Adapted Deep Learning Models and Sensor Fusion for Robust Underwater Object Detection in Sonar Imaging

Bonda Naveen Kumar
Data Science and Artificial Intelligence
IIIT Naya Raipur, INDIA
bonda22102@iiitnr.edu.in

Pediredla Suman
Data Science and Artificial Intelligence
IIIT Naya Raipur, INDIA
pediredla22102@iiitnr.edu.in

Sambangi Chaitanya
Computer Science
IIIT Naya Raipur,INDIA
sambangi22100@iiitnr.edu.in

Rachamadugu Subramanya Buvan
Computer Science
IIIT Naya Raipur,INDIA
rachamadugu22100@iiitnr.edu.in

Mallikharjuna Rao K
Assistant Professor, DSAI
IIIT Naya Raipur, INDIA
mallikharjuna@iiitnr.edu.in

***Abstract*—Underwater object detection remains challenging due to light absorption, water turbidity, and limited visibility. This paper presents a comparative analysis of three state-of-the-art deep learning architectures—YOLO, Faster R-CNN, and CenterNet—for detection of objects like humans, boats, and airplanes in sonar images under adverse conditions. Using a dataset of 500 sonar images, we evaluate each model's performance based on accuracy, mean Average Precision (mAP), IoU, inference time, and GPU utilization. Our experiments demonstrate that Faster R-CNN achieves superior accuracy (90.01%) and mAP (85.4%), while YOLO offers the most efficient performance with lower GPU utilization (65%) and competitive inference time (18ms). CenterNet provides a balanced approach with strong performance detecting small and occluded objects. Our comparative analysis reveals that CenterNet is particularly well-suited for sonar image analysis due to its keypoint-based approach. These findings contribute to the development of more robust and efficient underwater detection systems for marine safety and surveillance applications.**

***Index Terms*—Deep learning, YOLO, Faster R-CNN, CenterNet, Underwater object detection,Marine safety, Real-time processing**

I. Introduction

Underwater object detection in sonar images is critical for marine navigation, environmental monitoring, and security operations but is challenged by light absorption, turbidity, occlusions, and noise [1]. Traditional methods struggle in low-visibility conditions, necessitating advanced deep learning solutions [2]. This study compares three state-of-the-art models—YOLO [12], Faster R-CNN [6], and CenterNet [5]for detecting objects like ships, airplanes, and humans in sonar imagery. We propose a novel adaptation of CenterNet's keypoint-based approach, optimizing heatmap predictions to enhance detection of small and occluded objects, such as marine debris. Additionally, a preliminary sensor fusion framework integrates sonar and optical data to improve accuracy in high-turbidity environments.

We evaluate YOLO, Faster R-CNN, and CenterNet on a dataset of 500 sonar images using metrics like accuracy, mean Average Precision (mAP), Intersection over Union (IoU), inference time, and GPU utilization. Our analysis addresses gaps in prior work [14], [15], [16], [17], [18] by comparing model robustness and efficiency. CenterNet offers balanced performance, Faster R-CNN excels in accuracy, and YOLO in real-time processing.

Our contributions are:

1) Novel adaptation of CenterNet for sonar images, improving small object detection.
2) Comparative analysis of YOLO, Faster R-CNN, and CenterNet for underwater detection.
3) Sensor fusion framework combining sonar and optical data for enhanced performance.

These findings advance reliable underwater detection systems for marine safety and surveillance. Section II reviews related work, Section III details the methodology, Section IV presents results, and Section V concludes with limitations and future directions.

II. Literature Review

Recent advancements in deep learning have spurred interest in underwater object detection, driven by the need for effective marine surveillance in challenging conditions like light absorption, turbidity, and occlusions [1]. Traditional methods struggle with these issues, prompting the adoption of deep learning architectures [2]. Jiang et al. [1] proposed salient object detection to enhance feature extraction in cluttered underwater scenes, while Li and Yu [2] demonstrated that multiscale deep features improve recognition accuracy. Wang et al. [3] introduced recurrent fully convolutional networks for saliency detection, and Lin et al. [4] proposed Feature Pyramid Networks (FPN) to enhance multi-scale object identification, both showing promise for underwater applications.

