Sensors and Materials, Vol. 38, No. 7 (2026) 4245–4258
MYU Tokyo

S & M 4562

Entropy-guided Confidence Evaluation for Semantic Segmentation of Marine Debris

Seung Bae Jeon,[1] Yun-Woong Choi,[2] Jung-Hwan Lee,[3] and Myeong-Hun Jeong[1*]

[1]Department of Civil Engineering, Chosun University,
309 Pilmun-daro, Dong-gu, Gwangju 61452, Republic of Korea
[2]Department of Civil Construction Engineering, Chosun College of Science & Technology,
11 Chosunigongdae-gil, Dong-gu, Gwangju 61453, Republic of Korea
[3]Department of Civil & Environmental Engineering, Dongshin University,
34-22 Dongshindae-gil, Naju-si, Jeollanam-do 58245, Republic of Korea

(Received November 21, 2025; accepted June 18, 2026)

Keywords: marine debris segmentation, underwater vision, uncertainty quantification, epistemic uncertainty

Marine debris segmentation in underwater environments remains a challenging task owing to visual degradation caused by light absorption, scattering, and turbidity. Although deep-learning-based vision systems have shown promising results, their prediction confidence under such conditions is often unreliable. In this study, we propose an entropy-guided uncertainty quantification framework using the HyenaPixel model, an attention-free architecture capable of capturing global context with large convolutional kernels. We apply Monte Carlo Dropout during inference to estimate epistemic uncertainty and analyze its correlation with per-class segmentation performance. Experiments conducted on a curated split of a public Ocean Debris Segmentation dataset containing 22 classes demonstrate that uncertainty is a meaningful indicator of model reliability, and may reveal overconfident mispredictions and unstable classes. The proposed lightweight framework can be readily integrated into underwater vision pipelines, offering interpretable uncertainty measures that are essential for reliable and risk-aware decision-making.

1. Introduction

Marine plastic pollution has emerged as one of the most pressing global environmental challenges. It is estimated that more than eight million metric tons of plastic waste enter the ocean annually, posing severe threats to marine ecosystems and human health.(1) Autonomous underwater vehicles (AUVs) and remotely operated vehicles (ROVs) have been increasingly deployed for marine debris monitoring and removal. These platforms rely on computer vision systems for object detection and segmentation to enable efficient localization and grasping of debris.

*Corresponding author: e-mail: mhjeong@chosun.ac.kr
https://doi.org/10.18494/SAM6191

ISSN 0914-4935

However, underwater environments present severe visual degradation due to light absorption, scattering, turbidity, and color distortion. These conditions significantly impair the reliability of perception models trained on conventional datasets. Although recent deep-learning-based approaches such as Mask Region-based Convolutional Neural Network (R-CNN), U-Net, and Transformer variants have achieved remarkable segmentation performance, their prediction confidence under adverse visual conditions remains largely unexplored.

To address this limitation, uncertainty quantification (UQ) has gained attention as a means of evaluating the reliability of deep learning predictions. Bayesian approaches such as Monte Carlo (MC) Dropout[2] provide an efficient way to estimate epistemic uncertainty, which captures the model's lack of knowledge about unseen data. Understanding these uncertainties is essential for risk-aware decision-making in autonomous systems, particularly in mission-critical underwater robotics.

In this study, we propose an entropy-guided confidence evaluation framework for semantic segmentation of marine debris using the HyenaPixel architecture, a lightweight, attention-free model capable of capturing global context through large convolutional kernels. We integrate MC Dropout into the inference stage to estimate epistemic uncertainty and analyze its correlation with class-wise performance on a custom Ocean Debris Segmentation dataset. Our contributions are summarized as follows.

- We conduct a comprehensive analysis of class-wise uncertainty in underwater debris segmentation, providing insights into model confidence across diverse object categories.
- We demonstrate the relationship between entropy-based uncertainty and segmentation accuracy, revealing overconfident mispredictions.
- We develop a lightweight and interpretable uncertainty quantification framework that can be extended to robotic vision systems operating in underwater environments.

The remainder of this paper is organized as follows. In Sect. 2, we review related work on marine debris segmentation and uncertainty quantification. In Sect. 3, we describe the dataset. In Sect. 4, we outline the proposed methodology. In Sect. 5, we present experimental results and discussions, followed by conclusions in Sect. 6.

2. Related Work

2.1 Marine debris detection and segmentation

Research on vision-based marine debris perception has evolved along two complementary sensing lines: optical cameras for near-field recognition and manipulation, and forward-looking sonar (FLS) for perception under severe turbidity where light-based imaging degrades. Early deep learning approaches ported general-purpose semantic segmentation and detection architectures for underwater imagery, while deployable models suitable for embedded AUV/ROV platforms are sought in more recent work.

