Cross-modal Depth-privileged Knowledge Distillation for Underwater Salient Object Detection with Only RGB Images

Zhen Feng and Yuxiang Sun

2026 IEEE Canadian Atlantic Ocean Symposium (CAOS) | 979-8-3315-8293-7/26/$31.00 ©2026 IEEE | DOI: 10.1109/CAOS69069.2026.11666192

***Abstract*— Existing studies on RGB image-based underwater salient object detection (SOD) have made satisfactory progress. But using the single RGB modality is prone to degradation under challenging lighting conditions in underwater environments. So, many researchers resort to multi-modal data fusion to improve the detection performance. However, most existing multi-modal methods employ a dual-encoder architecture, which increases computational complexity, imposing high requirements on model deployment in real robots. To provide a solution to this issue, this paper proposes to transfer the capabilities of an RGB-Depth (RGB-D) multi-modal fusion-based network to a lightweight student network that uses only RGB images through knowledge distillation. Specifically, we use an existing RGB-D network as the teacher network, and a lightweight single-RGB-modal network as the student network. The decoder of the student network is designed with a dual-branch structure, with each branch dedicated to learning the feature extraction capabilities of the teacher network for the RGB data and the depth data. Experimental results on a public dataset demonstrate our effectiveness to improve the detection performance of the single-modal student network compared with supervised learning.**

I. Introduction

Underwater scene understanding is important for marine resource exploitation and underwater search and rescue [1]. As a fundamental task in scene understanding, salient object detection (SOD) provides support for target recognition and tracking, and has attracted increasing attention in the computer vision research community [2]. Currently, underwater SOD based on RGB images has been extensively studied [3]. However, most existing methods are prone to be degraded under challenging underwater lighting conditions caused by factors, such as water turbidity and color shifts [4], etc. To address this problem, many researchers have proposed multi-modal fusion methods (e.g., fusing RGB with depth data) to improve the detection performance [5]–[7]. The multi-modal methods leverage the strengths of different modalities to enable complementary feature learning [8]–[10].

However, most existing multi-modal methods employ a dual-encoder parallel architecture [5], [7], [11]–[18]. While this improves detection performance, it increases the amount of model parameters and computational cost, which imposes challenges for underwater robots with limited computational resources. To provide a solution to this issue, this paper proposes a cross-modal knowledge distillation framework for underwater SOD. The framework aims to transfer the capabilities of a multi-modal RGB-depth (RGB-D) fusion network to a single-RGB-modal network, thereby improving the detection performance of the single-modal network and avoiding the additional computational load caused by using the multi-modal network.

Specifically, our method employs a high-performance RGB-D network as the teacher network, and a lightweight RGB-only network as the student network. Through a knowledge distillation strategy, the accurate detection capabilities of the teacher network are transferred to the lightweight student network. In the student network, a dual-branch structure is proposed to separately learn the feature extraction capabilities for the RGB data, as well as the supplementary feature extraction capabilities for the depth data from the teacher network. Experimental results demonstrate that the proposed framework effectively improves the detection performance of the single-RGB-modal network compared with supervised learning, and our student network outperforms classical networks. Our code is open-sourced[1]. The contributions of this paper are summarized as follows:

- We propose a cross-modal distillation framework to transfer the capabilities of an RGB-D fusion network to a single-RGB-modal network, thereby improving the detection performance of the single-modal network.
- We propose a two-branch strategy in the student network to separately learn the feature extraction capabilities of the teacher network for both the RGB and depth data.

The remainder of this paper is organized as follows. Section II reviews the related work. Section III presents the details of our proposed framework. Section IV describes the experimental results. The last section concludes this paper.

II. Related Work

A. Underwater SOD

Many methods have been proposed for underwater SOD through multi-modal data fusion. For example, Zhou et. al [11] introduced the segmentation anything model (SAM) and designed a region-aware attention collaborative learning strategy to improve the performance of RGB-D SOD task. The RGB images are fed into SAM and the RGB-D images are feed into proposed auxiliary network to generate mask prompts for the SAM. Ma et. al [9] designed a multi-modal fusion dataset for underwater SOD with RGB images and polarization images. They designed a cross-modal fusion module based on Mamba [21] for fusing the features of RGB and polarization images. Liu et. al [4] proposed transmission

This work was supported by City University of Hong Kong under Grants 9610675 and 9231601. *(Corresponding author: Yuxiang Sun.)*

Zhen Feng and Yuxiang Sun are with the Department of Mechanical Engineering, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong (e-mail: zhen.feng@cityu.edu.hk; yx.sun@cityu.edu.hk).

