ULSNet: Underwater-Aware Preference–Value Attention with a Boundary-Refined Head for Underwater Instance Segmentation

Zhi-Dong Liang, De Li, Jin-Chun Piao*
Department of Computer Science and Technology
Yanbian University, Yanji, China
*jcpiao@ybu.edu.cn

***Abstract*—Underwater instance segmentation is critical for marine robotics and ecological monitoring. However, it remains challenging due to color cast, low contrast, scattering haze, and sensor noise, as well as frequent occlusions, ambiguous boundaries, and severe instance merging. To address the trade-off between accuracy and efficiency in existing methods, we propose ULSNet, an efficient two-stage underwater instance segmentation framework built upon an LSNet+FPN feature extractor. ULSNet introduces three lightweight, task-oriented modules to strengthen RoI representations and mask decoding under underwater degradations. Specifically, Underwater-Aware Feature Gating (UWAF) recalibrates RoI features via complementary spatial and channel gating guided by hetero-scale pyramid cues. Preference–Value Attention (PVA) further enhances instance evidence by modulating a shared value branch with content-adaptive preference cues from two lightweight heads. Finally, the UQBR-Head performs boundary-refined mask decoding with an explicit boundary branch to mitigate boundary uncertainty and instance merging. Extensive experiments on UIIS and USIS10K show that ULSNet improves the accuracy–efficiency trade-off, with consistent gains in mAP and AP_{75} under challenging underwater conditions.**

***Index Terms*—underwater instance segmentation, lightweight models, RoI feature gating, boundary-aware mask decoding**

I. Introduction

Underwater instance segmentation is fundamental for marine robotics, ecological monitoring, and underwater inspection, requiring instance-level localization and pixel-accurate masks for counting, interaction analysis, and precision operations. Compared with terrestrial scenes, it is significantly more challenging due to color cast, low contrast, scattering-induced haze, sensor noise, and the prevalence of densely distributed small or slender objects, frequent occlusions, ambiguous boundaries, and severe instance merging [1]. Existing methods exhibit a clear trade-off: large foundation models or heavyweight backbones are costly to deploy on resource-constrained platforms and often unstable under domain shift, while conventional lightweight designs suffer from limited receptive field and rigid local aggregation, leading to missed small instances and degraded mask quality at high IoU thresholds [2], [3]. This phenomenon is illustrated in Fig. 1, where methods lie along an accuracy–parameter frontier, whereas our ULSNet variants move consistently toward the upper-left region.

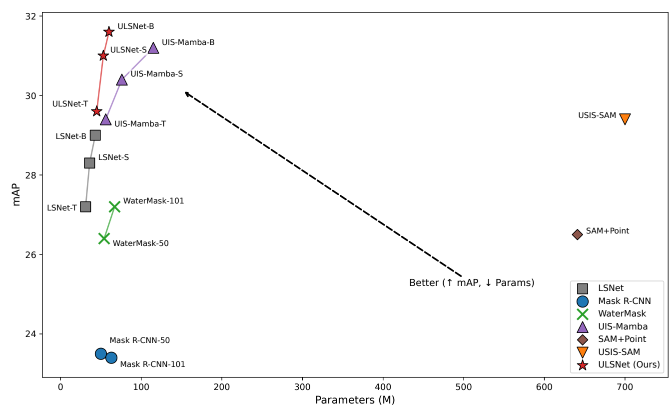

Fig. 1. Accuracy–Efficiency Trade-off on the UIIS dataset. Moving toward the upper-left (higher mAP and fewer parameters) indicates a better trade-off.

To achieve a better balance between accuracy and efficiency, we adopt LSNet [4] as the backbone. Its "See Large, Focus Small" design combines large-range perception with small-range adaptive aggregation, which is well suited for suppressing scattering-induced clutter while preserving fine structures. LSNet has also shown strong transferability to detection and instance segmentation tasks, making it a reliable foundation for underwater deployment. However, LSNet is domain-agnostic: degradation-induced spurious responses still contaminate multi-level and RoI features, and standard mask heads are insufficient for resolving boundary uncertainty and instance merging in turbid scenes. We therefore propose ULSNet, which preserves LSNet's efficiency while introducing underwater-oriented modules to recalibrate RoI features, enhance instance-relevant evidence selection, and progressively refine boundaries. The key contributions of this work are summarized as follows:

- We introduce an underwater-aware RoI recalibration module to suppress degradation-amplified noise while preserving useful context.
- We propose a preference–value attention with a shared value branch and two lightweight preference heads for robust evidence selection.
- We design a boundary-refined mask head to enhance boundary delineation and mitigate instance merging.