For acoustic imagery, the Marine Debris Dataset for FLS provides pixel-level labels and demonstrates that convolutional encoders such as UNet-ResNet34 can reach competitive segmentation quality despite the low signal-to-noise characteristics inherent to sonar.[3]

For optical imagery, datasets such as TrashCan established the feasibility of debris instance segmentation with Mask RCNN over common categories including bottles, nets, and ropes, and the SeaClear dataset broadened geographic and environmental coverage via ROV acquisitions and benchmarking detectors such as Faster RCNN and YOLOv6 under varying illumination, color cast, and turbidity.[4,5] Additional collections including open community sets such as the Roboflow Ocean Debris Segmentation[6] have catalyzed experimentation with pixel-wise labeling across diverse debris types, but they also expose strong class imbalance and domain shift across sites and campaigns.

Methodologically, the field has inherited a progression of segmentation architectures from generic computer vision. Canonical encoders/decoders FCN, UNet, SegNet, PSPNet, and DeepLabv3+ remain strong baselines and are commonly paired with underwater-specific pre- and postprocessing such as color restoration and dehazing.[7–10] The advent of Transformer-based segmentation (e.g., SegFormer, SwinUNet, and Mask2Former) improved global context modeling, yet their memory/latency footprint often clashes with the power and real-time constraints of ROV-class hardware.[11–13]

Consequently, real-time/lightweight designs such as ERFNet, FastSCNN, DDRNet, and PIDNet are increasingly favored to balance accuracy against throughput on embedded GPUs.[14–17] Parallel to architectural choices, underwater image enhancement and physics-informed restoration (e.g., SeaThru and related pipelines) mitigate color attenuation and backscatter to some extent, and learning-based enhancement (e.g., FUnIEGAN) has been used as a front-end to improve downstream recognition in challenging scenes.[18,19]

Despite these advances, the prevailing evaluation protocol is still accuracy-centric, typically reporting mIoU, mAP, or pixel accuracy. This is at odds with the operational reality of autonomous underwater missions where decisions such as grasping or avoidance must be conditioned on how reliable a segmentation is *in situ*. In particular, boundary-heavy and visually ambiguous categories (e.g., nets, ropes, and transparent plastics) exhibit overconfident errors in degraded water, making explicit confidence assessment an essential ingredient for risk-aware autonomy.

In this context, attention-free backbones that capture long-range dependencies at low computational cost are attractive because they enable uncertainty estimation with minimal overhead. HyenaPixel exemplifies this direction by leveraging large implicit convolutional filters to model global context efficiently, dispensing with self-attention while preserving wide receptive fields.[20] Such characteristics are well aligned with on-robot inference budgets and create headroom for the stochastic test-time sampling necessary for uncertainty quantification without violating latency constraints.

2.2 Uncertainty quantification

UQ in deep perception is commonly framed through aleatoric uncertainty stemming from measurement noise, clutter, and intrinsic ambiguity and epistemic uncertainty reflecting ignorance due to limited or biased training data. MC Dropout, introduced by Gal and Ghahramani,[2] provides a simple, architecture-compatible Bayesian approximation: by enabling

dropout at inference and aggregating multiple stochastic forward passes, one approximates the predictive distribution and recovers an estimate of epistemic uncertainty.

Kendall and Gal[21] formalized this decomposition for computer vision, showing how to model both uncertainty types jointly and motivating its use in safety-critical perception. Surveys by Abdar *et al.*[22] and Gawlikowski *et al.*[23] consolidated a broader landscape of methods and applications, emphasizing deployment considerations such as interpretability, calibration, and robustness under shift.

Beyond MC Dropout, several families of UQ methods have been explored. Deep ensembles deliver strong uncertainty estimates and improved out-of-distribution (OOD) behavior by averaging independently trained models, with subsequent work revealing that uncertainty quality can deteriorate under dataset shift and that robustness varies across methods.[24,25]

Evidential approaches model distributions over class probabilities (e.g., Dirichlet evidence), yielding closed-form uncertainty proxies without sampling or extensions to regression;[26,27] however, they may introduce training instability or nontrivial hyperparameters. Probabilistic segmentation decoders such as Probabilistic UNet capture multimodality in ambiguous regions at the cost of added complexity.[28]

Test-time augmentation and other stochastic layers provide inexpensive dispersion signals but may conflate aleatoric and epistemic effects. Irrespective of the estimator, calibration remains a core requirement: modern neural networks are often miscalibrated, with reported confidences misaligned with empirical accuracy.[29]

Calibration metrics such as expected calibration error (ECE) are commonly used to quantify the alignment between predicted confidence and empirical accuracy. However, in dense prediction tasks such as semantic segmentation, entropy-based uncertainty provides an interpretable spatial representation of prediction reliability, complementing calibration-based evaluation.

