technologies

MDPI

Article

YOLO-Driven Vessel Detection and Multi-Object Tracking in Fixed-Camera Marina Environments

Nikola Lopac [1,2,*], Karlo Severinski [1,2], Neven Grubišić [1] and Jonatan Lerga [2,3]

[1] Faculty of Maritime Studies, University of Rijeka, 51000 Rijeka, Croatia; karlo.severinski@pfri.uniri.hr (K.S.); neven.grubisic@pfri.uniri.hr (N.G.)
[2] Center for Artificial Intelligence and Cybersecurity, University of Rijeka, 51000 Rijeka, Croatia; jonatan.lerga@riteh.uniri.hr
[3] Faculty of Engineering, University of Rijeka, 51000 Rijeka, Croatia
* Correspondence: nikola.lopac@pfri.uniri.hr

Abstract

Vessel detection and multi-object tracking (MOT) in marinas remain challenging because of occlusions, small targets, cluttered backgrounds, and changing illumination. This paper presents a controlled, feasibility-focused comparative evaluation of YOLO-driven vessel detection and tracking in a fixed-camera marina environment, comparing YOLOv11 with the more attention-oriented YOLOv12 detector family. A dataset of 3546 annotated images was collected using a static ground-level camera at a marina in the northern Adriatic, Croatia. YOLOv11s, YOLOv11m, YOLOv12s, and YOLOv12m were fine-tuned using transfer learning, evaluated across five random seeds, and integrated into a common tracking-by-detection pipeline with Kalman filter (KF) and extended Kalman filter (EKF) motion models. Detection was assessed using COCO-style mAP metrics, while tracking was evaluated using MOT and HOTA-based metrics. YOLOv11m achieved the highest mean test-set mAP@50–95 (0.7458 ± 0.0027) and significantly outperformed YOLOv11s and YOLOv12s. Although YOLOv12m did not achieve the highest frame-level mAP, it obtained the highest mean AssA and IDF1 when averaged across the two motion models. KF achieved higher mean HOTA, DetA, AssA, MOTA, and IDF1 than EKF, whereas EKF achieved only marginally higher MOTP; the KF advantage was statistically significant for IDF1. Overall, the results show that frame-level detection accuracy does not necessarily determine downstream tracking performance, thereby supporting the joint evaluation of detector and motion-model choices.

Keywords: vessel detection; multi-object tracking; YOLOv11; YOLOv12; Kalman filter; marina surveillance

Check for updates

Academic Editor: Valeri Mladenov

Received: 1 July 2026
Revised: 27 July 2026
Accepted: 29 July 2026
Published: 3 August 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Advances in deep learning have substantially improved the accuracy and efficiency of object detection in real-world visual monitoring systems, increasing the applicability of vision-based perception in safety- and monitoring-critical environments [1]. In the maritime sector, machine learning and computer vision are increasingly used to support situational awareness, automation, and operational decision-making [2,3]. These developments form part of broader efforts toward digitalization and intelligent monitoring in maritime applications [4,5], where visual systems can contribute to vessel detection, traffic monitoring, and collision-risk awareness [6].

For continuous maritime monitoring, accurate frame-level detection alone is insufficient. Multi-object tracking (MOT) must also maintain consistent vessel identities over time despite missed detections, localization errors, and temporary loss of visibility. This requirement is particularly challenging in crowded marinas, where vessels appear at different scales and distances and are frequently affected by partial or near-complete occlusion. Dense structures such as docks, masts, ropes, and moored vessels produce persistent visual clutter, while glare, reflections, shadows, and changing weather conditions further reduce image consistency. In tracking-by-detection pipelines, these factors influence not only detection accuracy but also downstream association quality, track continuity, and identity preservation.

One-stage detectors from the You Only Look Once (YOLO) family are widely used in visual monitoring because they provide a favorable balance between detection performance and computational efficiency [7]. Recent YOLO generations differ in their approaches to multi-scale feature representation, contextual aggregation, and computational design. In particular, attention mechanisms originating from Transformer architectures [8] have increasingly been incorporated into visual models to improve contextual reasoning in cluttered scenes. Such architectural differences may affect both frame-level predictions and their temporal stability.