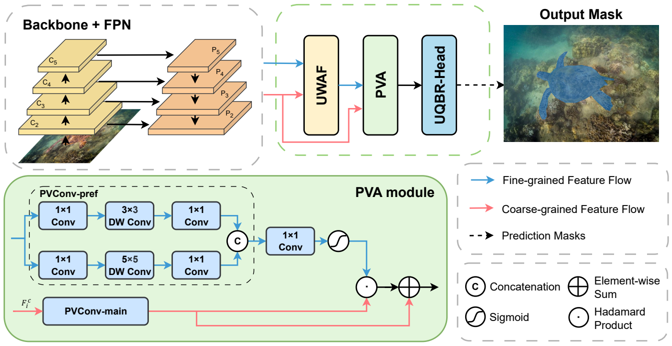

Fig. 2. Framework of ULSNet. Given RoI features extracted by an LSNet+FPN backbone, UWAF and PVA progressively refine instance evidence, and the UQBR-Head decodes boundary-aware masks for underwater instance segmentation.

II. Related Work

Underwater instance segmentation (UIS) targets instance-level masks with category distinction, while underwater salient instance segmentation (USIS) focuses on visually prominent instances, typically evaluated on UIIS and USIS10K [3]. Compared with terrestrial scenarios, UIS/USIS is more challenging due to attenuation- and scattering-induced degradations (e.g., color distortion, low contrast, and blurred contours) and complex structures such as occlusions and severe instance merging.

Existing methods mainly follow several representative directions. WaterMask adapts two-stage pipelines to underwater imagery by refining the head design, whereas USIS-SAM transfers large foundation models with underwater-adaptive components and prompting but is heavyweight for deployment and can be brittle in crowded scenes. UIS-Mamba explores state space models with dynamic scanning and hidden-state suppression to improve robustness under degradations with relatively low cost [5]; however, it mainly enhances backbone representation and does not explicitly address RoI-stage degradation interference or boundary-sensitive mask decoding for ambiguity and poor instance separability.

These gaps suggest that, beyond backbone robustness, UIS/USIS also requires RoI-stage degradation suppression and boundary-aware decoding under an efficient budget. Accordingly, we choose an efficient backbone that retains fine structures while maintaining large-range context. We adopt LSNet as the feature extractor due to its "See Large, Focus Small" LS convolution, which offers a favorable efficiency–accuracy trade-off for dense prediction [4]. Nevertheless, LSNet remains domain-agnostic and does not explicitly address RoI-stage degradation interference or boundary ambiguity, motivating task-oriented RoI evidence selection and boundary-oriented refinement modules on top of an efficient feature extractor.

III. Method

As illustrated in Fig. 2, ULSNet builds an efficient underwater instance segmentation pipeline upon an LSNet+FPN backbone. UWAF recalibrates RoI features, PVA enhances instance-relevant evidence, and UQBR-Head refines mask boundaries under underwater degradations.

A. Preliminaries

We follow a standard two-stage instance segmentation framework and keep the detector unchanged, focusing on enhancing RoI features and mask decoding under underwater degradations.

LSNet backbone. We adopt LSNet as the feature extractor for its efficient large-range perception and small-range aggregation. Given a token $\boldsymbol{x}_i$, LS convolution can be summarized as

$$\boldsymbol{y}_i = A\big(P(\boldsymbol{x}_i, N_P), N_A\big), \tag{1}$$

where $\mathcal{N}_P$ and $\mathcal{N}_A$ denote the perception and aggregation regions with $\mathcal{N}_P > \mathcal{N}_A$. In practice, LSNet uses large-kernel depth-wise operations for global context perception and lightweight content-adaptive aggregation for detail preservation.

FPN and RoI features. Multi-level features from LSNet are fed into FPN to produce $\{\boldsymbol{P}_2, \boldsymbol{P}_3, \boldsymbol{P}_4, \boldsymbol{P}_5\}$. For each RoI r_i, RoIAlign extracts $\boldsymbol{F}_i \in \mathbb{R}^{C\times H\times W}$. To better handle underwater degradations, we maintain two RoI feature flows: a fine-grained feature $\boldsymbol{F}_i^f$ for detail preservation and a coarse-grained feature $\boldsymbol{F}_i^c$ for context stability.

B. Underwater-Aware Feature Gating

Although LSNet is efficient, it is domain-agnostic and RoI features extracted from underwater scenes often contain haze-induced clutter and color-distortion responses. To adapt RoI representations for underwater instance segmentation, we introduce a lightweight *Underwater-Aware Feature Gating* module (UWAF) after RoIAlign, which performs hetero-scale recalibration using shallow features for spatial guidance and deep features for channel guidance.