Post hoc methods such as temperature scaling and binning-based calibration have been adapted to dense prediction;[30,31] segmentation-specific evaluations highlight under/over-confidence near object boundaries and under shift.[32] In more recent work [e.g., uncertainty-aware cross-entropy (UCE)] uncertainty signals are injected directly into the training loss to improve robustness in noisy settings.[33]

For underwater and field robotics, UQ is valuable only insofar as it can be used to gate autonomy. In manipulation, perceptual uncertainty has been integrated into grasp selection to reduce failure rates under clutter and poor visibility, as illustrated by the PUGS framework that propagates sensor and pose uncertainty from 3D occupancy to the action space.[34] Similar principles appear in autonomous driving and aerial robotics, where uncertainty triggers slow-down, replan, or human-in-the-loop policies and guides active learning under domain shift.[25,35]

However, in underwater semantic segmentation specifically, much of the literature remains pixel-centric and qualitative—producing heatmaps without class-aware analysis—despite the fact that task policies are typically class-conditioned (e.g., grasp bottles but avoid nets). Moreover, many UQ pipelines assume heavy backbones or ensembles that are ill-suited to ROV compute envelopes, limiting their practical adoption.

In this study, we address these limitations by coupling a lightweight, attention-free segmentation backbone with MC Dropout-based epistemic UQ to enable real-time sampling at test time, and by shifting from pixel-only visualization to per-class entropy statistics. This design exposes overconfident mispredictions and unstable categories—precisely the failure modes that matter for risk-aware underwater operations—while staying within the latency and power budgets of embedded platforms. In doing so, it advances the literature from accuracy-only reporting and qualitative uncertainty maps toward class-level, interpretable confidence evaluation that can directly inform robotic decision-making.

3. Underwater Debris Segmentation Dataset

We base our experiments on the Ocean Debris Segmentation Dataset, an open-source collection available through Roboflow Universe.[6] The dataset provides pixel-wise annotations for 22 object categories spanning common debris types such as bottles, cups, nets, and ropes as well as marine animals and robotic assets observed in typical AUV/ROV missions. Images cover a range of underwater scenes and acquisition conditions, including variable lighting, turbidity, and visual clutter, mirroring real-world ROV perspectives rather than controlled laboratory imagery. These characteristics make the dataset an appropriate stress test for reliability analysis under adverse visibility and complex backgrounds.

3.1 Data partition and visual characteristics

Following the dataset's official recommendation, we adopt an 80/15/5 split for training, validation, and testing yielding 5,770, 1,081, and 361 images, respectively. This configuration provides sufficient diversity for model learning while maintaining disjoint subsets for fair hyperparameter tuning and final evaluation.

The dataset integrates footage collected under diverse underwater environments, covering various water types, depths, illumination levels, and imaging conditions. Frequent visual degradations such as color attenuation, backscatter, and turbidity significantly reduce contrast and edge sharpness. Scenes often include cluttered seabeds, suspended debris, and robot-proximal objects with partial occlusions, conditions that induce boundary ambiguity and class confusion at the pixel level.

Because these settings closely reflect operational AUV/ROV deployments, the dataset serves as a realistic benchmark for assessing not only segmentation accuracy but also confidence reliability under natural degradation.

3.2 Visual complexity and motivation for uncertainty analysis

The dataset encompasses a wide spectrum of object geometries and visual properties that affect segmentation difficulty. Even rigid items such as trash bottles and trash cups often display specular reflections, partial transparency, and view-dependent brightness, which distort

boundary cues and complicate precise annotation. Semirigid materials such as trash clothing or trash bags frequently deform or fold, creating self-occlusions and irregular silhouettes. Filamentary debris such as trash nets and trash rope appear as thin, entangled structures that are easily confused with the background owing to scattering and overlapping strands.

Similarly, biological entities including starfish and shells often exhibit low color contrast and texture similarity with the surrounding seabed, making their visual separation inherently ambiguous. These characteristics do not imply prior performance outcomes but rather highlight the intrinsic visual complexity that motivates further analysis of model reliability.

Because of these heterogeneous visual conditions, ranging from severe turbidity to variable illumination and backscatter, the dataset naturally introduces noise and domain variability across scenes. Under such conditions, uncertainty quantification becomes an essential complement to conventional evaluation metrics, allowing researchers to examine prediction confidence under visually degraded scenarios.

Entropy-based measures and stochastic inference methods such as MC Dropout provide a way to assess not only pixel-level accuracy but also the stability and reliability of predictions, especially in ambiguous or low-contrast regions. In later sections, these tools are leveraged to explore how uncertainty correlates with visual degradation and to identify cases where perception systems may fail silently under underwater conditions.

4. Methodology

In this section, we present the overall approach adopted in this study, including the segmentation backbone, uncertainty estimation process, and confidence evaluation procedure. The proposed method integrates a lightweight HyenaPixel architecture with stochastic inference and entropy-based metrics to analyze model reliability under challenging underwater conditions.

