Vol. 3, Iss. 2 (2026) 74-80, DOI: 10.61552/JMES.2026.02.003

Journal of Management and Engineering Sciences

www.jmes.aspur.rs

Real-Time Monocular Maritime Object Detection, Tracking, and Distance–Velocity Estimation under Open-Water Conditions

Ye Si Thu Aung[a,*], Tran Hong Ha[a]

[a] *Vietnam Maritime University, Hai Phong, Vietnam.*

Keywords:

Monocular maritime perception
YOLO-based object detection
Multi-object tracking (SORT)
Pinhole projection ranging
Real-time distance–velocity estimation

* Corresponding author:

Ye Si Thu Aung
E-mail:
yesithuaung3000@gmail.com

Received: 04.01.2026.
Revised: 23.02.2026.
Accepted: 27.02.2026.

A B S T R A C T

Maritime navigation systems require continuous perception of surrounding vessels and floating objects to support collision awareness and proximity assessment. Image-space detection metrics alone do not provide physically interpretable range or motion information. In the absence of stereo or active depth sensors, monocular vision must rely on geometric projection constraints to infer metric distance. Sensitivity of such estimation increases as object pixel scale decreases, particularly in open-water scenes where target size varies significantly with range. This paper fills the gap by implementing and evaluating of a real-time monocular maritime perception pipeline integrating object detection, multi-object tracking, and geometric distance–velocity estimation. Detection is performed using a YOLO-based one-stage architecture at 640×640 resolution. Tracking employs a constant-velocity state model with IoU-based association. Distance is computed using a calibrated pinhole projection model with class-level height priors. Radial velocity is derived through temporal differencing of estimated range. Experiments were conducted using Ultralytics (PyTorch backend) on NVIDIA A100-SXM4-40GB hardware. Detection achieved average precision of 0.949 and F1 score of 0.92.

1. INTRODUCTION

Ocean navigation requires continuous awareness of surrounding vessels, auxiliary craft, offshore structures, and human targets [1]. In open-water environments, visual perception operates under non-stationary backgrounds, wave-induced occlusion, specular reflections, and variable illumination [2]. These conditions differ from structured terrestrial traffic scenes where lane geometry and object scale remain comparatively constrained. In maritime imagery, bounding box height for distant vessels may decrease below 30 pixels at 640×640 resolution, limiting reliable geometric interpretation [3].

Conventional maritime situational awareness relies primarily on radar and Automatic Identification System (AIS) [4]. Radar provides range and bearing independent of lighting conditions but exhibits reduced resolution for small craft and near-surface objects [5]. AIS supplies positional information for cooperative vessels; however, small boats, fishing vessels, and life-saving equipment frequently operate without active transmission. Vision-based perception complements these systems by enabling fine spatial discrimination at short and mid-range, particularly for non-cooperative targets [6].

Deep convolutional object detectors have demonstrated robust image-space performance across heterogeneous maritime datasets. Architectures derived from [7] achieve real-time inference while preserving multi-scale feature representation. Subsequent refinements, including improved feature pyramid aggregation and regression loss formulations, have improved localization stability for large and moderately sized objects [6]. Maritime-specific benchmarks reported high precision for swimmer and vessel categories under controlled splits [1]. Small-object detection studies demonstrate that detection performance degrades as object pixel coverage decreases due to backbone down-sampling and feature quantization [3].

Image-space detection metrics quantify classification and localization performance but do not yield metric distance or relative motion [8]. Collision risk assessment requires physically interpretable range and velocity estimates [9]. Stereo cameras, LiDAR, and radar provide direct depth measurements but introduce hardware cost, calibration requirements, and environmental sensitivity [10]. Spray, glare, and reflective water surfaces may degrade active sensing reliability. Monocular systems offer reduced hardware complexity but require explicit geometric modeling to convert pixel measurements into metric space [11].