For an RoI r_i, we extract multi-level RoI features $\boldsymbol{F}_i^{(l)} = \mathrm{RoIAlign}(\boldsymbol{P}_l, r_i)$, and use the baseline RoI feature $\boldsymbol{F}_i$ to compute two complementary gates.

Spatial gate. Shallow RoI features $(\boldsymbol{F}_i^{(2)}, \boldsymbol{F}_i^{(3)})$ provide fine-grained spatial cues for suppressing haze-induced noise:

$$\boldsymbol{W}_i^s = \sigma\Big(\mathrm{Conv}_{1\times1}\Big(\phi\big(\mathrm{GN}(\mathrm{DWConv}_{3\times3}([\boldsymbol{F}_i^{(2)}; \boldsymbol{F}_i^{(3)}]))\big)\Big)\Big). \tag{2}$$

Channel gate. Deeper RoI features $(\boldsymbol{F}_i^{(4)}, \boldsymbol{F}_i^{(5)})$ offer more stable semantic context for channel selection:

$$\boldsymbol{z}_i = \mathrm{AvgPool}\Big(\mathrm{Conv}_{1\times1}([\boldsymbol{F}_i^{(4)}; \boldsymbol{F}_i^{(5)}])\Big), \tag{3}$$

$$\boldsymbol{W}_i^c = \sigma(\mathrm{MLP}(\boldsymbol{z}_i)). \tag{4}$$

The spatial and channel gates are applied to the baseline RoI feature $\boldsymbol{F}_i$ via element-wise modulation with a residual formulation to obtain the recalibrated feature $\boldsymbol{F}_i^{\mathrm{uwaf}}$.

UWAF suppresses spatial clutter using shallow cues and enhances semantic discrimination using deep cues, while remaining lightweight and fully compatible with the LSNet pipeline.

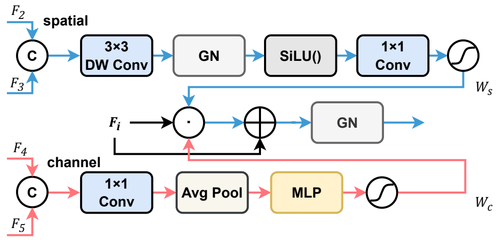

Fig. 3. Architecture of the proposed UWAF module. Shallow RoI features guide spatial gating, while deep RoI features guide channel gating for underwater-aware RoI recalibration.

C. Preference–Value Attention

UWAF suppresses degradation-amplified clutter in RoI features, yet underwater RoIs still vary significantly with visibility and background interference, where residual haze textures may be retained and thin structures can be weakened. Inspired by PVConv [6], we propose a lightweight *Preference–Value Attention* (PVA) module to further enhance instance-relevant evidence by fusing two complementary RoI flows.

Let $\boldsymbol{F}_i^f = \boldsymbol{F}_i^{\text{uwaf}}$ denote the fine-grained RoI feature and $\boldsymbol{F}_i^c$ denote the coarse-grained RoI feature. In practice, $\boldsymbol{F}_i^c$ is built by fusing deeper RoI features $\{\boldsymbol{F}_i^{(4)}, \boldsymbol{F}_i^{(5)}\}$, providing a more stable and context-rich reference. PVA first computes a shared value tensor via a lightweight projection (PVConv-main):

$$\boldsymbol{V}_i = \phi_v(\boldsymbol{F}_i^c), \quad \boldsymbol{V}_i \in \mathbb{R}^{C\times H\times W}, \tag{5}$$

where $\phi_v(\cdot)$ is implemented with efficient point-wise operations to preserve the computational profile of LSNet-based pipelines.

Next, PVA computes two preference maps (PVConv-pref) from the fine-grained feature. Each preference head adopts a depth-wise spatial mixer with a different kernel size to capture complementary local patterns:

$$\boldsymbol{Q}_i^{(k)} = \phi_p^{(k)}(\boldsymbol{F}_i^f), \quad k \in \{1, 2\}, \tag{6}$$

where $\phi_p^{(k)}(\cdot)$ consists of a 1×1 convolution, an $s_k\times s_k$ depth-wise convolution, and another 1×1 convolution, with $(s_1, s_2) = (3, 5)$. The two preferences are concatenated and converted into a gating map:

$$\boldsymbol{G}_i = \sigma\Big(\phi_g\big([\boldsymbol{Q}_i^{(1)}; \boldsymbol{Q}_i^{(2)}]\big)\Big), \quad \boldsymbol{G}_i \in \mathbb{R}^{C\times H\times W}, \tag{7}$$

where $\phi_g(\cdot)$ is a 1×1 convolution and $\sigma(\cdot)$ is the sigmoid function.