[1] https://github.com/lab-sun/CDKD-SOD

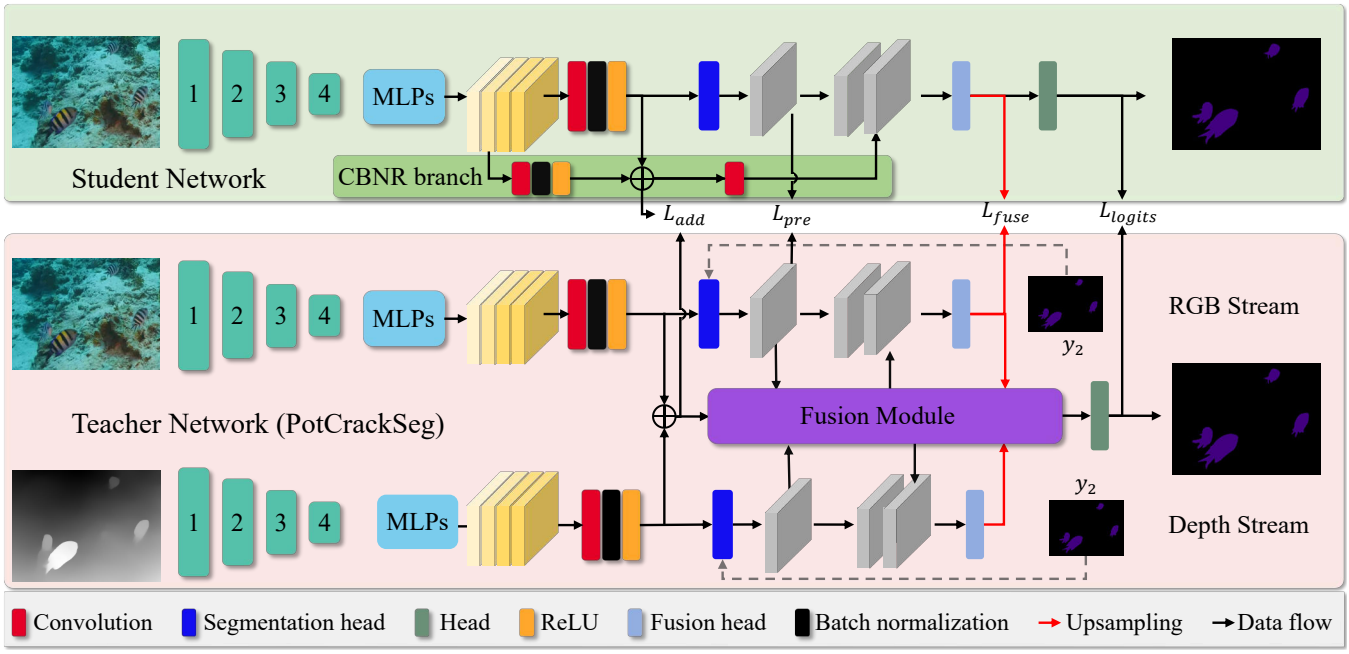

Fig. 1: The overview of our framework. The teacher network is PotCrackSeg [19]. The main structure of the student network draws on the RGB stream of PotCrackSeg. The student network uses SegFormer-B0 [20] for feature extraction, whereas PotCrackSeg uses SegFormer-B2 [20]. The figure is best viewed in color.

maps from RGB images to improve the performance of underwater SOD. They treated the transmission maps as modality data that are used to fuse with RGB images. Zha et al. [10] proposed a heterogeneous expert strategy to improve the performance of underwater SOD based on the frequency of feature maps.

Although multi-modal fusion is able to improve the performance, the dual-encoder architecture used in the multi-modal methods increases the computational load, placing high demands on their practical deployments.

B. Cross-Modal Knowledge Distillation

Knowledge distillation is commonly used to reduce the amount of model parameters while maintaining model performance [22], [23]. In recent years, with the development of multi-modal research, cross-modal distillation has been widely employed to transfer knowledge from multi-modal networks to single-modal networks, thereby improving the performance of the single-modal networks [24].

