotations, boundary information, and depth maps, allowing researchers to systematically evaluate salient object detection in underwater scenes [11]. Unlike existing RGB-only SOD methods, we focus on RGB-D underwater scenes and seek to use depth-structural information to compensate for degraded RGB appearance.

2.2 RGB-D Salient Object Detection and Cross-Modal Fusion

RGB-D salient object detection introduces depth maps to improve foreground-background discrimination. Early studies demonstrated that stereo depth and RGB-D data can provide effective geometric information for saliency analysis [29, 31]. Subsequent RGB-D methods began to explore more sophisticated cross-modal fusion strategies. DSA2F models RGB-D relationships through depth-sensitive attention and automatic multimodal fusion [37]; PCFNet enhances the complementarity between RGB and depth through progressive complementary-aware fusion [2]; CPFP integrates RGB-D features using contrast priors and a fluid pyramid structure [47]; and DMRA improves salient-region localization through depth-induced multiscale recurrent attention [32].

More recently, researchers have devoted increased attention to depth quality and cross-modal interactions. D3Net revisits the influence of depth quality on RGB-D SOD and proposes a depth-depurator mechanism [8]; JL-DCF promotes RGB-D representation learning through joint learning and densely cooperative fusion [9]; BBS-Net employs a bifurcated backbone strategy to strengthen hierarchical cross-modal features [42]; UC-Net addresses RGB-D saliency detection from the perspective of uncertainty modeling [43]; S2MA introduces selective self-mutual attention to improve cross-modal interactions [23]; HAINet uses hierarchical alternate interaction to model RGB-D feature relationships [20]; CalibNet mitigates potential biases in raw depth maps through depth calibration [15]; and RD3D uses 3D convolution to enhance RGB-depth interactions during encoding and decoding [3]. EATNet combines a Transformer-based dual-stream backbone with cross-modal feature enhancement and edge-guided decoding [24], whereas PENet couples asymmetric encoders with concise multimodal fusion and multiscale refinement to balance accuracy and efficiency [17]. These methods

demonstrate the importance of depth information, but most still fuse depth as a single auxiliary modality and lack explicit modeling of depth edges, local contrast, and the injection of unreliable depth. In contrast, we first structuralize a depth map into three task-relevant cues and then selectively inject trustworthy depth information through layer-wise reliability gating.

2.3 Vision Transformers, State-Space Models, and Mamba-Based Dense Prediction

Transformers have substantially strengthened the global modeling capabilities of visual models through self-attention. ViT demonstrates the effectiveness of a pure Transformer architecture for image recognition [6]; Swin Transformer adapts Transformers to dense visual tasks through a hierarchical shifted-window mechanism [27]; and PVT constructs a pyramid Vision Transformer to enhance multiscale representations [38]. However, self-attention typically incurs substantial computational overhead in high-resolution dense prediction tasks.

State-space models provide an alternative, efficient path for long-range modeling. Mamba performs linear-complexity sequence modeling through a selective state-space mechanism [10]; Vision Mamba introduces bidirectional state-space modeling into visual recognition [50]; and VMamba further constructs a visual state-space model through two-dimensional selective scanning [25]. These studies indicate the potential of the Mamba family for long-range dependency modeling and dense visual prediction. However, we argue that the key to RGB-D underwater salient object detection is not merely replacing the backbone, but combining Mamba's global modeling capability with reliable depth-structure modeling. We therefore employ VMamba for shared two-stream encoding and introduce compact Mamba context modules into both RGB-D fusion and pyramid decoding, forming a complete framework tailored to RGB-D underwater salient object detection.

3 Methodology

3.1 Overall Framework

We propose UDR-Mamba, a two-stream RGB-D underwater salient object detection framework based on a shared-weight VMamba backbone. As illustrated in Fig. 2, the model takes an RGB image and a raw depth map as input and outputs a saliency prediction map at the same spatial resolution as the original image. The network comprises four principal components: the Depth Structuralization Input Module (DSI), a two-stream shared-weight VMamba encoder, the Depth-Reliable Mamba Fusion Module (DRMF), and the Mamba Pyramid Decoder.

First, the raw depth map is fed into the DSI. Rather than using the single-channel depth map directly, the DSI transforms it into a three-channel structural representation comprising normalized depth, depth edges, and local depth contrast. This structured depth input retains the overall depth distribution while explicitly providing object-boundary and local geometric-variation information. In addition, the depth-edge map produced by the DSI is forwarded to the subsequent fusion modules and

the decoder's detail-enhancement stage to assist cross-modal fusion and boundary recovery.

The RGB image and three-channel depth-structure map are then passed separately through the VMamba encoder for feature extraction. To reduce parameter redundancy in the two-stream model, we employ a shared-weight VMamba backbone. Specifically, the RGB and depth branches perform two modality-specific forward passes but share the same VMamba encoder parameters. This design preserves the modeling capacity of the two-stream RGB-D architecture while substantially reducing the number of model parameters and training overhead. The encoder outputs four levels of multiscale features for each modality; their spatial resolutions decrease progressively, while their semantic representational capacity increases.

At each encoding stage, a DRMF fuses the RGB and depth features. Unlike direct concatenation or addition, the DRMF estimates a depth-reliability gate from the RGB features, depth features, cross-modal discrepancies, and depth-edge information. This gate determines whether the depth features at the current scale are trustworthy and selectively injects reliable depth information into the RGB representation. Meanwhile, compact Mamba context modeling within the DRMF enhances the fused features' ability to represent long-range dependencies.

Finally, the four levels of fused features are passed to the Mamba Pyramid Decoder. The decoder adopts a top-down pyramid structure that progressively upsamples and fuses features from different scales. At each decoding stage, we use Compact Mamba Context to enhance the features with global contextual information. At the end of the decoder, depth-edge information is further introduced for detail refinement, after which a single prediction head produces the final saliency map. Unlike methods that rely on multilevel auxiliary supervision, our model is trained using only the final prediction and introduces no additional depth, edge, or boundary loss, thereby keeping both the architecture and optimization objective simple.