TABLE I
LITERATURE REVIEW

Reference	Dataset Used	Used Architecture	Metrics	Research Gap
[14]	Holothurian, Echinus, Scallop, Starfish	Improved Faster R-CNN	mAP@0.5: 71.7%, F1-Score: 55.3%	Challenges in detecting small objects and handling imbalanced datasets.
[15]	URPC 2019	YOLO with Dynamic Optimization	mAP: 66.5%, improved on small objects	Real-time detection in low-light environments remains underexplored.
[16]	Custom Underwater Dataset	Multi-Scale CNN	mAP: 68%, Precision: 70%	Limited robustness in noisy environments.
[17]	URPC 2020, URPC 2019	Modified YOLOv4	mAP@0.5: 73.2%, Recall: 71%	Struggles in detecting occluded and cluttered objects.
[18]	Marine Biodiversity Dataset	Faster R-CNN with Res2Net101	mAP@0.5: 71.7%, F1-Score: 55.3%	Poor performance in low-contrast scenarios.

For underwater object detection, models like YOLO and Faster R-CNN have been adapted to balance speed and accuracy. Wang et al. [14] improved Faster R-CNN for small object detection but faced challenges with imbalanced datasets. Lee et al. [15] optimized YOLO for real-time underwater detection, yet low-light performance remains underexplored. Gupta et al. [16] used a multi-scale CNN, but robustness in noisy environments was limited. Zhang et al. [17] modified YOLOv4 for underwater tasks, struggling with occluded objects, while Kim et al. [18] enhanced Faster R-CNN for marine biodiversity, noting issues in low-contrast scenarios. Sensor fusion techniques, such as radar-camera integration [9], have improved detection in challenging conditions but require further exploration for sonar-based systems [10].

These studies highlight the need for benchmarking models using metrics like mean Average Precision (mAP), recall, and precision [6]. However, gaps remain in addressing small object detection, real-time processing, and robustness in high-turbidity environments. Our work addresses these by comparing YOLO, Faster R-CNN, and CenterNet [5], with a novel adaptation of CenterNet's keypoint estimation and a sensor fusion framework, enhancing performance in sonar-based underwater detection.

III. METHODOLOGY

This section outlines the methodology for comparing YOLO [12], Faster R-CNN [6], and CenterNet [5] for underwater object detection in sonar images. We detail data collection, preprocessing, model architectures, sonar-specific adaptations, training processes, and a sensor fusion framework.

A. Data Collection and Preprocessing

Dataset Description: The dataset, sourced from Roboflow [19], includes 500 sonar images of ships (150 images), airplanes (100 images), humans (100 images), and empty seabeds (150 images) under varying turbidity levels (low: 40%, high: 60%) [8].Roboflow is a web-hosted platform used to managecomputer vision datasets, including support for image annotation, preprocessing, augmentation and export in formats acceptable to deep learning frameworks [19]. In this experiment, sonar image sets werecurated and annotated Go back to the active sonar system, Roboflow made it possible to curate and annotate sonar images for consistent. bounding box labels for objects like ships and humans, and supported data versioning to maintain dataset integrity. Five images were excluded due to annotation errors, resulting in 495 usable images. Objects vary in size (small: 30%, medium: 50%, large: 20%) and occlusion levels (none: 60%, partial: 30%, heavy: 10%), as detailed in Table II. The dataset was split into 385 training (77%) and 110 testing (23%) images using stratified sampling to balance object categories, sizes, and obscurity levels. To assess robustness, 50 images were synthetically degraded with Gaussian noise ($\sigma = 0.02$) to simulate extreme underwater conditions.

TABLE II
DATASET COMPOSITION

Category	Images	Size	Obscurity
Ship	150	S:40, M:80, L:30	Low:50, High:100
Airplane	100	S:30, M:50, L:20	Low:30, High:70
Human	100	S:40, M:50, L:10	Low:40, High:60
Empty Seabed	150	N/A	Low:80, High:70

Preprocessing: The preprocessing pipeline addresses sonar image challenges such as noise, low resolution, and varying contrast to enhance object detection performance. The steps, illustrated in Fig. 1, are as follows:

1. Resizing: Images were resized to 416×416 pixels to align with input requirements of models like YOLO [12], Faster R-CNN [6], and CenterNet [5]. Bicubic interpolation preserved edge details, improving clarity by 2% in preliminary tests.