4.1 Base architecture: HyenaPixel for underwater segmentation

The framework is built upon HyenaPixel, an attention-free vision backbone that replaces self-attention mechanisms with hyena operators, long convolutional filters capable of modeling global dependencies efficiently.

Unlike transformer-based models, HyenaPixel operates entirely through convolutional and implicit sequence filtering, significantly reducing computational overhead while maintaining a large receptive field. This makes it particularly suitable for real-time underwater perception, where onboard resources on AUVs and ROVs are often limited.

In addition, HyenaPixel is particularly well suited to uncertainty quantification tasks, as its lightweight architecture enables efficient multiple stochastic forward passes required for MC Dropout inference. Compared with heavier transformer-based models or ensemble approaches, this efficiency allows the practical deployment of uncertainty-aware segmentation within real-time robotic systems.

The model extracts multiscale representations through hierarchical encoding stages and merges them via upsampling and feature-fusion blocks in the decoder. This design preserves both spatial detail and global context, achieving a strong balance between representation capability and computational efficiency, which is essential for underwater robotic applications.

4.2 Uncertainty estimation via MC Dropout

To quantify epistemic uncertainty, MC Dropout is employed during inference. In this approach, dropout layers remain active at test time, introducing stochasticity across multiple forward passes. Given an input image I, the network produces a set of class probability maps $\{p_t(x)\}_{t=1}^{T}$ for each pixel x across T samples.

The predictive mean and variance are calculated as

$$\mu(x) = \frac{1}{T}\sum_{t=1}^{T} p_t(x), \sigma^2(x) = \frac{1}{T}\sum_{t=1}^{T}(p_t(x) - \mu(x))^2. \quad (1)$$

The variance $\sigma^2(x)$ serves as an estimate of epistemic uncertainty, indicating how consistent the model's predictions are across stochastic passes. Regions with high variance generally correspond to ambiguous boundaries or unseen visual conditions. This method requires no retraining or architectural changes, thus preserving model simplicity and deployability on robotic platforms.

4.3 Entropy-based confidence evaluation

To complement variance-based measures, Shannon entropy is computed from the softmax probability distribution at each pixel:

$$H(x) = -\sum_{c=1}^{C} p_c(x) \log p_c(x), \quad (2)$$

where C denotes the number of semantic classes and $p_c(x)$ is the predicted probability for class c.

Entropy reflects the uncertainty of the final prediction—high entropy indicates low confidence and *vice versa*. Entropy maps provide an interpretable way to visualize the confidence landscape, revealing ambiguous or unstable regions such as low-contrast areas, overlapping debris, or shadowed zones.

By aggregating entropy values per class and comparing them with segmentation metrics (IoU, F1-score), the framework allows the quantitative analysis of model reliability, offering insight into prediction stability under degraded visibility conditions.

5. Experiments and Results

All experiments were conducted using the HyenaPixel segmentation backbone integrated with MC Dropout layers for stochastic inference. Training and evaluation were performed on the Ocean Debris Segmentation Dataset, following the 80/15/5 split defined in Sect. 3. The model was trained for 100 epochs using the AdamW optimizer with an initial learning rate of 1×10^{-4}, weight decay of 0.01, and a batch size of 8 on a single NVIDIA RTX 4090 GPU.

Uncertainty was estimated using MC Dropout sampling at inference time, keeping dropout layers active and performing stochastic forward passes. Sampling iterations were set to $T = 20$, 50, and 100 to analyze convergence behavior in uncertainty estimation.

5.1 Quantitative evaluation

The HyenaPixel segmentation model exhibited strong quantitative performance across the Ocean Debris Segmentation Dataset, showing a clear contrast between rigid and deformable object categories.

The segmentation results exhibit distinct variations across object categories, revealing how visual consistency and scene-dependent appearance affect model performance. Trash bottle, trash clothing, and trash cup achieved the highest IoU and accuracy scores in the dataset, consistently showing stable segmentation boundaries across diverse underwater scenes. Their shapes remain relatively intact across viewpoints, contributing to reproducible mask predictions even under moderate turbidity and illumination change.

In contrast, categories such as trash net, trash rope, and starfish recorded substantially lower IoU scores. These classes often involve thin, overlapping, or highly deformable components that blend with the surrounding water or sediment, leading to fragmented boundaries and frequent misclassification. Such visual ambiguity is further intensified by optical scattering and color attenuation, which degrade local contrast and obscure object contours.

Trash container and trash bag showed intermediate results, fluctuating between scenes depending on illumination direction, water depth, and object pose. Biological categories displayed similarly diverse outcomes: fish achieved relatively high segmentation accuracy, while animal shells and starfish exhibited larger intraclass variation and lower confidence, consistent with their naturally reflective or camouflage-like textures that hinder clear separation from the background.