3.2 Depth Structuralization Input Module

Although a raw depth map contains geometric information, directly feeding it into the network does not adequately represent object boundaries or local structural variations. We therefore design the Depth Structuralization Input Module (DSI), which transforms a raw depth map into a three-channel structural representation.

Given a depth map D, we first obtain its normalized depth representation:

$$D_n = \mathrm{Norm}(D) \tag{1}$$

We then use fixed Sobel operators to extract depth edges. Specifically, the depth gradients in the horizontal and vertical directions are computed as follows:

$$G_x = K_x * D_n \tag{2}$$

$$G_y = K_y * D_n \tag{3}$$

where K_x and K_y denote the Sobel-X and Sobel-Y kernels, respectively. The depth-edge map is defined as

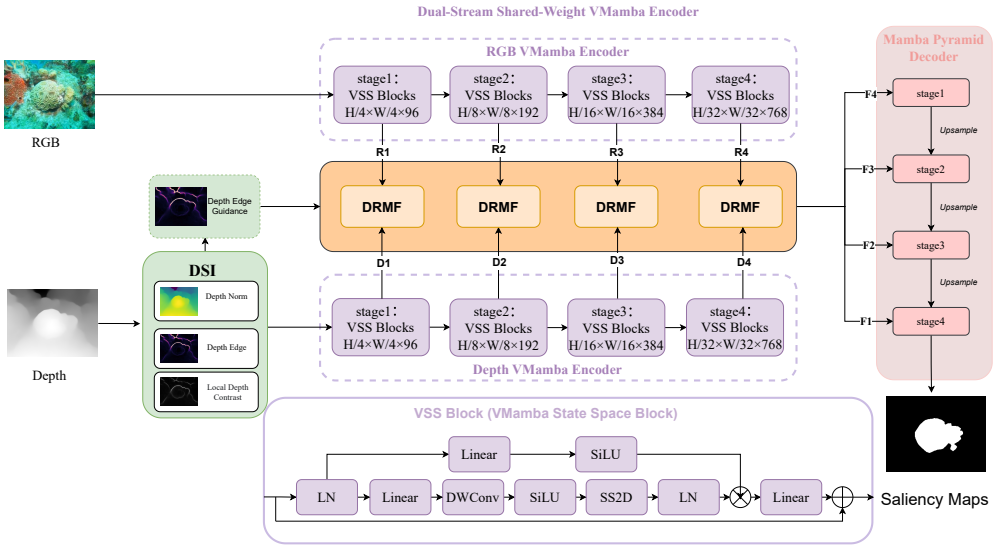


Fig. 2 Overall architecture of the proposed UDR-Mamba. The network takes an RGB image and a depth map as input. The depth map is first transformed into a structured depth representation by the DSI. The RGB image and structured depth map are then processed by the two-stream shared-weight VMamba encoder, and their features are fused at multiple stages by the DRMF modules. Finally, the Mamba Pyramid Decoder generates the saliency prediction map

$$E = \mathrm{Norm}(|G_x| + |G_y|) \tag{4}$$

This edge map explicitly characterizes regions with abrupt depth changes, which generally correspond to object contours or foreground-background boundaries.

In addition, we calculate local depth contrast to enhance local geometric differences. A 7×7 average-pooling operation first produces the local mean depth:

$$\bar{D} = \mathrm{AvgPool}_{7\times7}(D_n) \tag{5}$$

The absolute difference between the current depth and the local mean is then calculated as

$$C = \mathrm{Norm}\big(|D_n - \bar{D}|\big) \tag{6}$$

Finally, the DSI concatenates the three depth-structural cues into a three-channel depth-structure map:

$$D_s = \mathrm{Concat}(D_n, E, C) \tag{7}$$

where D_n represents the overall depth distribution, E denotes the depth edges, and C denotes local depth contrast. Equations (1)–(7) define the complete structuralization process illustrated in Fig. 3. This design enables the depth branch not only to exploit distance information but also to perceive object boundaries and local geometric variations explicitly. The depth-edge map E is additionally passed to the DRMF

modules and the decoder's detail-enhancement stage to assist cross-modal fusion and boundary recovery.

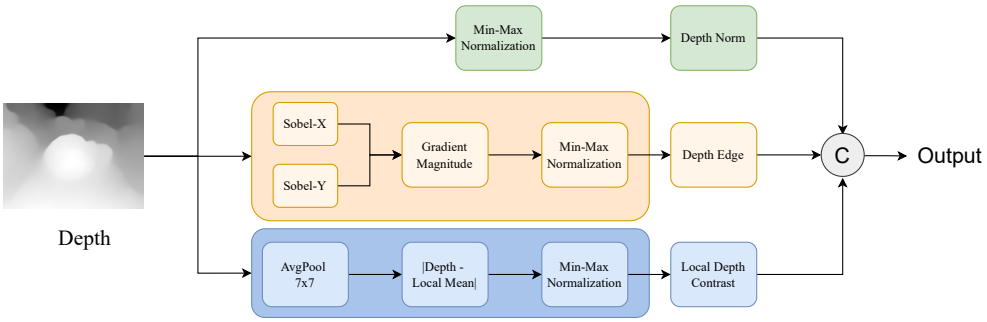


Fig. 3 Depth Structuralization Input Module. The three branches produce normalized depth, depth edges, and local depth contrast, which are concatenated into the structured depth representation

3.3 Depth-Reliable Mamba Fusion Module

Underwater depth maps may contain noise, missing regions, or inaccurate boundaries. We therefore avoid directly adding or concatenating the depth and RGB features and instead design the Depth-Reliable Mamba Fusion Module (DRMF) to estimate depth reliability at each layer and selectively inject depth information.