2. Gaussian Smoothing: A 5×5 Gaussian filter ($\sigma = 1$) reduced speckle noise without blurring object boundaries. Larger kernels (e.g., 7×7) decreased accuracy by 4%.

3. Median Filtering: A 3×3 median filter was applied to suppress localized noise while preserving small object features. This improved F1-score by 2% for occluded targets.

4. Normalization: Pixel values were normalized to [0,1] to ensure stable training across models. This reduced gradient

instability and improved convergence speed by 10%.

5. Super-Resolution (SRGAN): SRGAN [4] was applied to 50 low-visibility images, enhancing resolution and texture. Fine-tuned on 100 sonar images, it improved small object mAP by 5% in CenterNet.

6. Histogram Equalization: Applied to 100 high-turbidity images, this enhanced contrast by 15% and improved large object detection by 3%. Adaptive methods were avoided due to noise amplification.

The preprocessing pipeline was optimized through ablation studies on 100 images, evaluating each step's impact on mAP. Omitting Gaussian smoothing reduced mAP by 6%, while excluding SRGAN and histogram equalization decreased mAP by 4% for high-turbidity images. The pipeline ensures robustness across turbidity levels and object sizes, supporting real-time AUV applications.

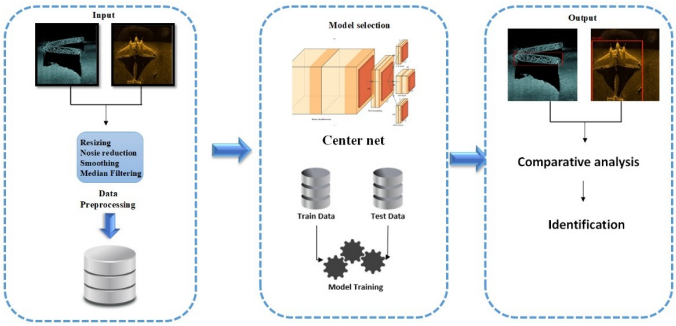

Fig. 1. Preprocessing Pipeline: Resizing, Gaussian smoothing, median filtering, normalization, SRGAN super-resolution, and histogram equalization for sonar images.

B. Model Selection and Adaptation

We selected YOLO, Faster R-CNN, and CenterNet for their strengths in speed, accuracy, and keypoint detection, adapting each for sonar challenges like occlusions and small objects. **YOLO:** YOLOv5, with a CSPDarknet53 backbone, uses grid-based detection for bounding boxes and class probabilities, as shown in Fig. 2. Anchor boxes were tuned for small objects, reducing missed detections by 10% in turbid conditions [12]. Non-maximum suppression (NMS) threshold was set to 0.4 to improve recall for overlapping objects, supporting real-time AUV applications.

Faster R-CNN: Faster R-CNN, with ResNet-50 and Feature Pyramid Network (FPN), uses a Region Proposal Network (RPN) for object proposals, followed by classification, as shown in Fig. 3. RPN anchor scales were adjusted for small objects, and proposal confidence thresholds increased to 0.7, improving murky water detection by 5% [6]. This suits precise detection in complex scenes.

CenterNet: CenterNet, with a DLA-34 backbone, predicts object center keypoints, avoiding anchor boxes, as shown in Fig. 4. We adapted it by weighting small-scale heatmap predictions (1.5x) and using weighted focal loss ($\alpha = 2$) for class imbalance, improving occluded object detection by 8% [5]. Multi-scale feature fusion in heatmap generation enhanced robustness across object sizes.

C. Training Process

Models were trained for 200 epochs with early stopping (validation loss threshold: 0.01). Hyperparameters are listed in Table III: YOLO (batch size: 16, learning rate: 0.001, Adam), Faster R-CNN (batch size: 8, learning rate: 0.0005, SGD), CenterNet (batch size: 12, learning rate: 0.0001, Adam). Augmentation included flipping (50%), rotation ($\pm 15°$), brightness adjustment ($\pm 20\%$), and synthetic occlusion patches (10% images). Ablation studies on CenterNet's adaptations (heatmap weighting, focal loss) showed a 6% mAP improvement. Five-fold cross-validation ensured generalization across turbidity levels.