Finally, PVA modulates the shared value with the learned gate and applies a residual fusion:

$$\boldsymbol{F}_i^{\text{pva}} = \boldsymbol{V}_i + \boldsymbol{V}_i \odot \boldsymbol{G}_i, \tag{8}$$

where $\odot$ denotes element-wise multiplication. With a shared value branch and depth-wise operators, PVA remains efficient while improving instance-relevant evidence selection for subsequent boundary-refined decoding.

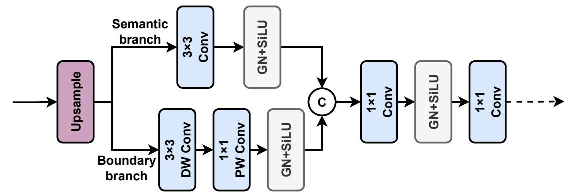

Fig. 4. Structure of UQBR-Head. A semantic branch and a boundary branch are fused by channel concatenation.

D. Underwater Quality Boundary Refinement Head

After UWAF and PVA, RoI features become cleaner and more instance-relevant by suppressing degradation-amplified clutter and reweighting useful evidence. Nevertheless, underwater masks are still prone to boundary uncertainty and instance merging, because scattering blur weakens contours and small/slender structures are easily overwhelmed by residual textures. To enhance boundary sensitivity with a lightweight design, we introduce UQBR-Head, as shown in Fig.4.

Given the RoI feature from PVA, $\boldsymbol{F}_i^{\text{pva}} \in \mathbb{R}^{C\times H\times W}$, we first upsample it to a higher-resolution mask feature:

$$\widetilde{\boldsymbol{F}}_i = \text{Up}(\boldsymbol{F}_i^{\text{pva}}), \quad \widetilde{\boldsymbol{F}}_i \in \mathbb{R}^{C\times H'\times W'}, \tag{9}$$

where $H' = W' = 28$ in our implementation.

UQBR-Head then decouples mask decoding into a semantic branch and a boundary branch. The semantic branch aggregates region-level cues to stabilize interior predictions:

$$\boldsymbol{F}_i^s = \phi\Big(\text{GN}\Big(\text{Conv}_{3\times3}(\widetilde{\boldsymbol{F}}_i)\Big)\Big), \tag{10}$$

where $\phi(\cdot)$ denotes the SiLU activation. The boundary branch emphasizes contour-sensitive responses using depth-wise filtering followed by point-wise mixing:

$$\boldsymbol{F}_i^b = \phi\Big(\text{GN}\Big(\text{PWConv}_{1\times1}\Big(\text{DWConv}_{3\times3}(\widetilde{\boldsymbol{F}}_i)\Big)\Big)\Big). \tag{11}$$

We fuse the two branches by channel-wise concatenation and lightweight 1×1 projections:

$$\boldsymbol{F}_i^u = \phi\big(\text{GN}\big(\text{Conv}_{1\times1}([\boldsymbol{F}_i^s; \boldsymbol{F}_i^b])\big)\big), \tag{12}$$

and obtain the stage-1 mask logits by a final 1×1 convolution:

$$\boldsymbol{M}_i^{(1)} = \text{Conv}_{1\times1}(\boldsymbol{F}_i^u). \tag{13}$$

Together with UWAF and PVA, UQBR-Head improves boundary delineation and instance separation under degradations.

E. Loss Function

All modules are trained end-to-end under the standard two-stage instance segmentation objective. We keep the detection losses unchanged and apply auxiliary supervision to the stage-1 mask M_1 in UQBR-Head. The overall loss is

$$L = L_{\text{det}} + \lambda_m L_{\text{mask}}(\boldsymbol{M}_2) + \lambda_a L_{\text{mask}}(\boldsymbol{M}_1), \tag{14}$$

where L_{det} denotes the RPN and box-head losses, and $\boldsymbol{M}_2$ is the final refined mask prediction. The mask loss L_{mask} is implemented as a combination of binary cross-entropy and Dice loss.

IV. Experiments

A. Datasets and Evaluation Metrics

Dataset. We conduct experiments on two representative underwater instance-level benchmarks. UIIS [2] targets general underwater instance segmentation with category-aware instance masks under challenging conditions such as color cast, low contrast, and frequent occlusions. USIS10K [3] focuses on underwater salient instance segmentation, where visually prominent foreground instances are annotated with pixel-accurate masks. For both datasets, we follow the official training, validation, and test splits and the standard evaluation protocols to ensure fair comparisons.