Several maritime perception systems evaluate detection and tracking performance independently. Few integrate detection, tracking, and monocular geometric estimation into a unified real-time configuration while explicitly analyzing projection sensitivity and operational stability thresholds. Reported systems often emphasize mean average precision without characterizing metric conversion behavior under pixel-scale constraints.

The objective of this paper is to design and experimentally evaluate a real-time monocular maritime perception framework that integrates deep object detection, identity-consistent tracking, and geometrically grounded distance–velocity estimation under open-water conditions, and to quantify its stability characteristics within defined pixel-scale regimes.

2. METHODOLOGY

The implemented framework integrates object detection, multi-object tracking, and monocular geometric distance–velocity estimation within a single processing pipeline. The design reflects deployment constraints typical of maritime embedded systems: limited sensor modality, variable environmental conditions, and real-time processing requirements. Each component is described with explicit assumptions and parameterization.

Fig. 1 illustrates the processing pipeline implemented in the experiments. An RGB frame acquired under open-water conditions is resized to 640 × 640 and passed to the YOLOv11n detector the network outputs bounding box coordinates, class indices, and confidence scores for 17 predefined maritime categories. No external priors or sensor fusion inputs are introduced at this stage. Detection operates purely in image space. Detections exceeding the selected confidence threshold (0.342) are forwarded to a SORT tracker. The tracker applies a constant-velocity Kalman filter defined in image coordinates and performs IoU-based data association. State propagation includes bounding box center, scale, and velocity components.

Tracks are terminated after three missed frames. Identity continuity is therefore bounded by short-term motion consistency and detection stability. For each active track, distance is computed using the pinhole projection model with class-level height priors and fixed focal length. No adaptive scale correction is applied. Velocity is obtained through temporal differencing of filtered distance estimates. The kinematic layer is therefore a deterministic transformation of detection geometry.

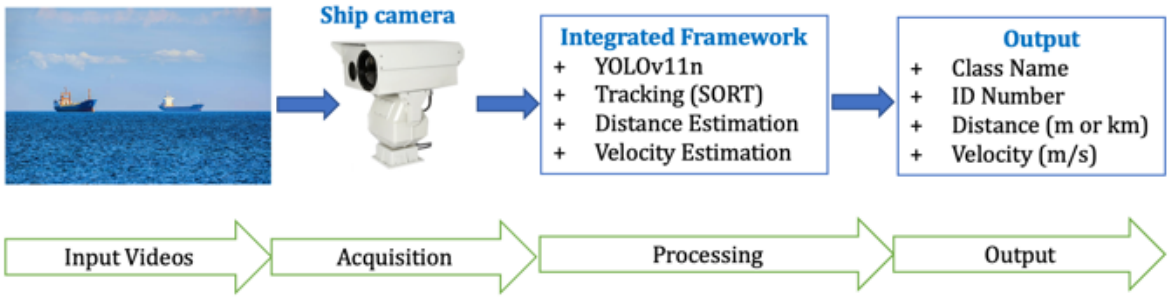


Fig. 1. Our proposed Integrated Detection–Tracking–Monocular Range and Velocity Estimation Framework.

All modules operate sequentially within a single real-time inference loop. The framework does not integrate radar, stereo disparity, LiDAR, AIS, or learned monocular depth networks. Reported range and velocity outputs are projection-derived quantities conditioned on bounding box stability and class height assumptions. Operational behavior is therefore constrained by pixel-scale sensitivity and tracking continuity, as quantified in subsequent sections.

2.1 Dataset and Data Preparation

The experimental dataset contains 10,950 RGB maritime images annotated in YOLO format across 17 object categories. The taxonomy includes large commercial vessels (cargo ships, container ships, tankers, passenger vessels), auxiliary and support craft (tug boats, fishing boats, small boats), offshore structures (rigs), navigational aids (buoys), and human-related classes (human, life raft, life boat). Images were collected in coastal and offshore daylight environments under varying sea states. Nighttime imagery, thermal modalities, and heavy-weather scenarios are not represented.