At the i-th encoding stage, given an RGB feature R_i, a depth feature D_i, and the depth-edge map E, we first align the channels of the depth feature:

$$D_i' = \phi_a(D_i) \tag{8}$$

where $\phi_a(\cdot)$ denotes a depth-alignment operation comprising a 1×1 convolution, GroupNorm, and GELU. We then feed the RGB feature, aligned depth feature, cross-modal discrepancy, and resized depth edge jointly into the reliability-gating branch:

$$\begin{aligned} G_i = \sigma\Big(\phi_g\big(\mathrm{Concat}\big(R_i, D_i', \\ |R_i - D_i'|, E_i\big)\big)\Big) \end{aligned} \tag{9}$$

where E_i denotes the depth-edge map resized to the current feature scale, $\phi_g(\cdot)$ is a gate-generation function comprising a 1×1 convolution, a 3×3 convolution, and another 1×1 convolution, and $\sigma(\cdot)$ is the Sigmoid function. The gating map G_i represents the reliability weights of the depth feature at the current scale.

We then use this gate to inject the depth feature selectively:

$$\widehat{D}_i = G_i \odot D_i' \tag{10}$$

where $\odot$ denotes element-wise multiplication. The RGB feature and reliable depth feature are then fused:

$$M_i = \phi_m\Big(\mathrm{Concat}(R_i, \widehat{D}_i)\Big) \tag{11}$$

where $\phi_m(\cdot)$ denotes a feature-mixing operation.

To further enhance the global-context representation of the fused feature, we introduce Compact Mamba Context into the DRMF. This module first projects the feature onto a compact state grid, performs long-range dependency modeling with VMamba/VSSBlock, and finally restores the feature to its original spatial scale:

$$\widetilde{M}_i = M_i + \mathrm{MambaContext}(M_i) \tag{12}$$

Finally, the fused feature is produced through a residual connection:

$$F_i = R_i + \phi_r(\widetilde{M}_i) \tag{13}$$

where $\phi_r(\cdot)$ denotes a two-layer convolutional refinement operation. Equations (8)–(13) summarize the DRMF computation, whose reliability-gating and feature-fusion paths are shown in Fig. 4. This residual design retains the RGB representation as a stable main branch while selectively incorporating depth-structural information through reliability gating, thereby preventing low-quality depth maps from degrading the prediction.

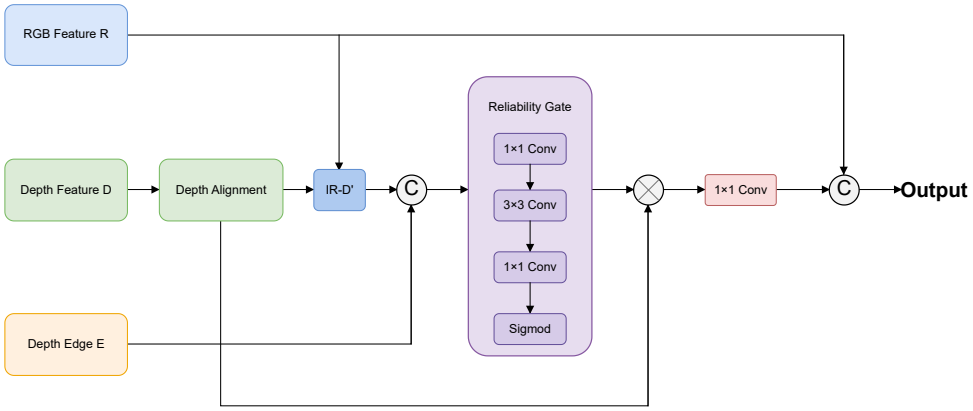


Fig. 4 Depth-Reliable Mamba Fusion Module. The gate is estimated from RGB features, aligned depth features, their discrepancy, and depth-edge cues; it modulates the aligned depth before fusion with the RGB representation

3.4 Loss Function

We supervise only the final saliency prediction. Given the model output P and binary saliency annotation Y, the total objective is the sum of the Binary Cross-Entropy (BCE) loss and the soft Intersection over Union (IoU) loss. This formulation is defined by Eqs. (14)–(16) and does not introduce auxiliary depth, edge, boundary, or multiscale prediction losses.

$$\mathcal{L} = \mathcal{L}_{\mathrm{bce}} + \mathcal{L}_{\mathrm{iou}} \tag{14}$$

The BCEWithLogits loss is defined as

$$\mathcal{L}_{\text{bce}} = -\frac{1}{N}\sum_{j=1}^{N}\Big[Y_j \log\big(\sigma(P_j)\big) + (1-Y_j)\log\big(1-\sigma(P_j)\big)\Big] \tag{15}$$

where $\sigma(\cdot)$ denotes the Sigmoid function and N is the number of valid pixels. The soft IoU loss is defined as

$$\mathcal{L}_{\text{iou}} = 1 - \frac{\sum_{j=1}^{N}\sigma(P_j)Y_j + 1}{\sum_{j=1}^{N}\left[\sigma(P_j) + Y_j - \sigma(P_j)Y_j\right] + 1} \tag{16}$$

It should be emphasized that we do not use depth supervision, edge supervision, boundary supervision, or multiscale auxiliary prediction losses. This design avoids manually tuning the weights of multiple auxiliary losses, ensuring that the performance gains arise primarily from the structured depth input, reliability-gated fusion, and Mamba context modeling.

4 Experiments

4.1 Datasets

Compared with salient object detection (SOD) in terrestrial scenes, underwater salient object detection (USOD) still suffers from relatively limited data resources, with few large-scale, high-quality public datasets available for research. The USOD dataset is an underwater saliency benchmark with pixel-level annotations [13]. It contains 300 real underwater images covering diverse natural scenes and provides pixel-level saliency annotations for evaluation.

USOD10K is the first large-scale, high-quality dataset with pixel-level annotations for USOD [11]. It contains 10,255 real underwater images spanning 12 representative underwater scenes and includes salient objects from 70 different categories. Saliency-boundary annotations and estimated depth maps are provided for every image.

USOD10K is divided into a training set of 9,229 images and a test set of 1,026 images. UDR-Mamba is trained on the USOD10K training set and evaluated on both the USOD10K test set and the USOD dataset.