Evaluation Metrics. We adopt the COCO-style instance segmentation metrics. Performance is primarily reported by mAP averaged over IoU thresholds from 0.50 to 0.95, together with AP50 and AP75. Since underwater scenes often exhibit blurred contours and severe instance merging, AP75 is emphasized to reflect boundary quality and high-IoU mask accuracy. When appropriate, we also report AP for different object scales (small/medium/large) following the standard definition.

B. Implementation Details

All experiments are implemented in PyTorch based on the MMDetection framework. We train the models on a single NVIDIA A40 GPU with a batch size of 2, using SGD as the optimizer with an initial learning rate of 2.5e-3. For fair comparisons, all methods are trained and evaluated under the same dataset splits and evaluation protocols.

C. Experimental Results

We evaluate ULSNet on UIIS for underwater instance segmentation and on USIS10K for salient instance segmentation under both class-agnostic and multi-class protocols.

Instance Segmentation. Table I compares ULSNet with recent methods on UIIS under the same two-stage pipeline. Across all LSNet scales, ULSNet improves both mAP and AP_{75}. With LSNet-T, ULSNet raises LSNet+Mask R-CNN from 27.2 mAP and 29.0 AP_{75} to 29.6 mAP and 31.8 AP_{75}, yielding +2.4 mAP and +2.8 AP_{75} respectively. With LSNet-S, it reaches 31.0 mAP and 34.5 AP_{75}, exceeding WaterMask R-CNN (ResNet-101) by +3.8 mAP and exceeding Mask2Former by +5.3 mAP. With LSNet-B, ULSNet obtains 31.6 mAP and 35.5 AP_{75} respectively, outperforming UIS-Mamba-B by +0.4 mAP and +1.0 AP_{75} while using fewer parameters (60M versus 115M).

Class-Agnostic Salient Instance Segmentation. Table II (Class-Agnostic) shows that ULSNet generalizes well to salient instance segmentation and achieves strong performance across different LSNet scales. ULSNet-T attains 63.2 mAP and 72.0 AP_{75} respectively, improving upon LSNet-T and also surpasses UIS-Mamba-T on AP_{75}. ULSNet-S further improves to 63.9 mAP and 72.8 AP_{75} respectively, exceeding UIS-Mamba-S, and ULSNet-B achieves the best performance with 65.7 mAP, demonstrating a favorable balance between accuracy and efficiency compared with large SAM-based baselines.

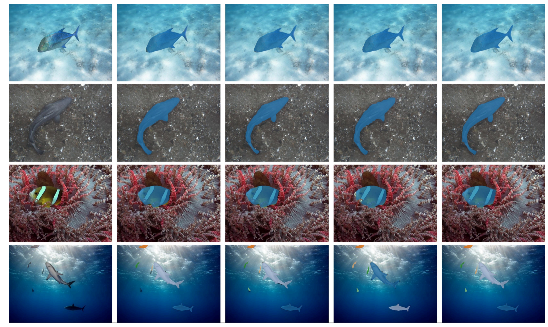

Fig. 5. Visualization of instance segmentation results on the UIIS and USIS10K datasets.

Multi-Class Salient Instance Segmentation. Under the multi-class setting in Table II, ULSNet also delivers consistent gains. ULSNet-T achieves 44.0 mAP and 50.1 AP_{75}, improving over LSNet-T and UIS-Mamba-T, while ULSNet-S reaches 45.9 mAP and 53.2 AP_{75} respectively, surpassing UIS-Mamba-S. With the B backbone, ULSNet-B attains 47.7 mAP and 54.6 AP_{75} respectively, outperforming UIS-Mamba-B and substantially exceeding LSNet-B. These results verify that progressively improving RoI evidence via UWAF and PVA, together with boundary refinement by UQBR-Head, is effective across both class-agnostic and category-aware salient segmentation.

Fig. 5 shows visualization comparisons of representative methods and our approach on UIIS and USIS10K. Our method produces more complete masks with cleaner boundaries under typical underwater degradations (e.g., blur, low contrast). Notably, UIS-Mamba under-segments thin structures (e.g., incomplete fish-tails in row 2), while WaterMask misses small targets (e.g., undetected fishes in row 4). These observations confirm that the proposed RoI evidence enhancement and boundary-refined decoding improve robustness against instance merging and fine-structure loss in challenging underwater scenes.