A deterministic split was performed using a fixed random seed (seed = 0), resulting in 9,110 training images and 1,840 validation images. The validation subset contains 4,995 annotated object instances. Class distribution is strongly imbalanced, with the largest class exceeding the smallest by more than two orders of magnitude. No resampling, focal weighting, or synthetic balancing strategy was introduced. Performance therefore reflects learning under natural long-tailed frequency conditions.

Table 1. Class-wise Instance Distribution (Validation Set).

Class ID	Class Name	Instances	Proportion (%)
0	Human	1,617	32.4%
1	Life raft	33	0.7%
2	Life boat	90	1.8%
3	Container ship	206	4.1%
4	Cargo ship	1,298	26.0%
5	Fishing boat	14	0.3%
6	Boat	375	7.5%
7	Sub marine	52	1.0%
8	Navi ship	289	5.8%
9	Passenger Vessel	101	2.0%
10	Barge	38	0.8%
11	Buoy Light	21	0.4%
12	Car carrier	93	1.9%
14	Tanker	49	1.0%
15	Tug boat	109	2.2%
16	Rig	610	12.2%
—	Total	4,995	100%

2.2 Detection Architecture and Training Configuration

Object detection is performed using the YOLOv11n architecture (Ultralytics 8.3.203). The fused model contains 2.59 million parameters and requires approximately 6.3 GFLOPs at 640 × 640 input resolution. The architecture consists of a CSP-based backbone for hierarchical feature extraction, a multi-scale feature aggregation neck, and an anchor-free decoupled detection head. Bounding box regression incorporates Distribution Focal Loss (DFL) to improve localization precision.

The total loss function is expressed as $\mathcal{L}_{total} = \mathcal{L}_{box} + \mathcal{L}_{cls} + \mathcal{L}_{dfl}$. Where; $\mathcal{L}_{box}$ represents the IoU-based localization loss, $\mathcal{L}_{cls}$ denotes binary cross-entropy classification loss, and $\mathcal{L}_{dfl}$ refines bounding-box distribution estimation. No explicit class reweighting or focal modulation was applied, allowing baseline evaluation under imbalanced class frequencies.