For example, Feng et al. [25] designed a Knowledge distillation framework to transfer the edge detection capability from an RGB-thermal network to a thermal-only network. Liu et al. [24] proposed a frequency-decoupled strategy to transfer knowledge from cross-modal data. Feng et al. [26] designed a distillation framework to transform knowledge from a higher-resolution teacher network to a lower-resolution student network. Ren et al. [27] adopted a cross-modal knowledge distillation strategy to improve depth estimation by transfer knowledge from a multi-view teacher network to an event-based student network.

III. The Proposed Framework

A. The Framework

The overview of our cross-modal distillation framework is shown in Fig. 1. As we can see, it contains an RGB-D fusion network as the teacher network and an RGB-only network as the student network.

We directly adopt the existing multi-modal fusion network PotCrackSeg [19] as the teacher network. PotCrackSeg fuses features from two modalities using a dual-complementary strategy. Specifically, it first extracts features from RGB images and depth images using two independent Segformer-B2 encoders [20]. Subsequently, semantic predictions are made for the features of each modality (which can be referred to as the original semantic information). Then, it calculates the difference between the ground truth and the semantic predictions for each modality, and extracts this difference from the other modality. Finally, the difference is fused with the original semantic features of each modality to improve the segmentation accuracy.

For the student network, our idea is to align with the structure of the teacher network. We use a light-weight SegFormer-B0 [20] as the encoder for the student network. The decoder of the student network is designed based on the RGB decoder of the teacher network. In addition, since the student network lacks the depth branch, we add an additional convolution-batch-normalization-ReLU (CBNR) branch during the decoding stage as the depth branch. We expect the features extracted by this added branch to capture pseudo-depth features from the RGB images. We fuse the results of the added CBNR branch with those of the original CBNR branch in the student decoder, ensuring consistency with

TABLE I: Ablation study results (%) on loss functions. The symbols – and ✓respectively indicate that the loss is not used and the loss is used. The best results are highlighted in bold font.

Loss	Loss components				Background		Object		mIoU	mF1
	$\mathcal{L}_{add}$	$\mathcal{L}_{pre}$	$\mathcal{L}_{logits}$	$\mathcal{L}_{seg}$	IoU	F1	IoU	F1		
Supervised	–	–	–	✓	95.63	97.77	74.20	85.19	84.92	91.48
A	–	–	✓	✓	96.10	98.01	78.93	88.22	87.51	93.12
B	–	✓	–	✓	96.49	98.21	79.95	88.86	88.22	93.54
C	✓	–	–	✓	**96.65**	**98.29**	80.44	89.16	88.54	93.73
D	–	✓	✓	✓	96.51	98.22	80.20	89.01	88.36	93.62
E	✓	✓	–	✓	96.47	98.20	80.36	89.11	88.42	93.66
F	✓	✓	✓	✓	96.57	98.26	**80.55**	**89.22**	**88.56**	**93.74**

TABLE II: Comparative results (%) on the USOD10K dataset. ms refers to the time required to process an image in milliseconds. FPS refers to frames per second. The best results are highlighted in bold font.

Network	Background				Objects				mPre	mAcc	mIoU	mF1	ms	FPS
	Pre	Acc	IoU	F1	Pre	Acc	IoU	F1						
DeepLabV3+	95.10	92.16	87.99	93.61	59.74	71.00	48.02	64.89	77.42	81.58	68.00	79.25	14.45	69.20
UNet	94.65	97.29	92.21	95.95	80.05	66.39	56.96	72.58	87.35	81.84	74.59	84.27	**8.29**	**120.63**
Ours	**98.09**	**98.43**	**96.57**	**98.26**	**90.20**	**88.27**	**80.55**	**89.22**	**94.14**	**93.35**	**88.56**	**93.74**	15.04	66.51

the fusion results in the teacher network. Subsequently, the fused results are first convolved and then concatenated with semantic-level features.

B. The Losses

To enable the student network to better learn from the teacher network, we adopt four distillation-related losses, as well as one cross-entropy loss supervised by ground truth. Specifically, we first let the student network learn semantic features from the RGB stream of the teacher network. The purpose of this step is to ensure that the student network can learn, as much as possible, the capability of the teacher network to extract features from RGB images. We denote this loss function as $\mathcal{L}_{\text{pre}}$.