D. Ablation Studies

Effectiveness of Individual Modules. We evaluate the contribution of each component on UIIS using LSNet-B as the backbone. As shown in Table III, the baseline achieves 29.0 mAP, 47.2 AP_{50}, and 31.9 AP_{75} respectively. Adding UWAF alone improves the results to 30.0 mAP, 47.6 AP_{50}, and 33.1 AP_{75} respectively, indicating that underwater-aware RoI recalibration helps suppress degradation-amplified clutter. Introducing PVA alone also yields consistent gains, reaching 29.8 mAP, 47.5 AP_{50}, and 32.8 AP_{75} respectively, suggesting more instance-relevant evidence selection. In contrast, UQBR-Head brings the largest improvement on AP_{75}, achieving 34.1 AP_{75} together with 30.1 mAP, which aligns with its boundary-refinement goal. When all three modules are enabled, the full model attains 31.6 mAP, 49.8 AP_{50}, and 35.5 AP_{75} respectively, confirming that UWAF, PVA, and UQBR-Head are complementary.

Effectiveness of the two UWAF branches. We further ablate the two gating branches in UWAF. As shown in Table IV,

TABLE I
COMPARISON WITH EXISTING METHODS ON THE UIIS DATASET.

Method	Backbone	Params (M)	mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Mask R-CNN [7]	ResNet-50	50	23.5	42.3	23.7	7.8	19.3	34.9
WaterMask R-CNN [2]	ResNet-50	54	26.4	43.6	28.8	9.1	21.1	38.1
UIS-Mamba [5]	UIS-Mamba-T	56	29.4	46.7	31.3	10.1	22.5	41.9
LSNet + Mask R-CNN [4]	LSNet-T	31	27.2	46.0	29.0	9.4	23.4	40.2
ULSNet(Ours)	LSNet-T	45	**29.6**	**46.8**	**31.8**	**10.7**	**24.3**	**42.3**
Mask R-CNN [7]	ResNet-101	63	23.4	40.9	25.3	9.3	19.8	32.5
Mask Scoring R-CNN [8]	ResNet-101	79	24.6	41.9	26.5	8.4	20.0	34.3
Cascade Mask R-CNN [9]	ResNet-101	88	25.5	42.8	27.8	7.5	20.1	35.0
BMask R-CNN [10]	ResNet-101	66	22.1	36.2	24.4	5.8	17.5	35.0
PointRend [11]	ResNet-101	63	25.9	43.4	27.6	8.2	20.2	38.6
R^3-CNN [12]	ResNet-101	77	24.9	40.5	27.8	9.7	21.4	33.6
Mask Transfiner [13]	ResNet-101	63	24.6	42.1	26.0	7.2	19.4	36.1
Mask2Former [14]	ResNet-101	63	25.7	38.0	27.7	6.3	18.9	38.1
WaterMask R-CNN [2]	ResNet-101	67	27.2	43.7	29.3	9.0	21.8	38.9
UIS-Mamba [5]	UIS-Mamba-S	76	30.4	48.6	33.2	10.2	23.3	42.7
LSNet + Mask R-CNN [4]	LSNet-S	36	27.5	46.8	30.7	10.3	24.3	41.3
ULSNet(Ours)	LSNet-S	53	**31.0**	**48.4**	**34.5**	**11.1**	**25.5**	**42.5**
SAM+Point [15]	ViT-H	641	26.5	44.5	26.9	22.3	35.0	25.6
USIS-SAM [3]	ViT-H	700	29.4	45.0	32.3	9.8	22.1	42.0
UIS-Mamba [5]	UIS-Mamba-B	115	31.2	49.1	34.5	10.4	24.2	43.5
LSNet + Mask R-CNN [4]	LSNet-B	43	29.0	47.2	31.9	10.7	26.5	42.7
ULSNet(Ours)	LSNet-B	60	**31.6**	**49.8**	**35.5**	**11.6**	**26.8**	**46.9**

TABLE II
COMPARISON WITH EXISTING METHODS ON THE USIS10K DATASET UNDER CLASS-AGNOSTIC AND MULTI-CLASS SETTINGS.