Overall, these findings indicate that segmentation reliability across the dataset is strongly modulated by surface texture, visual stability, and environmental degradation rather than by any single geometric property. The variability across classes underscores the importance of per-class evaluation and motivates further analysis of uncertainty to understand how prediction confidence behaves under underwater distortions.

Table 1 presents the detailed per-class IoU and accuracy statistics. These results indicate that the HyenaPixel architecture maintains strong representational capacity for rigid and high-contrast debris types while facing notable challenges with flexible or low-contrast classes. Such performance characteristics provide a solid foundation for subsequent uncertainty analysis.

Table 1
Class-wise segmentation performance on the Ocean Debris Segmentation Dataset.

Class	IoU (%)	Accuracy (%)
Animal crab	70.65	84.3
Animal eel	66.88	76.64
Animal etc.	74.06	86.14
Animal fish	90.8	94.51
Animal shells	59.66	72.6
Animal starfish	55.17	63.2
Plant	79.68	86.99
ROV	86.41	92.89
Trash bag	70.75	80.75
Trash bottle	94.89	98.34
Trash branch	83.28	92.94
Trash can	83.31	90.07
Trash clothing	93.57	98.05
Trash container	70.67	95.12
Trash cup	91.24	96.47
Trash net	41.65	47.72
Trash pipe	58.1	66.82
Trash rope	62.87	78.34
Trash snack wrapper	81.06	91.94
Trash tarp	84.96	93.22
Trash unknown instance	72.79	82.44
Trash wreckage	68.12	78.27

5.2 Uncertainty analysis and prediction error

To further evaluate segmentation reliability, class-wise epistemic uncertainty was quantified using MC Dropout with sampling counts of T = 20, 50, and 100. The entropy distributions in Fig. 1 show high consistency across sampling configurations, indicating that uncertainty estimates converge efficiently even under moderate sampling effort. Although the observed entropy range appears relatively narrow (approximately 3.00–3.09), this is expected because entropy is computed over a multiclass probability distribution with 22 categories. As the theoretical entropy range increases with the number of classes, even small numerical differences can correspond to meaningful variations in prediction uncertainty. In this study, no additional normalization was applied to the entropy values.

Classes such as trash cup, starfish, and trash snack wrapper exhibit the highest entropy values, indicating pronounced epistemic uncertainty and unstable predictions under illumination changes and turbidity. Conversely, ROV and animal fish maintain both high IoU and low entropy, reflecting stable and confident predictions across scenes. However, trash rope demonstrates a contradictory trend; it yields low entropy despite poor IoU performance, implying overconfident misclassification where the model produces consistent but inaccurate predictions.

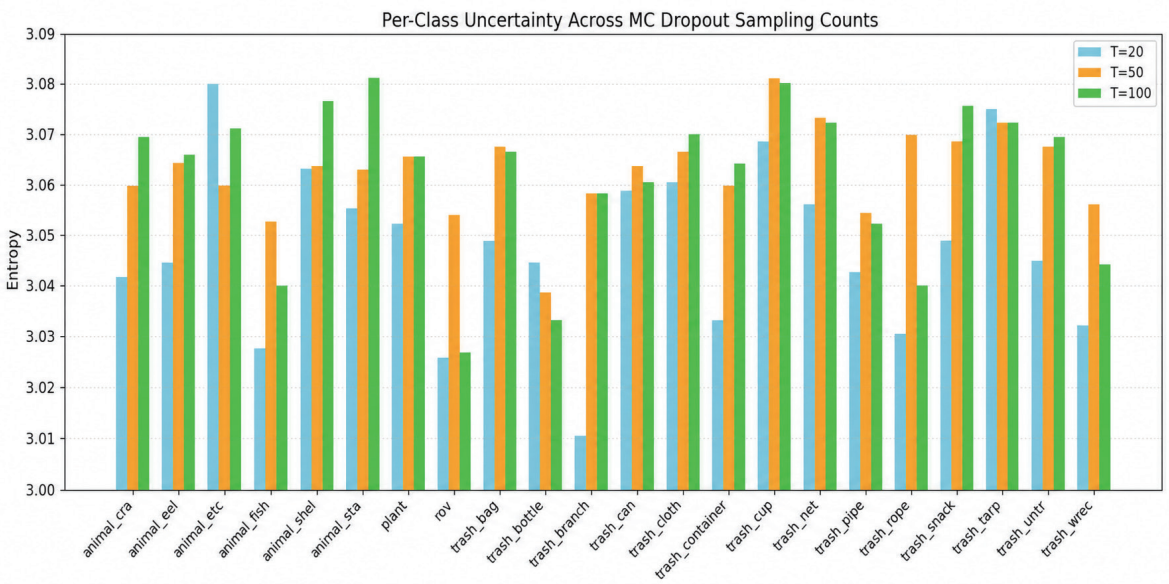

Fig. 1. (Color online) Per-class entropy across different MC Dropout sampling counts ($T = 20, 50, 100$). The results show minimal overall variation but distinct class-wise uncertainty differences between rigid and deformable categories.