4.2 Implementation Details

The experiments are implemented in Python 3.8 with PyTorch 1.13.1 and conducted on an NVIDIA GeForce RTX 3090 GPU. VMamba-Small is adopted as the encoder backbone, with the same set of VMamba backbone weights shared by the RGB and structured-depth branches. The input resolution is set to 448×448, and the batch size is 4. AdamW is used as the optimizer with an initial learning rate of 1×10^{-4}; the learning rate of the backbone is set to 0.1 times the base learning rate, and the weight decay is 5×10^{-4}. Warm-up and cosine learning-rate scheduling are employed during training, together with mixed-precision training to improve efficiency. Data augmentation comprises random horizontal flipping, random rotation, and color jittering.

Multi-scale training is performed using scale factors of 0.75, 1.0, and 1.25. For model selection, validation-set MAE is used as the primary criterion, and the checkpoint with the lowest MAE is retained as the final model for testing.

4.3 Evaluation Metrics

We quantitatively evaluate the model using metrics commonly adopted in salient object detection, including S-measure, maximum F-measure, maximum E-measure, and mean absolute error (MAE). S-measure assesses the structural similarity between a prediction map and its ground-truth annotation; F-measure jointly considers precision and recall; E-measure captures both local pixel-level matching and image-level statistics; and MAE measures the mean absolute difference between the predicted saliency map and the ground truth. Higher values indicate better performance for S-measure, F-measure, and E-measure, whereas a lower MAE indicates a more accurate prediction.

4.4 Comparison with State-of-the-Art Methods

To evaluate the effectiveness of the proposed method, we compare it with a range of existing state-of-the-art approaches on the USOD10K and USOD underwater salient object detection test sets, including PSGLoss [41], PFSNet [28], MFNet [33], CTD-Net [48], CATNet [36], SPDE [16], CE3USOD [40], Dual-SAM [45], TCUSOD [11], and HDANet [26]. The evaluation metrics are S_m, $E_\xi^{\max}$, maxF, and MAE. Higher S_m, $E_\xi^{\max}$, and maxF values indicate better performance, whereas a lower MAE indicates a smaller prediction error.

To examine the reliability of the experimental results, we further analyze the model's prediction errors across different datasets. In scenes with complex backgrounds or low contrast, the errors are concentrated primarily around object boundaries. Underwater image degradation often blurs these boundaries, making it difficult for the model to distinguish foreground objects from the background accurately.

As shown in Table 1, our method achieves the best overall performance on USOD10K. UDR-Mamba obtains the highest S_m, $E_\xi^{\max}$, and maxF scores, reaching 0.9332, 0.9713, and 0.9365, respectively. Compared with the strongest competing scores, our method improves S_m and maxF over HDANet by 0.0079 and 0.0070, respectively, and improves $E_\xi^{\max}$ over SPDE by 0.0023. For MAE, our method obtains 0.0192, only slightly higher than the 0.0190 achieved by Dual-SAM, while outperforming most other methods. These results demonstrate that the proposed depth-structuralized input and depth-reliable Mamba fusion effectively improve structural perception and salient-object localization in large-scale, challenging underwater scenes.

Our method is also highly competitive on the USOD dataset. UDR-Mamba achieves the highest $E_\xi^{\max}$ of 0.9573 and the lowest MAE of 0.0359, indicating a smaller overall discrepancy between its predicted saliency maps and the ground-truth annotations. It also attains a maxF of 0.9314, second only to HDANet's 0.9352 and higher than those of most competing methods. While Dual-SAM achieves the highest S_m and HDANet achieves the highest maxF on USOD, UDR-Mamba achieves the highest $E_\xi^{\max}$ and the lowest MAE. In addition, our method achieves the best results on

three metrics on the larger USOD10K dataset, demonstrating consistent performance across the two test sets.

The qualitative results further support these observations. As shown in Fig. 5, several competing methods tend to produce background false positives, missing object regions, or incomplete boundaries in underwater scenes characterized by low contrast, complex backgrounds, and blurred object contours. For example, some methods misclassify regions with complex underwater textures as foreground or generate discontinuous predictions for elongated or partially occluded objects. In contrast, UDR-Mamba produces saliency maps that are more complete and cleaner, with more consistent boundaries. Even when object contours are complex or RGB appearance is severely degraded, our method preserves the principal object structure more effectively. This advantage is attributable primarily to the explicit modeling of depth edges and local depth contrast by DSI and the selective injection of reliable depth information by DRMF.

Overall, both the quantitative and qualitative results show that UDR-Mamba delivers better overall performance than existing methods. Its advantages are reflected not only in the higher S_m, $E_\xi^{\max}$, and maxF scores on USOD10K, but also in the lower MAE and cleaner, more structurally consistent visual predictions on USOD. These findings validate the effectiveness of depth-structuralized representations, reliability-gated fusion, and Mamba-based long-range context modeling for RGB-D underwater salient object detection.

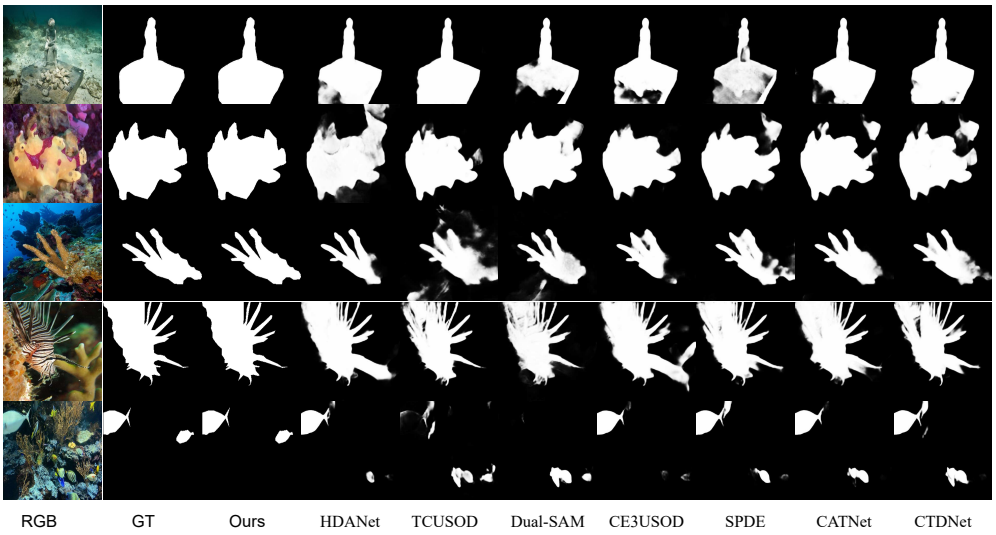


Fig. 5 Qualitative comparison on representative underwater scenes. Columns show the RGB image, ground truth, UDR-Mamba, and predictions from HDANet, TCUSOD, Dual-SAM, CE3USOD, SPDE, CATNet, and CTDNet

Table 1 Quantitative comparison with existing state-of-the-art methods on USOD10K and USOD.