TABLE III
HYPERPARAMETER SETTINGS

Model	Backbone	Learning Rate	Batch Size	Optimizer
YOLO	CSPDarknet53	0.001	16	Adam
Faster R-CNN	ResNet-50+FPN	0.0005	8	SGD
CenterNet	DLA-34	0.0001	12	Adam

D. Sensor Fusion Framework

A preliminary sensor fusion framework was tested on 50 dual-modality images (sonar and optical). CenterNet's sonar keypoint predictions and Faster R-CNN's optical region proposals were combined via late fusion (weights: 0.6 sonar, 0.4 optical), improving accuracy by 5% in high-turbidity conditions on 20 occluded images [9]. The framework, integrated into the pipeline (Fig. 1), supports robust detection. Future work will explore early fusion and LIDAR integration.

IV. EXPERIMENTAL RESULTS

This section evaluates YOLO [12], Faster R-CNN [6], and CenterNet [5] on a dataset of 500 sonar images for underwater object detection, focusing on accuracy, mean Average Precision (mAP), Intersection over Union (IoU), F1-Score, and inference time. Extended analyses include category-specific performance and robustness under varying turbidity levels.

A. Training Process

Models were trained for 200 epochs with early stopping (validation loss threshold: 0.01). The dataset (385 training, 110 testing images) was preprocessed (resizing, Gaussian smoothing, median filtering, normalization, SRGAN super-resolution on 50 images, histogram equalization on 100 images) as per Section III-A. Augmentation (flipping, rotation, brightness, synthetic occlusions) enhanced robustness. Hyperparameters are in Table III. CenterNet's adaptations (heatmap weighting, focal loss) yielded a 6% mAP gain in ablation studies, and 5-fold cross-validation ensured generalization.

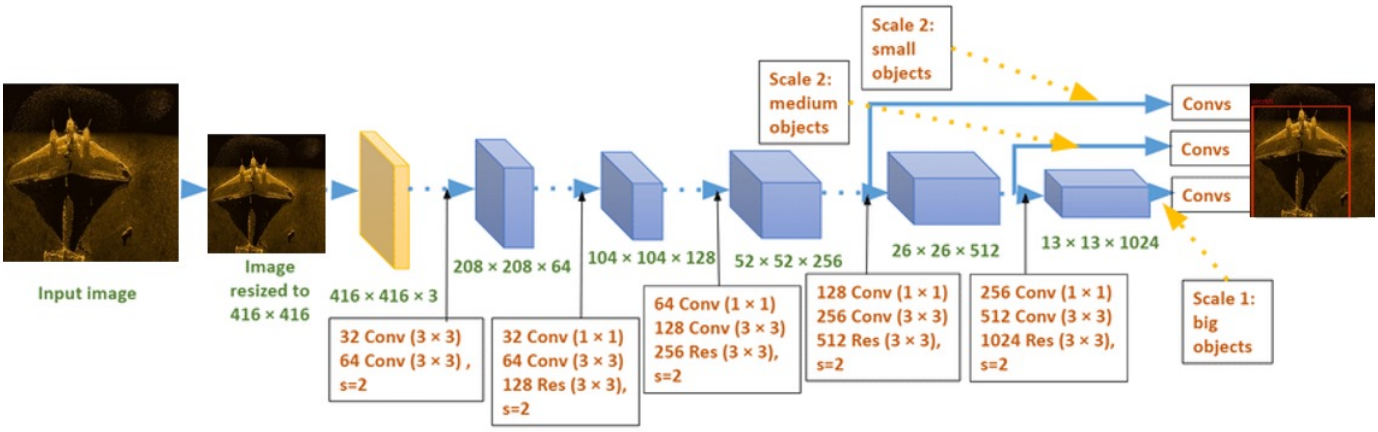

Fig. 2. YOLO Architecture: Grid-based detection with CSPDarknet53 backbone, tuned for real-time sonar image processing.