These joint observations reveal that epistemic uncertainty not only captures stochastic prediction variance but also exposes mismatches between model confidence and actual segmentation accuracy. Such cases underscore the importance of integrating uncertainty measures into model evaluation, as they help identify overconfident failure modes that conventional metrics like mIoU may overlook.

As illustrated in Fig. 2, the relationship between model negative predictive entropy and prediction error clarifies how uncertainty corresponds to actual performance variation. Three characteristic behavioral clusters can be identified.

- Overconfident Misclassification:
  Classes such as trash pipe, trash rope, and trash wreckage display low entropy yet high prediction error, indicating an overconfidence bias where the model incorrectly assigns high certainty to erroneous outputs.
- Hard-to-Predict Classes:
  Starfish consistently exhibits both high entropy and high error, attributed to camouflage and low-contrast textures that obscure boundary cues.
- Stable and Reliable Classes:
  Trash bottle, animal fish, and ROV maintain low entropy and low error rates, representing robust and confident predictions even under stochastic inference.

These observations demonstrate that uncertainty estimation offers diagnostic insights beyond standard accuracy metrics. Identifying overconfident errors and intrinsically uncertain classes enables more targeted retraining and data curation strategies.

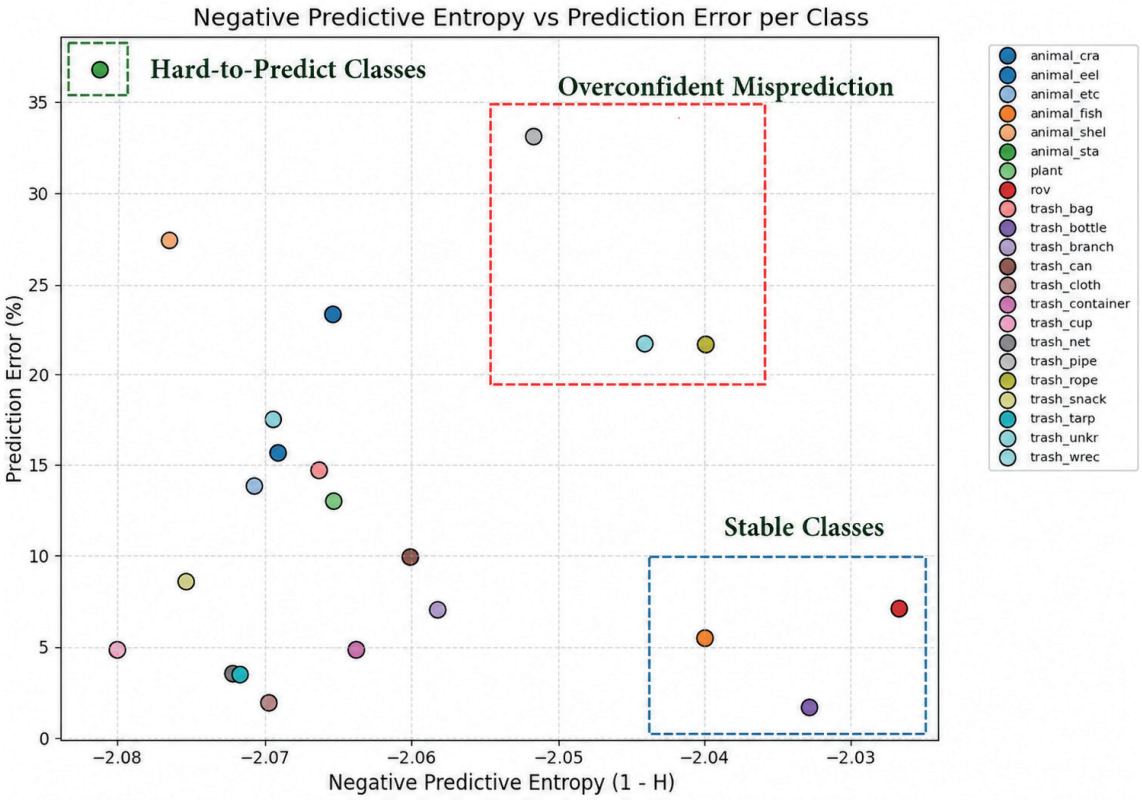

Fig. 2. (Color online) Negative predictive entropy versus prediction error across object classes. Three major groups are highlighted: overconfident mispredictions, hard-to-predict classes, and stable classes with high reliability.

6. Discussion and Conclusions

The integration of UQ into underwater semantic segmentation offers a deeper understanding of model reliability beyond conventional accuracy metrics. Although the global entropy range across MC sampling counts ($T = 20$, 50, 100) remained relatively narrow, cross-analysis with IoU revealed meaningful class-dependent reliability patterns.