Then, we expect the added CBNR branch to produce the pseudo-depth features like the real depth features in the teacher network. So, we enforce the fusion results of the two branches to be consistent with the fusion results in the teacher network. We use the cosine similarity to supervise the two fusion results, and denote this loss as $\mathcal{L}_{\text{add}}$, which is calculated as follows:

$$\mathcal{L}_{\text{add}} = 1 - \frac{1}{HW}\sum_{i=1}^{HW}\frac{\mathbf{f}_s^{(i)}\cdot\mathbf{f}_t^{(i)}}{\|\mathbf{f}_s^{(i)}\|_2\|\mathbf{f}_t^{(i)}\|_2}, \tag{1}$$

where H, W denote the height and width of the feature maps. $\mathbf{f}_s^{(i)}$ and $\mathbf{f}_t^{(i)}$ represent the C-dimensional feature vectors of the student and teacher at the i-th spatial location, respectively.

Finally, we use the output of the teacher network and the ground truth to supervise the output of the student network. These two loss functions are defined as $\mathcal{L}_{\text{logits}}$ and $\mathcal{L}_{\text{seg}}$. The calculations of $\mathcal{L}_{\text{pre}}$ and $\mathcal{L}_{\text{logits}}$ use the temperature-weighted cross-entropy loss from [25]. The loss function of the whole framework is obtained by adding up the 5 loss functions.

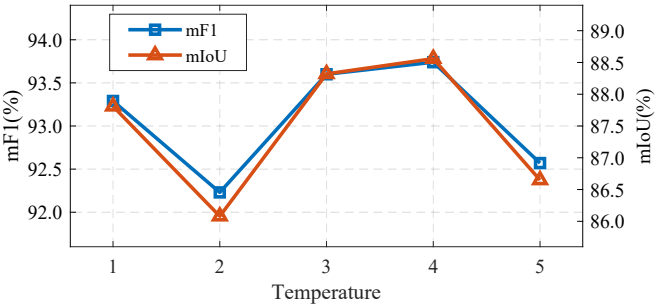

Fig. 2: The results of ablation study on distillation temperature. The figure is best viewed in color.

IV. Experimental Results and Discussions

A. Experimental Setup

Our framework is implemented with PyTorch. It is trained and tested on a PC with an NVIDIA A5000 (24 GB RAM) graphics card. The encoder parameters of our framework are initialized with the pre-trained weight of SegFormer. The learning rate is set with a warmup strategy, an initial value of 6×10^{-5}, and the polynomial decay strategy [20]. The student network, all the comparative networks and the variations are trained with 5 epochs on the USOD10K dataset [30]. The teacher network is trained to achieve the optimal value on the USOD10K dataset.

The USOD10K dataset contains $10,255$ pairs of RGB images and estimated depth images, covering 12 different underwater scenes. The images are with the resolution of 640×480, which are split into three sets: training ($7,178$ pairs of images), validation ($2,051$ pairs of images), and test ($1,026$ pairs of images).

We treat the background as a class that is needed to detect during the training and testing process. For quantitative evaluation, we employ the metrics Precision (Pre), Accuracy