Methods	Venue	USOD10K				USOD			
		$S_m \uparrow$	$E_\xi^{\max} \uparrow$	maxF↑	MAE↓	$S_m \uparrow$	$E_\xi^{\max} \uparrow$	maxF↑	MAE↓
PSGLoss	TIP	0.8640	0.9078	0.8508	0.0417	0.8911	0.9366	0.9108	0.0416
PFSNet	AAAI	0.8983	0.9421	0.8966	0.0370	0.9031	0.9387	0.9160	0.0443
MFNet	ICCV	0.8425	0.9146	0.8193	0.0512	0.8465	0.9103	0.8485	0.0722
CTDNet	ACM MM	0.9085	0.9531	0.9073	0.0285	0.9041	0.9397	0.9112	0.0406
CATNet	TMM	0.8898	0.9492	0.8876	0.0300	0.8868	0.9377	0.9026	0.0446
SPDE	TCSVT	0.9217	<u>0.9690</u>	0.9259	0.0200	0.9118	0.9497	0.9249	0.0362
CE3USOD	JOE	0.9020	0.9460	0.8712	0.0280	0.8821	0.9215	0.9100	0.0520
Dual-SAM	CVPR	0.9225	0.9684	0.9210	**0.0190**	**0.9334**	0.9527	0.9226	<u>0.0360</u>
TCUSOD	TIP	0.9217	0.9685	0.9238	0.0200	0.9019	0.9445	0.9136	0.0397
HDANet	TGRS	<u>0.9253</u>	0.9685	<u>0.9295</u>	0.0201	0.9223	<u>0.9548</u>	**0.9352**	0.0363
UDR-Mamba	Ours	**0.9332**	**0.9713**	**0.9365**	<u>0.0192</u>	<u>0.9273</u>	**0.9573**	<u>0.9314</u>	**0.0359**

4.5 Ablation Studies

To verify the effectiveness of the key components in UDR-Mamba, we conduct a series of ablation studies on USOD10K. Unless otherwise specified, all variants use identical training and testing settings. We analyze four aspects: the effectiveness of DSI, the contributions of its individual structural branches, the effectiveness of the DRMF fusion strategy, and the accuracy–efficiency trade-off under different backbone configurations.

We first compare the proposed Depth Structuralization Input Module (DSI) with the direct use of raw depth maps. For the Raw Depth variant, the normalized single-channel depth map is replicated as $[D_n, D_n, D_n]$ and fed directly into the depth branch. In contrast, DSI constructs the three-channel structural representation $[D_n, E, C]$, where D_n, E, and C denote normalized depth, depth edges, and local depth contrast, respectively.

As shown in Table 2, DSI consistently improves all four evaluation metrics. Compared with Raw Depth, DSI increases S_m from 0.9236 to 0.9332, $E_\xi^{\max}$ from 0.9618 to 0.9713, and maxF from 0.9274 to 0.9365, while reducing MAE from 0.0228 to 0.0192. These results indicate that direct use of the raw depth map does not fully exploit the structural value of depth information, whereas explicitly incorporating depth edges and local depth contrast provides more discriminative structural cues for underwater salient object detection.

Table 2 Ablation study on the effectiveness of the proposed DSI module

Method	Depth Input	USOD10K			
		$S_m \uparrow$	$E_\xi^{\max} \uparrow$	maxF↑	MAE↓
Raw Depth	$[D_n, D_n, D_n]$	0.9236	0.9618	0.9274	0.0228
DSI	$[D_n, E, C]$	**0.9332**	**0.9713**	**0.9365**	**0.0192**

To further analyze the role of each structural branch in DSI, we independently remove the depth-edge branch and the local-depth-contrast branch. The results are reported in Table 3. Removing either branch degrades performance. Without depth edges, S_m, $E_\xi^{\max}$, and maxF decrease to 0.9278, 0.9662, and 0.9312, respectively, while MAE increases to 0.0214. This finding indicates that depth edges are important for localizing object contours and perceiving foreground–background boundaries.

Removing local depth contrast also degrades performance, yielding S_m, $E_\xi^{\max}$, and maxF scores of 0.9291, 0.9674, and 0.9320, respectively, and an MAE of 0.0208. This result indicates that local depth contrast strengthens local geometric differences between foreground and background, helping the model distinguish salient-object regions more accurately. The complete DSI achieves the best result on every metric, confirming the complementarity of normalized depth, depth edges, and local depth contrast.

Table 3 Ablation study on the contributions of the structural branches in DSI

Method	Depth	Edge	Contrast	USOD10K			
				$S_m \uparrow$	$E_\xi^{\max} \uparrow$	maxF↑	MAE↓
w/o Edge	✓	×	✓	0.9278	0.9662	0.9312	0.0214
w/o Contrast	✓	✓	×	0.9291	0.9674	0.9320	0.0208
Full DSI	✓	✓	✓	**0.9332**	**0.9713**	**0.9365**	**0.0192**

To evaluate the effectiveness of the Depth-Reliable Mamba Fusion Module (DRMF), we compare it with three alternative fusion strategies in Table 4: direct addition, concatenation followed by convolution, and Mamba-based fusion without reliability gating. Add Fusion performs the worst, obtaining only $S_m = 0.9217$, $E_\xi^{\max} = 0.9596$, maxF=0.9259, and MAE=0.0236. This finding suggests that directly adding RGB and depth features can introduce unreliable depth noise and consequently impair the predictions.