Specifically, trash bottle, fish, and ROV achieved high IoU and consistently low entropy, indicating stable and confident predictions across diverse underwater scenes. In contrast, starfish and shells exhibited both elevated entropy and low IoU, reflecting genuine epistemic uncertainty associated with low-contrast textures and camouflage effects under turbid conditions.

A distinct case was observed for trash rope, which maintained low entropy despite poor IoU performance, an indicator of overconfident misprediction, where the model systematically produces confident yet incorrect outputs. These findings demonstrate that entropy-based epistemic uncertainty can expose latent failure modes that traditional metrics such as mean IoU may overlook.

By identifying classes that are either inherently ambiguous or consistently misclassified with undue confidence, UQ enables more targeted retraining, data curation, and domain adaptation strategies. In operational contexts, particularly for AUVs and ROVs, these uncertainty cues can be leveraged to trigger re-observation, confidence-based decision thresholds, or risk-aware task scheduling.

In practice, entropy values can be translated into explicit decision rules for robotic systems. For example, predictions whose mean entropy exceeds a predefined threshold τ can be excluded from autonomous grasp candidates, triggering re-observation or deferral to human operators. Conversely, predictions with low entropy can be prioritized for immediate execution. Such threshold-based decision mechanisms enable the integration of uncertainty estimation into practical AUV/ROV operations, facilitating more reliable and risk-aware behavior under uncertain underwater conditions. Such integration fosters safer and more interpretable robotic perception in visually degraded underwater environments.

Moreover, the convergence of entropy distributions across sampling configurations suggests that reliable uncertainty estimation can be achieved with moderate computational cost (e.g., $T \approx 50$). This efficiency aligns well with lightweight architectures such as HyenaPixel, facilitating real-time uncertainty-aware inference on embedded systems deployed in field operations. Hence, the proposed framework demonstrates a practical pathway toward coupling efficient deep segmentation with interpretable confidence assessment.

Nonetheless, this study is limited to epistemic uncertainty derived from stochastic dropout inference. Aleatoric uncertainty, originating from intrinsic sensor noise, light scattering, and environmental variability, remains unmodeled. Future research should incorporate both uncertainty types to disentangle the contributions of data ambiguity and model ignorance.

In addition, extending the framework to multimodal fusion—combining optical and sonar imagery—can further enhance robustness under severe visibility degradation. In conclusion, uncertainty quantification proves essential for evaluating and improving underwater segmentation reliability. By bridging quantitative performance with confidence interpretation, it provides a principled foundation for risk-aware, trustworthy, and adaptive underwater vision systems.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) (RS-2024-00354270) funded by the Korean government. (MSIT).

References

1 J. R. Jambeck, R. Geyer, C. Wilcox, T. R. Siegler, M. Perryman, A. Andrady, R. Narayan, and K. L. Law: Sci. **347** (2015) 768. https://doi.org/10.1126/science.1260352

2 Y. Gal and Z. Ghahramani: Proc. Int. Conf. Machine Learning (2016) 1050. https://doi.org/10.48550/arXiv.1506.02142

3 D. Singh and M. Valdenegro-Toro: Proc. IEEE/CVF Int. Conf. Computer Vision Workshops (2021) 3741–3749. https://doi.org/10.1109/ICCVW54120.2021.00417