Method	Backbone	Params (M)	Class-Agnostic			Multi-Class		
			mAP	AP_{50}	AP_{75}	mAP	AP_{50}	AP_{75}
S4Net [16]	ResNet-50	47	32.8	64.1	27.3	23.9	43.5	24.4
RDPNet [17]	ResNet-50	49	53.8	77.8	61.9	37.9	55.3	42.7
OQTR [18]	ResNet-50	50	56.6	79.3	62.6	19.7	30.6	21.9
WaterMask [2]	ResNet-50	54	58.3	80.2	66.5	37.7	54.0	42.5
UIS-Mamba [5]	UIS-Mamba-T	56	62.2	84.0	71.3	42.1	59.6	48.3
LSNet + Mask R-CNN [4]	LSNet-T	31	60.6	80.7	68.1	38.9	56.2	46.8
ULSNet(Ours)	LSNet-T	45	**63.2**	**83.7**	**72.0**	**44.0**	**59.5**	**50.1**
RDPNet [17]	ResNet-101	66	54.7	78.3	63.0	39.3	55.9	45.4
WaterMask [2]	ResNet-101	67	59.0	80.6	67.2	38.7	54.9	43.2
UIS-Mamba [5]	UIS-Mamba-S	76	63.1	85.1	72.0	44.5	61.5	51.1
LSNet + Mask R-CNN [4]	LSNet-S	36	61.3	81.5	69.8	40.8	57.7	49.0
ULSNet(Ours)	LSNet-S	53	**63.9**	**86.3**	**72.8**	**45.9**	**62.7**	**53.2**
SAM+BBox [15]	ViT-H	641	45.9	65.9	52.1	26.4	38.9	29.0
SAM+Mask [15]	ViT-H	641	55.1	80.2	62.8	38.5	56.3	44.0
RSPrompter [19]	ViT-H	632	58.2	79.9	65.9	40.2	55.3	44.8
USIS-SAM [3]	ViT-H	701	59.7	81.6	71.7	43.1	59.0	48.5
UIS-Mamba [5]	UIS-Mamba-B	115	63.8	86.0	72.8	46.2	63.2	53.4
LSNet + Mask R-CNN [4]	LSNet-B	43	64.7	83.6	71.1	44.7	60.5	51.1
ULSNet(Ours)	LSNet-B	60	**65.7**	**87.2**	**73.5**	**47.7**	**63.7**	**54.6**

TABLE III
ABLATION STUDY OF MODULES.

UWAF	PVA	UQBR-Head	mAP	AP_{50}	AP_{75}
			29.0	47.2	31.9
✓			30.0	47.6	33.1
	✓		29.8	47.5	32.8
		✓	30.1	47.3	34.1
✓	✓		30.8	48.5	34.1
✓		✓	31.0	48.2	34.9
	✓	✓	30.9	48.0	34.8
✓	✓	✓	31.6	49.8	35.5

enabling only the spatial gate improves the baseline to 29.7 mAP, 47.3 AP_{50}, and 32.6 AP_{75}, while enabling only the channel gate yields 29.6 mAP, 47.4 AP_{50}, and 32.3 AP_{75} respectively. Enabling both branches achieves the best performance with 30.0 mAP, 47.6 AP_{50}, and 33.1 AP_{75}, validating their complementarity.

Effectiveness of value-branch sharing in PVA. We further study the design choice of sharing the value branch in PVA. Specifically, we keep UWAF and UQBR-Head fixed and only replace the shared value tensor V with two independent value branches (non-shared) while leaving the preference heads unchanged. As shown in Table V, sharing V yields consistently better results: compared with the non-shared variant, the shared design improves mAP from 31.3 to 31.6, AP_{50} from

TABLE IV
ABLATION STUDY OF THE TWO UWAF BRANCHES.

Spatial gate	Channel gate	mAP	AP_{50}	AP_{75}
		29.0	47.2	31.9
✓		29.7	47.3	32.6
	✓	29.6	47.4	32.3
✓	✓	30.0	47.6	33.1

TABLE V
ABLATION STUDY OF VALUE-BRANCH SHARING.

Value branch in PVA	mAP	AP_{50}	AP_{75}
Non-shared (V non-shared)	31.3	49.0	35.1
Shared (V shared)	31.6	49.8	35.5

49.0 to 49.8, and AP_{75} from 35.1 to 35.5 respectively. This indicates that a shared value branch provides a more stable instance-evidence basis, enabling preference-guided reweighting to focus on underwater-relevant cues without fragmenting the evidence representation.

Generality of PVA and UQBR-Head. To verify that PVA and UQBR-Head are backbone-agnostic, we integrate them into Mask R-CNN with a ResNet-101 backbone. As shown in Table VI, PVA alone improves the baseline from 23.4 mAP, 40.9 AP_{50}, and 25.3 AP_{75} to 24.6 mAP, 42.2 AP_{50}, and 26.5 AP_{75}, while UQBR-Head alone increases AP_{75} to 28.0. Combining both modules further improves the results to 26.1 mAP, 43.0 AP_{50}, and 29.2 AP_{75} respectively, demonstrating their plug-and-play effectiveness across backbones.