Concat Fusion improves upon Add Fusion, indicating that learnable feature fusion is preferable to fixed element-wise addition. Introducing Mamba-based contextual modeling while removing the reliability gate yields a further improvement, showing that long-range context modeling facilitates cross-modal feature interaction. Nevertheless, the complete DRMF achieves the best results, with $S_m = 0.9332$, $E_\xi^{\max} = 0.9713$, maxF=0.9365, and MAE=0.0192. Compared with the w/o Gate variant, DRMF improves S_m by 0.0064 and reduces MAE by 0.0015. These results show that reliability gating and Mamba-based context modeling play complementary roles: the former suppresses unreliable depth responses, whereas the latter strengthens long-range cross-modal interactions.

Finally, Table 5 analyzes how backbone capacity and weight sharing affect model performance and efficiency. The RGB Baseline records the lowest accuracy. The RGB-D VMamba-Tiny configuration yields higher scores, increasing S_m from 0.9142 to 0.9245 and reducing MAE from 0.0268 to 0.0224; however, because this comparison

Table 4 Ablation study of different RGB-D fusion strategies

Method	Fusion Strategy	USOD10K			
		$S_m \uparrow$	$E_\xi^{\max} \uparrow$	maxF↑	MAE↓
Add Fusion	$R + D$	0.9217	0.9596	0.9259	0.0236
Concat Fusion	Conv($[R, D]$)	0.9251	0.9641	0.9290	0.0221
w/o Gate	Conv($[R, D]$) + Mamba	0.9268	0.9674	0.9328	0.0207
DRMF	Gate + Mamba	**0.9332**	**0.9713**	**0.9365**	**0.0192**

changes both the input modalities and backbone capacity, it should not be interpreted as an isolated estimate of the depth contribution.

VMamba-Small provides stronger feature representations than VMamba-Tiny and therefore achieves higher accuracy. The independent dual-backbone VMamba-Small configuration is marginally more accurate than its shared-weight counterpart, obtaining S_m, $E_\xi^{\max}$, maxF, and MAE values of 0.9340, 0.9725, 0.9371, and 0.0190, respectively. However, this configuration requires 133.76M parameters and 173.76G FLOPs, resulting in substantially greater computational and storage costs. In contrast, shared-weight VMamba-Small uses only 83.85M parameters and 160.34G FLOPs while achieving very similar performance, with S_m, $E_\xi^{\max}$, maxF, and MAE values of 0.9332, 0.9713, 0.9365, and 0.0192, respectively. Thus, weight sharing reduces the parameter count by approximately 49.91M with virtually no loss in detection accuracy. This result indicates that a shared VMamba backbone substantially reduces model redundancy while retaining the representational capacity of the RGB-D dual-stream architecture, thereby providing a more favorable accuracy–efficiency trade-off.

Overall, the ablation studies provide strong evidence for the effectiveness of each core component. DSI supplies more discriminative structural cues than raw depth input; DRMF enables reliable, adaptive RGB-D fusion; and the shared VMamba backbone substantially reduces the parameter count with almost no loss in accuracy. Together, these results demonstrate the importance of the proposed depth-structuralization scheme, reliability-gated fusion, and shared Mamba encoding strategy for RGB-D underwater salient object detection.

Table 5 Backbone and efficiency analysis on USOD10K. The RGB-D and Shared columns indicate the use of depth input and encoder weight sharing, respectively

Method	Encoder	RGB-D	Shared	USOD10K					
				Params (M)↓	FLOPs (G)↓	$S_m \uparrow$	$E_\xi^{\max} \uparrow$	maxF↑	MAE↓
RGB Baseline	VMamba-S	×	×	56.47	80.77	0.9142	0.9483	0.9185	0.0268
RGB-D Tiny	VMamba-T	✓	✓	65.32	115.28	0.9245	0.9625	0.9286	0.0224
RGB-D Small	VMamba-S	✓	×	133.76	173.76	0.9340	0.9725	0.9371	0.0190
RGB-D Small	VMamba-S	✓	✓	83.85	160.34	0.9332	0.9713	0.9365	0.0192

5 Conclusion

This paper presents UDR-Mamba, a depth-structuralization and reliability-guided Mamba network for RGB-D underwater salient object detection. To address color degradation, low contrast, complex backgrounds, and blurred object boundaries in underwater imagery, we improve both depth-structure modeling and reliable cross-modal fusion. First, the proposed Depth Structuralization Input Module converts a raw depth map into a structured representation comprising normalized depth, depth edges, and local depth contrast, thereby exploiting geometric and boundary cues in the depth map more fully. Second, the proposed Depth-Reliable Mamba Fusion Module adaptively assesses the reliability of depth features through reliability gating and combines this mechanism with Mamba-based contextual modeling to fuse RGB and depth features effectively. In addition, a shared-weight VMamba backbone and a single-head Mamba pyramid decoder reduce model redundancy while preserving dual-stream RGB-D modeling capacity and avoiding complex auxiliary-loss designs.

Experiments on the USOD10K and USOD datasets demonstrate that the proposed method is highly competitive with existing state-of-the-art approaches. UDR-Mamba achieves the best S_m, $E_\xi^{\max}$, and maxF scores on USOD10K. On USOD, it achieves the highest $E_\xi^{\max}$, the lowest MAE, and the second-highest maxF. The ablation studies further validate the effectiveness of DSI, DRMF, and the shared VMamba backbone. The qualitative results also show that the proposed method produces more complete and structurally consistent saliency maps in underwater scenes with complex backgrounds, low contrast, and blurred boundaries.

Future work will investigate more lightweight Mamba architectures to facilitate deployment on real-time platforms such as underwater robots. We also plan to develop stronger depth-quality assessment mechanisms to improve robustness to low-quality or missing depth information.

Statements and Declarations

Funding

No funds, grants, or other support was received.