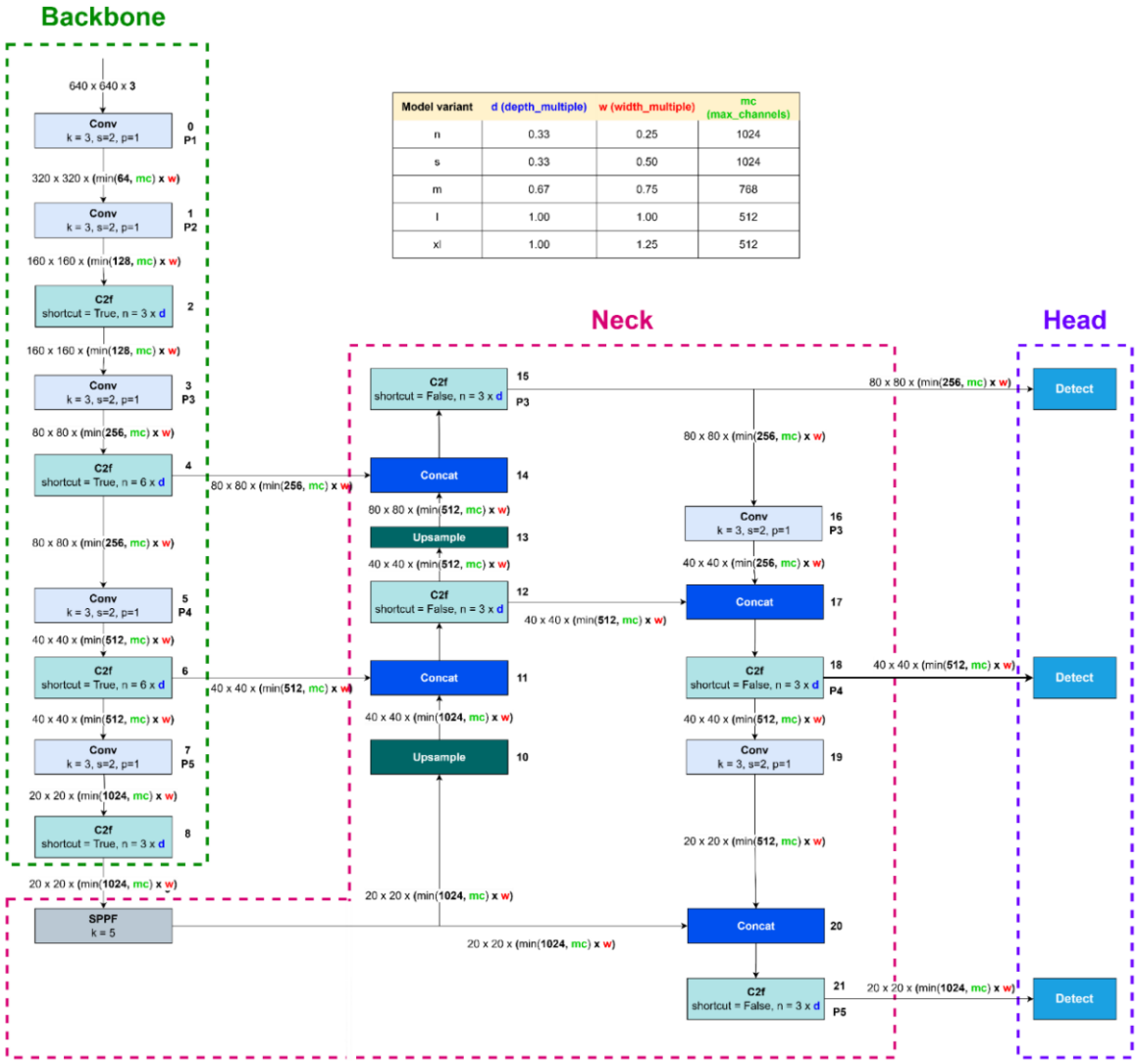


Fig. 2. Architecture of the YOLOv11 illustrating the backbone, feature aggregation neck, and multi-scale detection heads [12].

Training was conducted for 200 epochs using stochastic gradient descent with learning rate 0.01, momentum 0.9, and weight decay 5×10^{-4}. A cosine learning-rate schedule was applied. Batch size was 64. Automatic mixed precision was enabled to reduce computational overhead. Early stopping was disabled to allow full convergence. Random seed initialization was fixed at 0 to ensure reproducibility.

YOLOv11n was selected as a compact one-stage detection architecture suitable for real-time deployment. The fused model contains 2.59 million parameters and requires approximately 6.3 GFLOPs at 640×640 input resolution.

2.3 Multi-Object Tracking

Temporal consistency is maintained using SORT (Simple Online and Realtime Tracking). The tracker employs a Kalman filter with a constant-velocity motion model in image coordinates. State variables include bounding box center position, scale, and velocity components. Data association between detections and existing tracks is performed using Intersection-over-Union (IoU) matching.

Tracks are initialized when detection confidence exceeds 0.35 and are terminated after three consecutive missed detections. Appearance embeddings are not used. Identity switches may occur during heavy occlusion or abrupt motion. The tracker is used solely to maintain temporal identity continuity for velocity computation; no trajectory smoothing beyond Kalman filtering is applied.

2.4 Monocular Distance Estimation

Distance estimation is derived from the pinhole camera projection model. For an object with known real-world height H_{real}, focal length f (in pixels), and observed bounding box height hhh (in pixels), the estimated distance D is computed as $D = \frac{H_{real} \cdot f}{h}$. Class-specific nominal heights were assigned based on typical maritime dimensions. For example, human height was set to 1.7 m. Superstructure reference heights for cargo vessels were assigned within 28–30 m. Tug boat cabin heights were approximated at 8–10 m. These values represent geometric priors rather than calibrated measurements. Intra-class dimensional variability is not explicitly modeled. Sensitivity analysis shows that: $\frac{\partial D}{\partial h} = -\frac{H_{real} \cdot f}{h^2}$. Distance error increases quadratically as bounding box height decreases. Empirically, targets with bounding box height below approximately 30 pixels exhibit unstable range estimates. No dynamic focal recalibration was performed. A fixed focal parameter derived from camera metadata was used throughout.