TABLE VI
GENERALITY OF PVA AND UQBR-HEAD WITH MASK R-CNN.

PVA	UQBR-Head	mAP	AP_{50}	AP_{75}
		23.4	40.9	25.3
✓		24.6	42.2	26.5
	✓	24.8	41.6	28.0
✓	✓	26.1	43.0	29.2

V. CONCLUSION

In this paper, we presented ULSNet, an efficient underwater instance segmentation framework built upon LSNet and strengthened by three lightweight modules: UWAF for underwater-aware RoI recalibration, PVA for preference–value evidence selection, and UQBR-Head for boundary-refined mask decoding. Extensive experiments on UIIS and USIS10K demonstrate that ULSNet consistently improves the accuracy–efficiency trade-off, yielding more complete masks and cleaner boundaries under underwater degradations. Future work will explore stronger cross-domain generalization and further acceleration for real-time deployment on marine platforms.

ACKNOWLEDGEMENT

This work was supported in part by the Jilin Provincial Nature Science Foundation, China (No.YDZJ202201ZYTS566); in part by Education Department of Jilin Province of China (No.JJKH20240682KJ); and in part by National Natural Science Foundation of China (No.62062064).

REFERENCES

[1] R. Kaneko, Y. Sato, T. Ueda, H. Higashi, and Y. Tanaka, "Marine snow removal benchmarking dataset," in *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2023, pp. 771–778.

[2] S. Lian, H. Li, R. Cong, S. Li, W. Zhang, and S. Kwong, "Watermask: Instance segmentation for underwater imagery," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 1305–1315.

[3] S. Lian, Z. Zhang, H. Li, W. Li, L. T. Yang, S. Kwong, and R. Cong, "Diving into underwater: Segment anything model guided underwater salient instance segmentation and a large-scale dataset," in *Forty-first International Conference on Machine Learning*, 2024.

[4] A. Wang, H. Chen, Z. Lin, J. Han, and G. Ding, "Lsnet: See large, focus small," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 9718–9729.

[5] R. Cong, Z. Yu, H. Fang, H. Sun, and S. Kwong, "Uis-mamba: exploring mamba for underwater instance segmentation via dynamic tree scan and hidden state weaken," in *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025, pp. 343–352.

[6] W. Peng, B. Li, P. Wang, H. Huang, Y. Zou, and X. Qiao, "Pvconv: Enhancing depthwise separable convolution via preference-value learning for similar-feature discrimination," *Electronics*, vol. 14, no. 24, 2025.

[7] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[8] Z. Huang, L. Huang, Y. Gong, C. Huang, and X. Wang, "Mask scoring r-cnn," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6409–6418.

[9] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6154–6162.

[10] T. Cheng, X. Wang, L. Huang, and W. Liu, "Boundary-preserving mask r-cnn," in *European conference on computer vision*. Springer, 2020, pp. 660–676.

[11] C. Li, C. Guo, W. Ren, R. Cong, J. Hou, S. Kwong, and D. Tao, "An underwater image enhancement benchmark dataset and beyond," *IEEE transactions on image processing*, vol. 29, pp. 4376–4389, 2019.

[12] L. Rossi, A. Karimi, and A. Prati, "Recursively refined r-cnn: Instance segmentation with self-roi rebalancing," in *International Conference on Computer Analysis of Images and Patterns*. Springer, 2021, pp. 476–486.

[13] L. Ke, M. Danelljan, X. Li, Y.-W. Tai, C.-K. Tang, and F. Yu, "Mask transfiner for high-quality instance segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4412–4421.

[14] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 1290–1299.

[15] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4015–4026.

[16] R. Fan, M.-M. Cheng, Q. Hou, T.-J. Mu, J. Wang, and S.-M. Hu, "S4net: Single stage salient-instance segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6103–6112.

[17] Y.-H. Wu, Y. Liu, L. Zhang, W. Gao, and M.-M. Cheng, "Regularized densely-connected pyramid network for salient instance segmentation," *IEEE Transactions on Image Processing*, vol. 30, pp. 3897–3907, 2021.

[18] J. Pei, T. Cheng, H. Tang, and C. Chen, "Transformer-based efficient salient instance segmentation networks with orientative query," *IEEE Transactions on Multimedia*, vol. 25, pp. 1964–1978, 2022.

[19] K. Chen, C. Liu, H. Chen, H. Zhang, W. Li, Z. Zou, and Z. Shi, "Rsprompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–17, 2024.