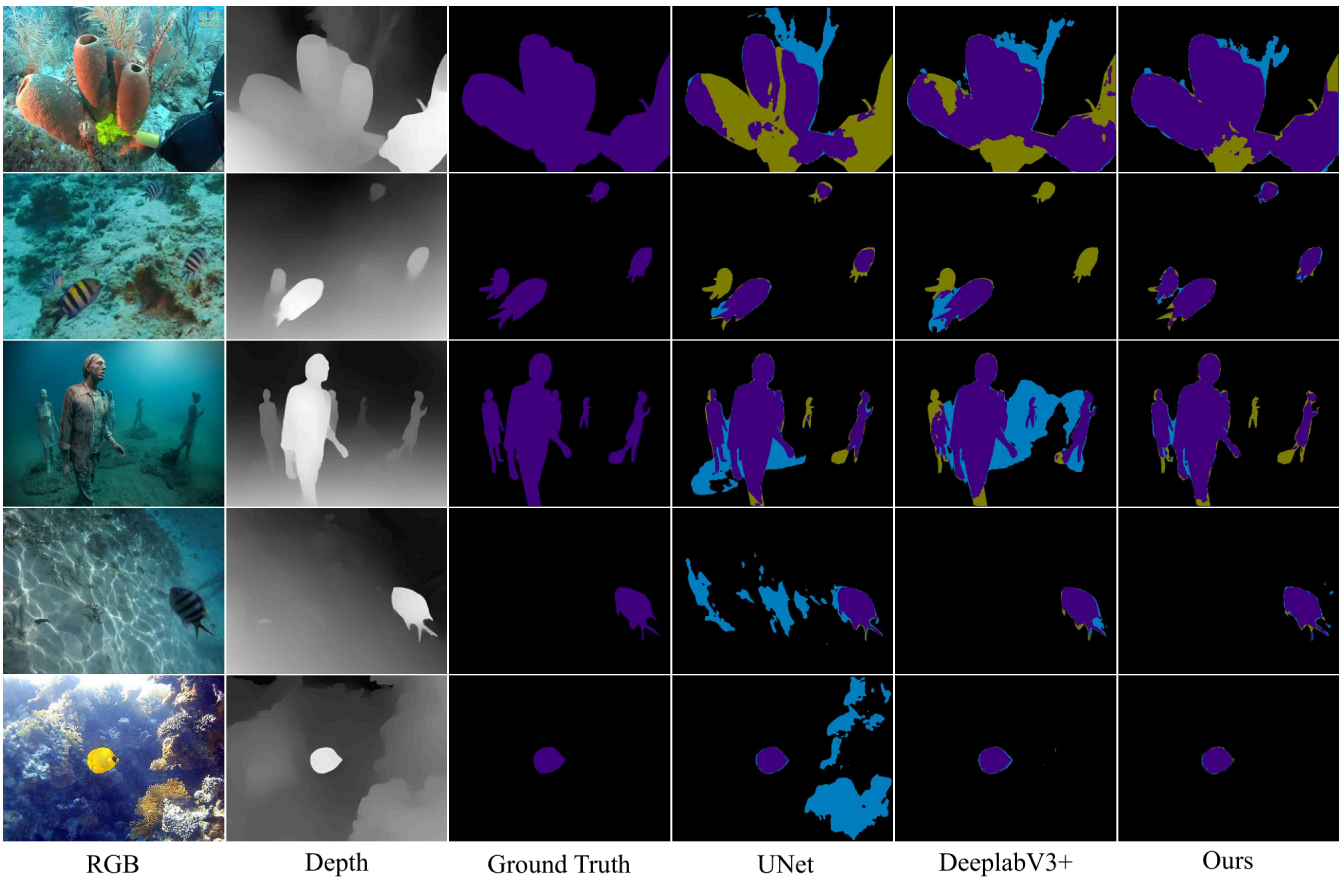

Fig. 3: Sample qualitative demonstrations for the detection of salient objects on the USOD10K dataset. The first column to the last column are respectively the RGB images, depth images, ground truth, the results of UNet [28], the results of DeepLabV3+ [29], and the results of our student network. The colors ■, ■, and ■ represent ground truth, false positives, and false negatives, respectively. The figure is best viewed in color.

(Acc), Intersection-over-Union (IoU), F1-score (F1), as well as their mean counterparts (i.e., mPre, mAcc, mIoU, and mF1) across all the classes including the background [31].

B. Ablation Studies

1) Ablation on Distillation Temperature: We design several variants to determine the optimal distillation temperature. Specifically, we set the distillation temperature to the values ranging from 1 to 5. The results of the variants are displayed in Fig. 2. As shown in Fig. 2, the student network achieves the optimal value when the distillation temperature is set to 4.

2) Ablation on Loss Functions: We design several variants to demonstrate the role of each loss function. Specifically, we first set up a supervised-learning variant, in which only $\mathcal{L}_{\text{seg}}$ is used during training. Then, we utilize a cross-modal knowledge distillation framework to remove each loss function from the set of all loss functions, to verify the role of each loss function. The loss functions used in each variant and their results are displayed in Tab. I.

The experimental results show that when training the student network using the supervised-learning method, its performance is inferior to that of the student network trained using our proposed cross-modal knowledge distillation framework. This demonstrates the effectiveness of the proposed framework. Furthermore, the other experimental results indicate that the combined use of the four loss functions enables more effective knowledge transfer from the teacher network to the student network.

C. Comparative Study

1) Quantitative Results: We compare our student network with the classic networks: UNet [28] and DeeplabV3+ [29]. The results of the networks are displayed in Tab. II. The results demonstrate the superiority of our student network that is trained with our proposed framework.