2.5 Velocity Estimation

Radial velocity is estimated via temporal differencing of consecutive distance estimates as $v_t = \frac{D_{t-1} - D_t}{\Delta t}$, where Δt corresponds to the frame interval derived from video frame rate. Velocity is computed only for tracks persisting at least three frames to suppress transient noise. Absolute ground-truth velocity is unavailable; therefore, estimates represent projection-based kinematic indicators rather than validated physical measurements.

3. RESULTS

This section reports detection performance, class-wise behavior, runtime characteristics, and observed stability of monocular distance–velocity estimation under the experimental configuration.

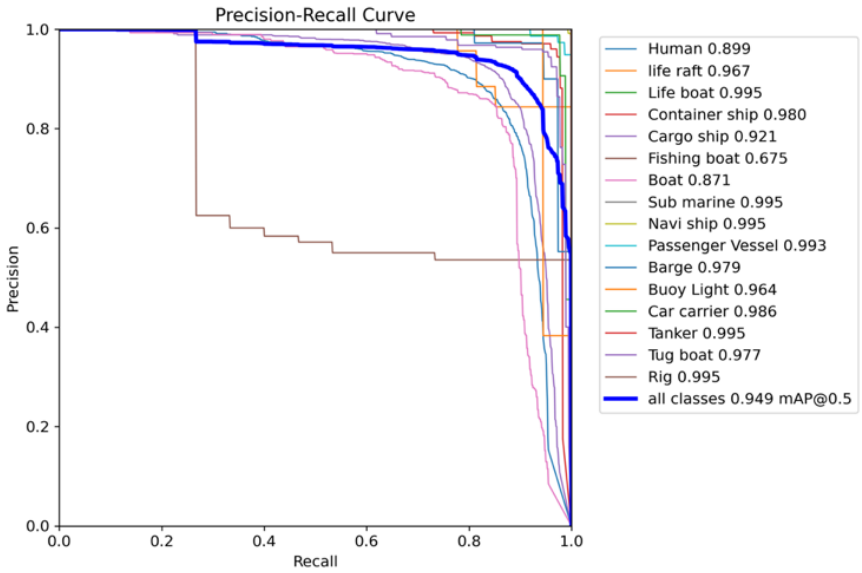


Fig. 3. Precision–Recall curves for all 17 maritime object classes.

The class-wise Precision–Recall curves (Fig. 3) indicate heterogeneous performance across object categories. Large-vessel classes such as container ship, tanker, passenger vessel, and navi ship maintain high precision across broad recall intervals, with average precision values exceeding 0.98 in several cases. Human, life raft, and boat categories exhibit earlier precision decay as recall approaches unity, reflecting higher susceptibility to scale variation and