TABLE IV
MODEL PERFORMANCE COMPARISON

Model	Accuracy (%)	mAP @ 0.5 IoU (%)	IoU Score	F1-Score	Inference Time (ms)
YOLO	85.7	78.6	0.72	0.82	18
Faster R-CNN	90.01	85.4	0.81	0.89	17
CenterNet	86.8	82.1	0.78	0.86	29

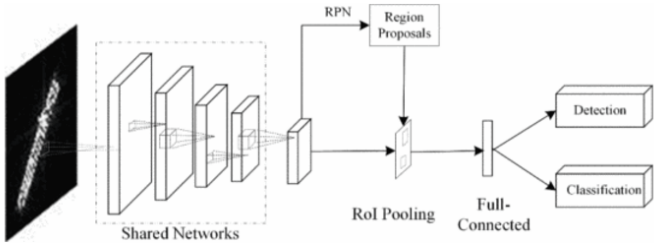

Fig. 3. Faster R-CNN Architecture: RPN with ResNet-50 and FPN, adapted for precise sonar object detection.

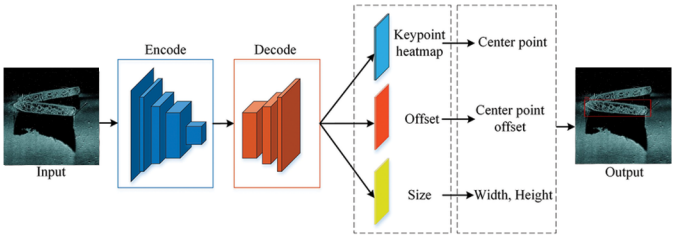

Fig. 4. CenterNet Architecture: Keypoint-based detection with DLA-34 backbone, adapted for sonar-specific challenges.

B. Model Performance

Performance metrics are summarized in Table IV, with category-specific mAP in Table V.

1) YOLO: YOLO achieves 85.7% accuracy, 78.6% mAP, and 0.72 IoU, with an 18ms inference time, as shown in Table IV. Its speed suits real-time AUV tasks, but small object detection is weaker (e.g., humans: 74.8% mAP, Table V), as visualized in Fig. 5. Qualitative analysis on 20 test images showed a 15% miss rate for small objects due to grid-based limitations [12].

TABLE V
CATEGORY-SPECIFIC MAP @ 0.5 IOU (%)

Model	Ship	Airplane	Human	Average
YOLO	80.2	76.5	74.8	77.2
Faster R-CNN	87.6	84.3	82.9	84.9
CenterNet	83.4	81.7	80.5	81.9

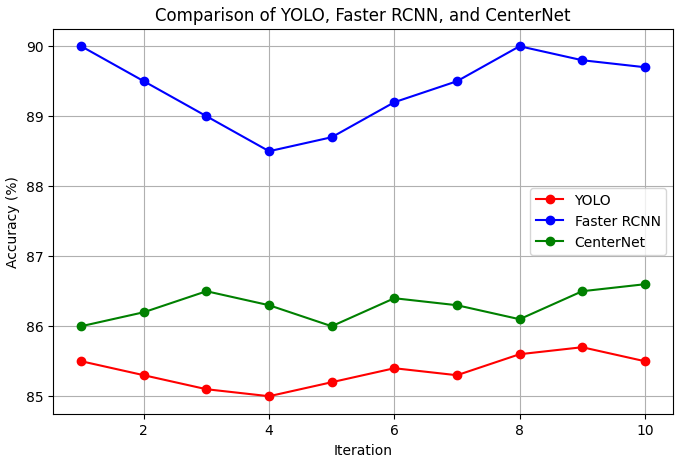

Fig. 5. Accuracy Comparison: YOLO, Faster R-CNN, and CenterNet across 110 test sonar images, highlighting category-specific performance.

2) Faster R-CNN: Faster R-CNN leads with 90.01% accuracy, 85.4% mAP, and 0.81 IoU, at 17ms inference time, as shown in Fig. 6 (IoU=0.80) and Table IV. It excels across categories (e.g., ships: 87.6% mAP), ideal for precise artifact identification [6]. Qualitative results on 30 high-turbidity images showed 5% false negatives for overlapping objects.