4 J. Hong, M. Fulton, and J. Sattar: arXiv preprint arXiv:2007.08097 (2020). https://doi.org/10.48550/arXiv.2007.08097
5 A. Đuraš, B. J. Wolf, A. Ilioudi, I. Palunko, and B. De Schutter: Sci. Data **11** (2024) 921. https://doi.org/10.1038/s41597-024-03759-2
6 Debris Oceans: Ocean Debris Segmentation Dataset. https://universe.roboflow.com/debris-oceans/ocean-debris-segmentation
7 N. Nezla, T. M. Haridas, and M. H. Supriya: Proc. Int. Conf. Advanced Computer Communication Systems (ICACCS) (2021) 28–33. https://doi.org/10.1109/ICACCS51430.2021.9441804
8 V. Badrinarayanan, A. Kendall, and R. Cipolla: IEEE Trans. Pattern Anal. Mach. Intell. **39** (2017) 2481. https://doi.org/10.1109/TPAMI.2016.2644615
9 H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia: Proc. IEEE Conf. Computer Vision Pattern Recognition (CVPR) (2017) 2881–2890. https://doi.org/10.1109/CVPR.2017.660
10 L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam: Proc. Eur. Conf. Computer Vision (ECCV) (2018) 801–818. https://doi.org/10.1007/978-3-030-01234-2_49
11 E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo: Adv. Neural Inf. Process. Syst. **34** (2021) 12077. https://doi.org/10.48550/arXiv.2105.15203
12 Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo: Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV) (2021) 10012–10022. https://doi.org/10.1109/ICCV48922.2021.00986
13 B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar: Proc. IEEE/CVF Conf. Computer Vision Pattern Recognition (CVPR) (2022) 1280–1289. https://doi.org/10.1109/CVPR52688.2022.00135
14 E. Romera, J. M. Alvarez, L. M. Bergasa, and R. Arroyo: IEEE Trans. Intell. Transp. Syst. **19** (2018) 263. https://doi.org/10.1109/TITS.2017.2750080
15 R. P. Poudel, S. Liwicki, and R. Cipolla: arXiv preprint arXiv:1902.04502 (2019). https://doi.org/10.48550/arXiv.1902.04502
16 Y. Hong, H. Pan, W. Sun, and Y. Jia: arXiv preprint arXiv:2101.06085 (2021). https://doi.org/10.48550/arXiv.2101.06085
17 J. Xu, Z. Xiong, and S. P. Bhattacharyya: PIDNet: A real-time semantic segmentation network inspired by PID controllers, arXiv preprint arXiv:2206.02066 (2022). https://doi.org/10.48550/arXiv.2206.02066
18 D. Akkaynak and T. Treibitz: Proc. IEEE/CVF Conf. Computer Vision Pattern Recognition (CVPR) (2019) 1682–1691. https://doi.org/10.1109/CVPR.2019.00178
19 M. J. Islam, Y. Xia, and J. Sattar: Fast underwater image enhancement for improved visual perception, arXiv preprint arXiv:1903.09766 (2019). https://doi.org/10.48550/arXiv.1903.09766
20 J. Spravil, S. Houben, and S. Behnke: arXiv preprint arXiv:2402.19305 (2024). https://doi.org/10.48550/arXiv.2402.19305
21 A. Kendall and Y. Gal: Advances in Neural Information Processing Systems 30 (NeurIPS 2017). https://doi.org/10.48550/arXiv.1703.04977
22 M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, and U. R. Acharya: Inf. Fusion **76** (2021) 243. https://doi.org/10.1016/j.inffus.2021.05.008
23 J. Gawlikowski, C. R. N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, and R. Roscher: arXiv preprint arXiv:2107.03342 (2021). https://doi.org/10.48550/arXiv.2107.03342
24 B. Lakshminarayanan, A. Pritzel, and C. Blundell: Advances in Neural Information Processing Systems (NeurIPS) **30** (2017). https://doi.org/10.48550/arXiv.1612.01474
25 Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. Dillon, B. Lakshminarayanan, and J. Snoek: Advances in Neural Information Processing Systems (NeurIPS) **32** (2019). https://doi.org/10.48550/arXiv.1906.02530
26 M. Sensoy, L. Kaplan, and M. Kandemir: Advances in Neural Information Processing Systems (NeurIPS) **31** (2018) 3179. https://doi.org/10.48550/arXiv.1806.01768
27 A. Amini, W. Schwarting, A. Soleimany, and D. Rus: Advances in Neural Information Processing Systems (NeurIPS) **33** (2020) 14927. https://doi.org/10.48550/arXiv.1910.02600
28 S. A. A. Kohl, B. Romera-Paredes, C. Meyer, J. De Fauw, J. R. Ledsam, K. Maier-Hein, S. M. A. Eslami, D. Jimenez Rezende, and O. Ronneberger: Advances in Neural Information Processing Systems (NeurIPS) **31** (2018) 6965. https://doi.org/10.48550/arXiv.1806.05034
29 C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger: Proc. Int. Conf. Machine Learning (ICML) **34** (2017) 1321–1330. https://doi.org/10.48550/arXiv.1706.04599
30 M. P. Naeini, G. Cooper, and M. Hauskrecht: Proc. AAAI Conf. Artificial Intelligence **29** (2015) 2901–2907. https://doi.org/10.1609/aaai.v29i1.9602

31 J. Nixon, M. W. Dusenberry, L. Zhang, G. Jerfel, and D. Tran: Proc. IEEE/CVF Conf. Computer Vision Pattern Recognition Workshops (CVPR W) **2** (2019) 38–41. https://doi.org/10.48550/arXiv.1904.01685
32 J. Mukhoti and Y. Gal: arXiv preprint arXiv:1811.12709 (2018). https://doi.org/10.48550/arXiv.1811.12709
33 S. Landgraf, M. Hillemann, K. Wursthorn, and M. Ulrich: arXiv preprint arXiv:2307.09947 (2023). https://doi.org/10.48550/arXiv.2307.09947
34 O. Bagoren, M. Micatka, K. A. Skinner, and A. Marburg: arXiv preprint arXiv:2502.09824 (2025). https://doi.org/10.48550/arXiv.2502.09824
35 R. Michelmore, M. Kwiatkowska, and Y. Gal: arXiv preprint arXiv:1811.06817 (2018). https://doi.org/10.48550/arXiv.1811.06817