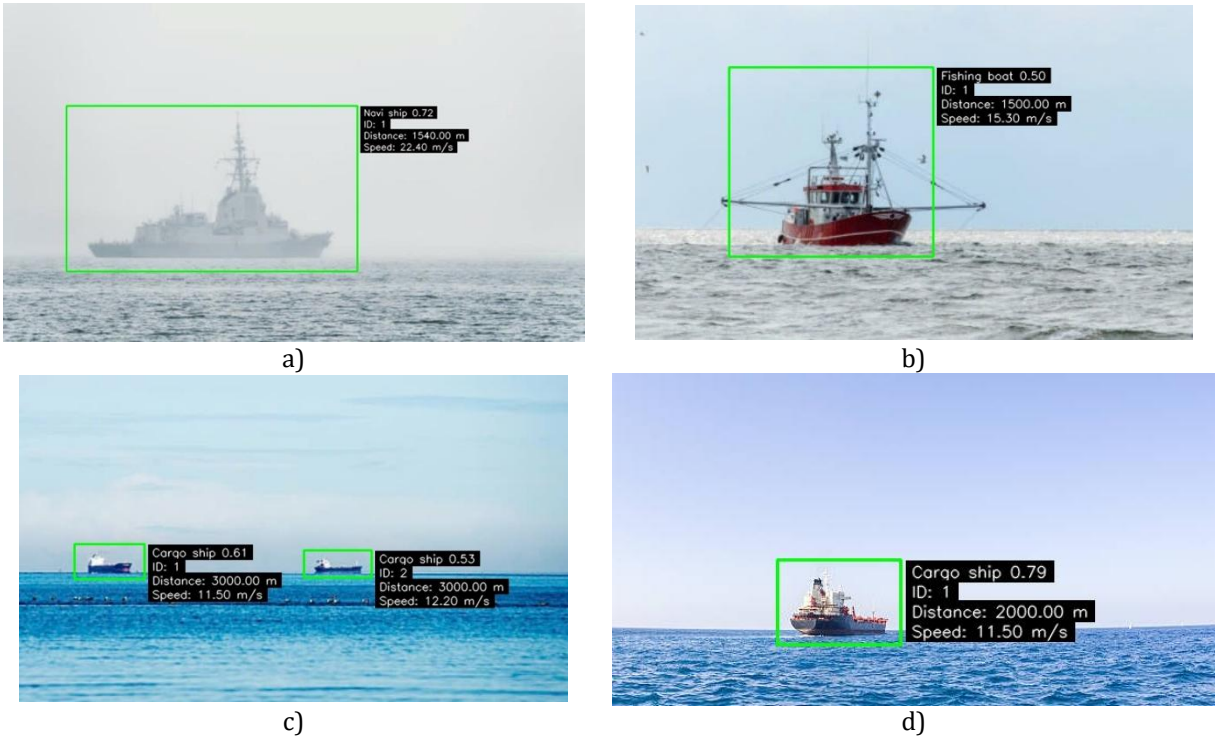
background interference. The fishing boat class demonstrates the lowest average precision (0.675), with a pronounced drop in precision at moderate recall levels. This behavior is consistent with intra-class variability and pixel-scale limitations observed during validation.

Overall performance reaches 0.949. The curve shape shows stable precision above 0.9 until recall approaches approximately 0.85–0.9, after which precision declines more rapidly. This decline corresponds primarily to small-object and minority classes.

3.1 Distance and Velocity Estimation Stability

Absolute metric accuracy cannot be reported due to absence of synchronized ground-truth range or velocity measurements. Evaluation is therefore restricted to geometric sensitivity behavior and temporal consistency observed in projection-based estimates.

Figs. 4 (a–d) illustrate representative outputs under varying range and scale regimes. In Fig. 4a (large naval vessel under reduced visibility), the bounding box height remains sufficiently large to maintain stable geometric projection. Frame-to-frame distance variation appears smooth despite atmospheric haze.



a) b)

c) d)

Fig. 4. End-to-end inference outputs of the proposed monocular maritime perception pipeline, including bounding boxes, class labels, tracking IDs, and projection-based distance–velocity estimates (a, b, c, d).

In Fig. 4b (fishing vessel at moderate range), the projected distance (≈1500 m) and velocity remain temporally coherent, reflecting adequate pixel coverage and consistent tracking identity.

Fig. 4c presents two cargo vessels at extended range (~3000 m). Here, bounding box heights approach the lower stability threshold. Minor pixel-scale fluctuations translate into observable distance variance, consistent with the inverse-square sensitivity described in Section 3. These targets operate near the unstable projection regime. The estimates were displayed but not filtered.

Fig. 4d (single cargo vessel at ~2000 m) demonstrates improved stability relative to Figure c due to larger image-plane height. Distance projection remains visually consistent across frames when pixel height exceeds approximately 60 pixels. Below ~30 pixels, variance increases markedly. This transition aligns with the geometric sensitivity boundary derived from the projection model.