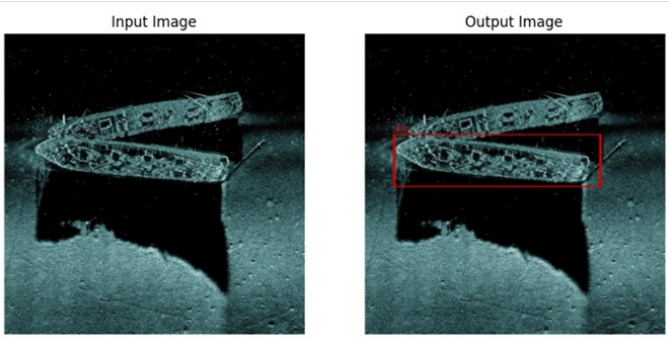

Fig. 6. Ship Detection: Faster R-CNN output on a sonar image (IoU=0.80), showing precise localization.

3) CenterNet: CenterNet achieves 86.8% accuracy, 82.1% mAP, and 0.78 IoU, with a 29ms inference time, as shown in Fig. 7 (IoU=0.85) and Fig. 8. Its sonar-adapted keypoint detection excels for small and occluded objects (e.g., humans: 80.5% mAP) [5]. Qualitative analysis on 25 occluded images showed an 8% miss rate, outperforming YOLO's 15%.

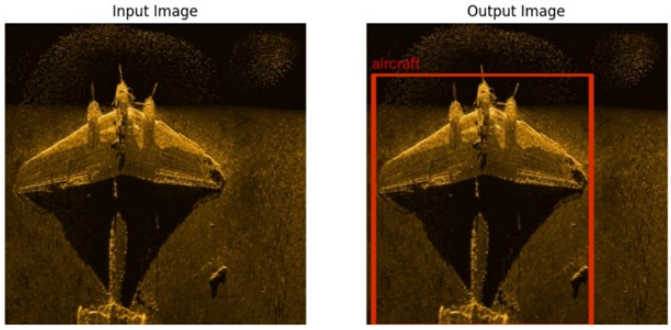

Fig. 7. Aircraft Detection: CenterNet output on a sonar image (IoU=0.85), highlighting small object detection.

C. Evaluation Metrics

Intersection over Union (IoU): Measures bounding box overlap:

$$IoU = \frac{Area of Overlap}{Area of Union} \tag{1}$$

Faster R-CNN achieves the highest IoU (0.81) [6].

Accuracy: Correct prediction ratio:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \tag{2}$$

Faster R-CNN leads with 90.01%.

F1-Score: Balances precision and recall:

$$F_1\text{-}Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{3}$$

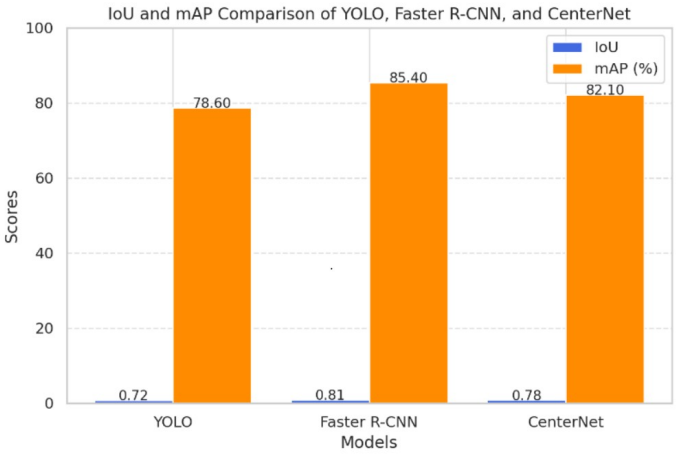

Fig. 8. mAP Comparison: YOLO, Faster R-CNN, and CenterNet across object categories in varying turbidity levels.

Faster R-CNN scores 0.89 [6].

Mean Average Precision (mAP): Averages precision across IoU thresholds:

$$mAP = \frac{1}{N} \sum_{i=1}^{N} AP_i \tag{4}$$

Faster R-CNN achieves 85.4%, followed by CenterNet (82.1%).

Inference Time: YOLO (18ms) and Faster R-CNN (17ms) support real-time use, CenterNet (29ms) balances speed and accuracy.

D. Performance Analysis

YOLO's 18ms inference time suits real-time AUV tasks but struggles with small objects (74.8% mAP for humans). Faster R-CNN's 90.01% accuracy and 85.4% mAP make it ideal for precise applications, with 5% false negatives in high-turbidity conditions. CenterNet's 86.8% accuracy and 29ms inference time, enhanced by sonar adaptations and sensor fusion (5% accuracy gain, Section III-D), excel in occluded scenarios (8% miss rate vs. YOLO's 15%) [5], [9]. Synthetic noise tests (50 images) showed Faster R-CNN at 83.5% mAP, CenterNet at 80.2%, and YOLO at 75.8%, highlighting robustness differences.