We also test the inference speed of the classical single-modal networks and our student network on a PC with an NVIDIA RTX 4060 Ti graphics card. As shown in Tab. II, although our student network just achieves inference speeds comparable to DeeplabV3+, it significantly outperforms the network in terms of SOD accuracy.

2) Qualitative Demonstrations: Fig. 3 shows sample qualitative demonstration results of the student network trained with our proposed cross-modal knowledge distillation framework, as well as the qualitative demonstration results of the

compared networks. As we can see, the student network achieves generally better performance in the edge regions of the salient objects.

V. Conclusions and Future Work

This paper proposed a cross-modal knowledge distillation framework for underwater SOD. It is designed to transfer the capabilities of an RGB-D fusion-based teacher network to a single-RGB-modal student network via knowledge distillation. To better learn the capabilities of the teacher network, we proposed a dual-branch learning strategy to learn the feature extraction capabilities of the RGB data and the depth data separately. The experimental results demonstrate that our proposed framework outperforms supervised-learning method and achieves superior performance compared to the comparative networks. In the future, we would like to further introduce a distillation strategy for the edge detection capability to improve the detection accuracy in the edge regions for the student network.

References

[1] L. Chen, Y. Huang, J. Dong, Q. Xu, S. Kwong, H. Lu, H. Lu, and C. Li, "Underwater optical object detection in the era of artificial intelligence: current, challenge, and future," *ACM Computing Surveys*, vol. 58, no. 3, pp. 1–34, 2025.

[2] Z. Su, L. Liu, M. Müller, J. Zhang, D. Wofk, M.-M. Cheng, and M. Pietikäinen, "Rapid salient object detection with difference convolutional neural networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[3] Y. M. Alsakar, N. A. Sakr, S. El-Sappagh, T. Abuhmed, and M. Elmogy, "Underwater image restoration and enhancement: a comprehensive review of recent trends, challenges, and applications," *The Visual Computer*, vol. 41, no. 6, pp. 3735–3783, 2025.

[4] Y. Liu, X. Zhang, J. Zhu, B. Ma, Y. Duan, and P. Tan, "Hdanet: Enhancing underwater salient object detection with physics-inspired multimodal joint learning," *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[5] T. Wang, J. Lu, B. Wan, R. Lu, Y. Sun, D. Wu, Y. Liu, and C. Yan, "Feature boosting and scale-aware network with multi-modal information for underwater salient object detection," *Engineering Applications of Artificial Intelligence*, vol. 174, p. 114489, 2026.

[6] S. Zheng, X. Zhou, L. Bao, X. Hu, and J. Zhang, "Depth-guided cross-modal fusion network for underwater salient instance segmentation," *Symmetry*, vol. 18, no. 5, p. 799, 2026.

[7] F. Guo, P. Ren, and C. Luo, "Utnet: event-rgb multimodal fusion model for underwater transparent organism detection," *Intelligent Marine Technology and Systems*, vol. 3, no. 1, p. 18, 2025.

[8] W. Xu, C. Wang, D. Liang, Z. Zhao, X. Jiang, P. Zhang, and X. Bai, "Nautilus: A large multimodal model for underwater scene understanding," *Advances in Neural Information Processing Systems*, vol. 38, pp. 25 079–25 105, 2026.

[9] Q. Ma, X. Li, B. Li, Z. Zhu, J. Wu, F. Huang, and H. Hu, "Stamf: synergistic transformer and mamba fusion network for rgb-polarization based underwater salient object detection," *Information Fusion*, vol. 122, p. 103182, 2025.

[10] M. Zha, G. Wang, Y. Pei, T. Li, X. Tang, C. Li, Y. Yang, and H. T. Shen, "Heterogeneous experts and hierarchical perception for underwater salient object detection," *IEEE Transactions on Image Processing*, 2025.

[11] W. Zhou, B. Tang, X. Dong, and F. Qiang, "Prompt then refine: Prompt-free sam-enhanced collaborative learning network for detecting salient objects in underwater images," *IEEE Transactions on Neural Networks and Learning Systems*, 2026.

[12] W. Zhou, B. Tang, R. Cong, and Q. Jiang, "Turbidity–similarity decoupling: Feature-consistent mutual learning for underwater salient object detection," *IEEE Transactions on Image Processing*, 2026.