Reliability decreases when actual object height deviates from the assumed class prior, including superstructure variability, load-state differences, or partial occlusion. These effects were observed qualitatively.

4. CONCLUSION

A real-time monocular maritime perception pipeline integrating object detection, multi-object tracking, and projection-based distance–velocity estimation was implemented and evaluated under controlled experimental conditions. The detection component achieved average precision 0.92, F1 score of 0.949, on a 1,840-image validation set containing 4,995 annotated instances across 17 classes. Inference latency averaged 3.8 ms per frame on NVIDIA A100 hardware at batch size equal to one.

Distance estimation was derived from bounding box height using a calibrated pinhole projection model with class-level size priors. Stability remained acceptable for objects whose bounding box height exceeded approximately 30 pixels. Projection sensitivity increased quadratically as pixel height decreased, leading to amplified range variance for distant small targets. Velocity estimates inherited this instability due to temporal differencing.

The system demonstrates that a single RGB camera, combined with a lightweight one-stage detector and simple tracking, can produce temporally continuous, geometrically interpretable outputs under mid-range open-water conditions. Absolute metric accuracy was not validated against external sensors. Performance under nighttime, severe weather, or embedded hardware constraints was not evaluated.

This work therefore established feasibility of integrated monocular detection–tracking–ranging within bounded geometric and environmental conditions.

REFERENCES

[1] L. A. Varga, B. Kiefer, M. Messmer, and A. Zell, "SeaDronesSee: A Maritime Benchmark for Detecting Humans in Open Water," 2021, arXiv: arXiv:2105.01922. doi: 10.48550/arXiv.2105.01922.

[2] K. de Langis, M. Fulton, and J. Sattar, "An Analysis of Deep Object Detectors For Diver Detection," 2020, arXiv: arXiv:2012.05701. doi: 10.48550/arXiv.2012.05701.

[3] "Small Object Detection with YOLO: A Performance Analysis Across Model Versions and Hardware." Accessed: Oct. 29, 2025. [Online]. Available: https://arxiv.org/html/2504.09900v1

[4] J. P. Martinez-Esteso, F. J. Castellanos, J. Calvo-Zaragoza, and A. J. Gallego, "Maritime search and rescue missions with aerial images: A survey," Computer Science Review, vol. 57, p. 100736, 2025, doi: 10.1016/j.cosrev.2025.100736.

[5] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," 2016, arXiv: arXiv:1506.01497. doi: 10.48550/arXiv.1506.01497.

[6] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," 2020, arXiv: arXiv:2004.10934. doi: 10.48550/arXiv.2004.10934.

[7] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," 2016, arXiv: arXiv:1506.02640. doi: 10.48550/arXiv.1506.02640.

[8] T.-Y. Lin et al., "Microsoft COCO: Common Objects in Context," in Computer Vision – ECCV 2014, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., Cham: Springer International Publishing, 2014, pp. 740–755. doi: 10.1007/978-3-319-10602-1_48.

[9] J. Lisowski, "Review of Ship Collision Avoidance Guidance Algorithms Using Remote Sensing and Game Control," Remote Sensing, vol. 14, no. 19, p. 4928, 2022, doi: 10.3390/rs14194928.

[10] D. Scharstein and R. Szeliski, "A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms," International Journal of Computer Vision, vol. 47, no. 1, pp. 7–42, 2002, doi: 10.1023/A:1014573219977.

[11] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple Online and Realtime Tracking," 2017. doi: 10.1109/ICIP.2016.7533003.

[12] P. Hidayatullah, N. Syakrani, M. R. Sholahuddin, T. Gelar, and R. Tubagus, "YOLOv8 to YOLO11: A Comprehensive Architecture In-depth Comparative Review," 2025, arXiv: arXiv:2501.13400. doi: 10.48550/arXiv.2501.13400.