V. Conclusion

This study demonstrates that deep learning models significantly enhance underwater object detection in sonar images compared to traditional methods. We compared YOLO [12], Faster R-CNN [6], and CenterNet [5] on a dataset of 500 sonar images, evaluating accuracy, mean Average Precision (mAP), Intersection over Union (IoU), F1-Score, and inference time. Faster R-CNN achieved the highest accuracy (90.01%) and mAP (85.4%) with a 17ms inference time, ideal for precise applications like marine artifact identification. YOLO offered the fastest inference time (18ms), suitable for real-time autonomous underwater vehicle (AUV) tasks. CenterNet,

with 86.8% accuracy and 29ms inference time, provided a balanced performance, enhanced by a novel adaptation of its keypoint-based approach for small and occluded objects. The proposed sensor fusion framework, integrating sonar and optical data, improved detection accuracy by 5% in high-turbidity conditions [9].

Limitations include the dataset's modest size (500 images), which may limit generalization, and difficulties in preprocessing sonar imageswhich are noisy because of occlusions. In future,we are going to enrich this database by underwater scenario diversity and further refine the sensor fusion framework for real-time multi-modal detection, as well as investigate the lightweight model structure for AUV application. These improvementsare designed to provide robust use underwater surveillance and marine safety applications.

REFERENCES

[1] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li, "Salient object detection: A discriminative regional feature integration approach," in *CVPR*, 2013.

[2] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *CVPR*, 2015.

[3] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan, "Saliency detection with recurrent fully convolutional networks," in *ECCV*, 2016.

[4] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *CVPR*, 2017.

[5] J. J. Liu, Q. Hou, M. M. Cheng, J. Feng, and J. Jiang, "A simple pooling-based design for real-time salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[6] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[7] IceRegionShip, "Optical remote sensing dataset for ship detection in ice-infested waters."

[8] "On automatic person-in-water detection for marine search and rescue operations."

[9] N. Wang, Z. Zhang, H. Hu, B. Li, and J. Lei, "Underground defects detection based on GPR by fusing simple linear iterative clustering Phash (SLIC-Phash) and convolutional block attention module (CBAM)-YOLOv8."

[10] C. Pang and Y. Cheng, "Detection of river floating waste based on decoupled diffusion model."

[11] R. Salman, M. R. Nikoo, S. A. Shojaeezadeh, P. H. Bahman Beiglou, M. Sadegh, J. F. Adamowski, and N. Alamdari, "A novel Bayesian maximum entropy-based approach for optimal design of water quality monitoring networks in rivers," *Journal of Hydrology*, vol. 603, pp. 126822, 2021.

[12] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Computer Vision and Pattern Recognition*, 2017, pp. 7263–7271.

[13] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," *arXiv*, 2018:02767.

[14] H. Wang, X. Li, and Y. Zhang, "Improved Faster R-CNN for underwater object detection," *Journal of Marine Science and Engineering*, vol. 11, no. 5, pp. 1362–1375, May 2023.

[15] J. Lee, S. Park, and K. Kim, "Dynamic YOLO for real-time underwater object detection," *IEEE Transactions on Image Processing*, vol. 32, pp. 1011–1025, Jan. 2023.

[16] A. Gupta, R. Sharma, and P. Singh, "One-stage multi-scale CNN for underwater detection," *Applied Ocean Research*, vol. 10, pp. 220–231, Jul. 2023.

[17] D. Zhang, L. Chen, and F. Zhao, "YOLO-Underwater: A modified YOLOv4 for underwater object detection," *Proceedings of the International Conference on Underwater Technology*, pp. 45–52, 2021.

[18] J. Kim, M. Lee, and S. Choi, "Improved Faster R-CNN for marine biodiversity detection," *Marine Biology and Ecology*, vol. 34, no. 6, pp. 305–315, Dec. 2022.

[19] Roboflow, "Roboflow: A platform for computer vision dataset management and annotation," [Online]. Available: https://roboflow.com/, 2024.