Competing interests

The authors have no relevant financial or non-financial interests to disclose.

Ethics approval

Not applicable.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Data availability

The experiments use the publicly available USOD10K and USOD datasets described in Refs. [11, 13]. The USOD10K data and benchmark resources are available at https://github.com/LinHong-HIT/USOD10K, and the USOD dataset is available at https://irvlab.cs.umn.edu/resources/usod-dataset. No new benchmark dataset was created for this study.

Materials availability

Not applicable.

Code availability

The source code developed for this study is not publicly available.

Author contributions

Ning Li and Jiaxin Li made the principal contributions to the conceptualization of the study, development of the proposed solutions, design of the experiments, organization of the results, and writing of the manuscript. All authors read and approved the final manuscript.

Declaration of generative AI and AI-assisted technologies

During the preparation of this work, the authors used OpenAI Codex to assist with English translation, language and readability improvement, manuscript organization, figure-format conversion, and LaTeX formatting. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

[1] Achanta R, Hemami S, Estrada F, et al (2009) Frequency-tuned salient region detection. In: Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 1597–1604, https://doi.org/10.1109/CVPR.2009.5206596

[2] Chen H, Li Y (2018) Progressively complementarity-aware fusion network for RGB-D salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 3051–3060, https://doi.org/10.1109/CVPR.2018.00322

[3] Chen Q, Liu Z, Zhang Y, et al (2021) RGB-D salient object detection via 3D convolutional neural networks. Proceedings of the AAAI Conference on Artificial Intelligence 35(2):1063–1071. https://doi.org/10.1609/aaai.v35i2.16191

[4] Cheng MM, Zhang GX, Mitra NJ, et al (2011) Global contrast based salient region detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 409–416, https://doi.org/10.1109/CVPR.2011.5995344

[5] Deng Z, Hu X, Zhu L, et al (2018) R^3Net: Recurrent residual refinement network for saliency detection. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI), pp 684–690, https://doi.org/10.24963/ijcai.2018/95

[6] Dosovitskiy A, Beyer L, Kolesnikov A, et al (2021) An image is worth 16×16 words: Transformers for image recognition at scale. In: Proceedings of the International Conference on Learning Representations (ICLR), URL https://openreview.net/forum?id=YicbFdNTTy

[7] Du D, Pan L, Liang Y, et al (2025) Underwater image restoration based on a dual-branch super-resolution residual network. Multimedia Systems 31:286. https://doi.org/10.1007/s00530-025-01861-y

[8] Fan DP, Lin Z, Zhang Z, et al (2021) Rethinking RGB-D salient object detection: Models, data sets, and large-scale benchmarks. IEEE Transactions on Neural Networks and Learning Systems 32(5):2075–2089. https://doi.org/10.1109/TNNLS.2020.2996406

[9] Fu K, Fan DP, Ji GP, et al (2020) JL-DCF: Joint learning and densely-cooperative fusion framework for RGB-D salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 3049–3059, https://doi.org/10.1109/CVPR42600.2020.00312

[10] Gu A, Dao T (2024) Mamba: Linear-time sequence modeling with selective state spaces. In: Proceedings of the First Conference on Language Modeling (COLM), URL https://openreview.net/forum?id=tEYskw1VY2

[11] Hong L, Wang X, Zhang G, et al (2025) USOD10K: A new benchmark dataset for underwater salient object detection. IEEE Transactions on Image Processing 34:1602–1615. https://doi.org/10.1109/TIP.2023.3266163

[12] Hou Q, Cheng MM, Hu X, et al (2017) Deeply supervised salient object detection with short connections. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 5300–5309, https://doi.org/10.1109/CVPR.2017.563

[13] Islam MJ, Wang R, Sattar J (2022) SVAM: Saliency-guided visual attention modeling by autonomous underwater robots. In: Robotics: Science and Systems XVIII, https://doi.org/10.15607/RSS.2022.XVIII.048

[14] Itti L, Koch C, Niebur E (1998) A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(11):1254–1259. https://doi.org/10.1109/34.730558

[15] Ji W, Li J, Yu S, et al (2021) Calibrated RGB-D salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 9466–9476, https://doi.org/10.1109/CVPR46437.2021.00935

[16] Jin J, Jiang Q, Wu Q, et al (2025) Underwater salient object detection via dual-stage self-paced learning and depth emphasis. IEEE Transactions on Circuits and Systems for Video Technology 35(3):2147–2160. https://doi.org/10.1109/TCSVT.2024.3491907

[17] Kou C, Han J, Wang M (2026) Rethinking RGB-D salient object detection. Multimedia Systems 32:113. https://doi.org/10.1007/s00530-025-02163-z

[18] Li C, Guo C, Ren W, et al (2020) An underwater image enhancement benchmark dataset and beyond. IEEE Transactions on Image Processing 29:4376–4389. https://doi.org/10.1109/TIP.2019.2955241

[19] Li G, Yu Y (2015) Visual saliency based on multiscale deep features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 5455–5463, https://doi.org/10.1109/CVPR.2015.7299184

[20] Li G, Liu Z, Chen M, et al (2021) Hierarchical alternate interaction network for RGB-D salient object detection. IEEE Transactions on Image Processing 30:3528–3542. https://doi.org/10.1109/TIP.2021.3062689

[21] Liu JJ, Hou Q, Cheng MM, et al (2019) A simple pooling-based design for real-time salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 3912–3921, https://doi.org/10.1109/CVPR.2019.00404

[22] Liu N, Han J, Yang MH (2018) PiCANet: Learning pixel-wise contextual attention for saliency detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 3089–3098, https://doi.org/10.1109/CVPR.2018.00326

[23] Liu N, Zhang N, Han J (2020) Learning selective self-mutual attention for RGB-D saliency detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 13753–13762, https://doi.org/10.1109/CVPR42600.2020.01377

[24] Liu X, Liu C, Zhou X, et al (2025) EATNet: Edge-aware and transformer-based network for RGB-D salient object detection. Multimedia Systems 31:128. https://doi.org/10.1007/s00530-025-01717-5