[13] H. Li, H. K. Chu, and Y. Sun, "Spatial balancing for rgb-thermal semantic segmentation in autonomous driving: A study from analysis to improvement," *IEEE Transactions on Robotics*, vol. 42, pp. 1840–1855, 2026.

[14] Y. Wu, W. Wang, C. Lin, M. Hou, and M. Liu, "Towards multimodal underwater object detection: A bidirectional feature recomposition network and visual-sonar dataset," *Expert Systems with Applications*, p. 131710, 2026.

[15] Z. Feng, Y. Guo, R. Fan, and Y. Sun, "Mmfseg: Multi-structure multi-feature fusion for segmentation of road potholes," *IEEE Transactions on Automation Science and Engineering*, 2025.

[16] Z. Feng, Y. Guo, D. Navarro-Alarcon, Y. Lyu, and Y. Sun, "Inconseg: Residual-guided fusion with inconsistent multi-modal data for negative and positive road obstacles segmentation," *IEEE Robotics and Automation Letters*, vol. 8, no. 8, pp. 4871–4878, 2023.

[17] Y. Sun, W. Zuo, P. Yun, H. Wang, and M. Liu, "Fuseseg: Semantic segmentation of urban scenes based on rgb and thermal data fusion," *IEEE Transactions on Automation Science and Engineering*, vol. 18, no. 3, pp. 1000–1011, 2021.

[18] Y. Sun, W. Zuo, and M. Liu, "Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes," *IEEE Robotics and Automation Letters*, vol. 4, no. 3, pp. 2576–2583, 2019.

[19] Z. Feng, Y. Guo, and Y. Sun, "Segmentation of road negative obstacles based on dual semantic-feature complementary fusion for autonomous driving," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 4, pp. 4687–4697, 2024.

[20] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in neural information processing systems*, vol. 34, pp. 12 077–12 090, 2021.

[21] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[22] Y. Himeur, N. Aburaed, O. Elharrouss, I. Varlamis, S. Atalla, W. Mansoor, and H. Al-Ahmad, "Applications of knowledge distillation in remote sensing: A survey," *Information Fusion*, vol. 115, p. 102742, 2025.

[23] C. Yang, Y. Zhu, W. Lu, Y. Wang, Q. Chen, C. Gao, B. Yan, and Y. Chen, "Survey on knowledge distillation for large language models: methods, evaluation, and application," *ACM Transactions on Intelligent Systems and Technology*, vol. 16, no. 6, pp. 1–27, 2025.

[24] J. Liu, Y. Zhang, T. Huang, W. Xu, and R. Yang, "Distilling cross-modal knowledge via feature disentanglement," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, no. 28, 2026, pp. 23 739–23 747.

[25] Z. Feng, Y. Guo, and Y. Sun, "Cekd: Cross-modal edge-privileged knowledge distillation for semantic scene understanding using only thermal images," *IEEE Robotics and Automation Letters*, vol. 8, no. 4, pp. 2205–2212, 2023.

[26] Z. Feng, Y. Guo, and Y. Sun, "Cpkd: Channel and position-wise knowledge distillation for segmentation of road negative obstacles," in *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2023, pp. 3110–3115.

[27] Y. Ren, J. Zhu, K. Chen, Z. Li, J. Ou, Z. Cao, T. Hua, P. Shi, Y. Fu, W. Zhao *et al.*, "Eventvggt: Exploring cross-modal distillation for consistent event-based depth estimation," *arXiv preprint arXiv:2603.09385*, 2026.

[28] T. Falk, D. Mai, R. Bensch, Ö. Çiçek, A. Abdulkadir, Y. Marrakchi, A. Böhm, J. Deubner, Z. Jäckel, K. Seiwald *et al.*, "U-net: deep learning for cell counting, detection, and morphometry," *Nature methods*, vol. 16, no. 1, pp. 67–70, 2019.

[29] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[30] L. Hong, X. Wang, G. Zhang, and M. Zhao, "Usod10k: a new benchmark dataset for underwater salient object detection," *IEEE transactions on image processing*, vol. 34, pp. 1602–1615, 2023.

[31] Z. Feng, Y. Guo, Q. Liang, M. U. M. Bhutta, H. Wang, M. Liu, and Y. Sun, "Mafnet: Segmentation of road potholes with multimodal attention fusion network for autonomous vehicles," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–12, 2022.