[25] Liu Y, Tian Y, Zhao Y, et al (2024) VMamba: Visual state space model. In: Advances in Neural Information Processing Systems, pp 103031–103063, https://doi.org/10.52202/079017-3273

[26] Liu Y, Zhang X, Zhu J, et al (2025) HDANet: Enhancing underwater salient object detection with physics-inspired multimodal joint learning. IEEE Transactions on Geoscience and Remote Sensing 63:1–14. https://doi.org/10.1109/TGRS.2025.3565579

[27] Liu Z, Lin Y, Cao Y, et al (2021) Swin Transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp 9992–10002, https://doi.org/10.1109/ICCV48922.2021.00986

[28] Ma M, Xia C, Li J (2021) Pyramidal feature shrinking for salient object detection. Proceedings of the AAAI Conference on Artificial Intelligence 35(3):2311–2318. https://doi.org/10.1609/aaai.v35i3.16331

[29] Niu Y, Geng Y, Li X, et al (2012) Leveraging stereopsis for saliency analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 454–461, https://doi.org/10.1109/CVPR.2012.6247708

[30] Pang Y, Zhao X, Zhang L, et al (2020) Multi-scale interactive network for salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 9410–9419, https://doi.org/10.1109/CVPR42600.2020.00943

[31] Peng H, Li B, Xiong W, et al (2014) RGBD salient object detection: A benchmark and algorithms. In: Fleet D, Pajdla T, Schiele B, et al (eds) Computer Vision – ECCV 2014, Lecture Notes in Computer Science, vol 8691. Springer, Cham, p 92–109, https://doi.org/10.1007/978-3-319-10578-9_7

[32] Piao Y, Ji W, Li J, et al (2019) Depth-induced multi-scale recurrent attention network for saliency detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp 7253–7262, https://doi.org/10.1109/ICCV.2019.00735

[33] Piao Y, Wang J, Zhang M, et al (2021) MFNet: Multi-filter directive network for weakly supervised salient object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp 4116–4125, https://doi.org/10.1109/ICCV48922.2021.00410

[34] Qin X, Zhang Z, Huang C, et al (2019) BASNet: Boundary-aware salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 7471–7481, https://doi.org/10.1109/CVPR.2019.00766

[35] Qin X, Zhang Z, Huang C, et al (2020) U^2-Net: Going deeper with nested U-structure for salient object detection. Pattern Recognition 106:107404. https://doi.org/10.1016/j.patcog.2020.107404

[36] Sun F, Ren P, Yin B, et al (2024) CATNet: A cascaded and aggregated transformer network for RGB-D salient object detection. IEEE Transactions on Multimedia 26:2249–2262. https://doi.org/10.1109/TMM.2023.3294003

[37] Sun P, Zhang W, Wang H, et al (2021) Deep RGB-D saliency detection with depth-sensitive attention and automatic multi-modal fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 1407–1417, https://doi.org/10.1109/CVPR46437.2021.00146

[38] Wang W, Xie E, Li X, et al (2021) Pyramid Vision Transformer: A versatile backbone for dense prediction without convolutions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp 548–558, https://doi.org/10.1109/ICCV48922.2021.00061

[39] Wei J, Wang S, Huang Q (2020) F^3Net: Fusion, feedback and focus for salient object detection. Proceedings of the AAAI Conference on Artificial Intelligence 34(7):12321–12328. https://doi.org/10.1609/aaai.v34i07.6916

[40] Wu Q, Xie J, Fu Z, et al (2025) CE^3USOD: Channel-enhanced, efficient, and effective network for underwater salient object detection. IEEE Journal of Oceanic Engineering 50(2):941–954. https://doi.org/10.1109/JOE.2024.3523356

[41] Yang S, Lin W, Lin G, et al (2021) Progressive self-guided loss for salient object detection. IEEE Transactions on Image Processing 30:8426–8438. https://doi.org/10.1109/TIP.2021.3113794

[42] Zhai Y, Fan DP, Yang J, et al (2021) Bifurcated backbone strategy for RGB-D salient object detection. IEEE Transactions on Image Processing 30:8727–8742. https://doi.org/10.1109/TIP.2021.3116793

[43] Zhang J, Fan DP, Dai Y, et al (2020) UC-Net: Uncertainty inspired RGB-D saliency detection via conditional variational autoencoders. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 8579–8588, https://doi.org/10.1109/CVPR42600.2020.00861

[44] Zhang P, Wang D, Lu H, et al (2017) Amulet: Aggregating multi-level convolutional features for salient object detection. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp 202–211, https://doi.

org/10.1109/ICCV.2017.31

[45] Zhang P, Yan T, Liu Y, et al (2024) Fantastic animals and where to find them: Segment any marine animal with Dual SAM. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 2578–2587, https://doi.org/10.1109/CVPR52733.2024.00249

[46] Zhang S, Chen Y, Meng Y, et al (2023) A multi-level feature weight fusion model for salient object detection. Multimedia Systems 29:887–895. https://doi.org/10.1007/s00530-022-01018-1

[47] Zhao JX, Cao Y, Fan DP, et al (2019) Contrast prior and fluid pyramid integration for RGBD salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 3922–3931, https://doi.org/10.1109/CVPR.2019.00405

[48] Zhao Z, Xia C, Xie C, et al (2021) Complementary trilateral decoder for fast and accurate salient object detection. In: Proceedings of the 29th ACM International Conference on Multimedia, pp 4967–4975, https://doi.org/10.1145/3474085.3475494

[49] Zheng Z, Yu W (2025) RG-YOLO: Multi-scale feature learning for underwater target detection. Multimedia Systems 31:26. https://doi.org/10.1007/s00530-024-01617-0

[50] Zhu L, Liao B, Zhang Q, et al (2024) Vision Mamba: Efficient visual representation learning with bidirectional state space model. In: Proceedings of the 41st International Conference on Machine Learning, Proceedings of Machine Learning Research, vol 235. PMLR, pp 62429–62442, URL https://proceedings.mlr.press/v235/zhu24f.html