Maritime vessel detection has been investigated using satellite, Synthetic Aperture Radar (SAR), thermal, and unmanned aerial vehicle (UAV) imagery. Although these sensing modalities are important for wide-area maritime monitoring, their viewpoints, target appearances, and scene structures differ substantially from those of fixed-camera marina surveillance. Ground-level cameras can often be integrated into existing infrastructure and provide continuous coverage of docks, berths, and vessel movements. However, fixed-camera monitoring in crowded marinas has received comparatively less attention than aerial, remote-sensing, open-water, and general port-surveillance scenarios [9]. Moreover, many existing evaluations emphasize frame-level detection, traffic counting, or distance and speed estimation, whereas persistent vessel monitoring also requires metrics that quantify temporal association quality and identity preservation.

Motivated by these considerations, this paper presents a controlled comparative evaluation of YOLO-based vessel detection and MOT in a fixed-camera marina environment. The contribution lies in the matched experimental assessment of recent detector variants and motion models on a newly collected dataset representative of this comparatively underexplored setting. The dataset of 3546 annotated images was collected using a static ground-level camera in a marina in the northern Adriatic, Croatia. The dataset captures substantial variation in vessel scale, illumination, background complexity, overlap, and occlusion. Additional dedicated tracking sequences were prepared to assess temporal association under predefined levels of vessel overlap and occlusion.

The study compares the small and medium variants of YOLOv11 and YOLOv12, two recent detector families with different architectural emphases, particularly the comparatively stronger attention-oriented design of YOLOv12. Each detector is integrated into a unified tracking-by-detection framework and evaluated separately with Kalman filter (KF) and extended Kalman filter (EKF) motion models. All detector–motion-model combinations use the same intersection-over-union (IoU)-based association procedure and a common tracking operating point, enabling a controlled comparison of detector and motion-model selection. Detection performance is evaluated across five independent training seeds, while tracking performance is assessed using conventional MOT and Higher Order Tracking Accuracy (HOTA)-based metrics. Pipeline processing rate, latency, and peak GPU memory allocation are also reported to characterize the computational performance and resource usage of the evaluated configurations under the tested hardware and software setup.

The main contributions of this work are summarized as follows:

1. A newly collected and annotated fixed-camera marina dataset comprising 3546 images and additional dedicated tracking sequences. The dataset represents a comparatively underexplored maritime-monitoring setting characterized by dense infrastructure, reflections, scale variation, vessel overlap, and partial or near-complete occlusion.
2. A multi-seed comparative evaluation of YOLOv11s, YOLOv11m, YOLOv12s, and YOLOv12m under common training and evaluation conditions. Five independent training runs are used to assess detection performance, run-to-run variability, and the statistical significance of differences among detector variants on the proposed dataset.
3. A joint analysis of frame-level detection and downstream tracking performance, demonstrating that the detector with the highest mean frame-level mAP does not necessarily achieve the strongest identity-preservation performance.
4. A controlled comparison of KF- and EKF-based motion models using identical detector outputs, the same IoU-based association procedure, and a common tracking operating point, enabling the isolated assessment of motion-model effects on tracking performance.

The remainder of this paper is organized as follows. Section 2 reviews related work on vessel detection and tracking across sensing modalities and operational settings. Section 3 provides an overview of the evaluated object detectors and motion models. Section 4 describes the dataset, training procedure, and tracking framework. Section 5 presents and discusses the detection, tracking, statistical, and computational performance results. Finally, Section 6 summarizes the main findings and outlines directions for future work.

2. Related Work

Recent computer vision methods have enabled vessel detection and tracking across several maritime sensing modalities, including satellite and SAR imagery, thermal and nighttime remote sensing, UAV-based monitoring, and shore-based camera surveillance. These modalities differ substantially in viewpoint, object scale, background structure, temporal continuity, and deployment constraints. As a result, performance trends observed in one sensing setting are not necessarily transferable to another. This section reviews related work on vessel detection and tracking, with emphasis on how sensing modality and operational context influence model design, evaluation protocols, and practical deployment trade-offs.

2.1. Vessel Detection from Aerial and Remote-Sensing Imagery

A considerable body of research has addressed vessel detection from overhead and remote-sensing imagery. In SAR-based detection, Zhao et al. [10] proposed CRAS-YOLO, a YOLOv5s-based model for multi-category vessel detection and classification. Their approach combines an attention module with an improved feature-fusion neck and is evaluated on the ArtificialSAR-Vessel dataset, constructed from FUSAR-Ship instances [11] and SAR sea-surface backgrounds using SimpleCopyPaste augmentation [12]. The model reports improved performance over the YOLOv5s baseline and competitive generalization on SSDD [13]. Jiang et al. [14] similarly focused on SAR imagery, proposing a preprocessing strategy based on Non-Subsampling Laplacian Pyramid Decomposition and a lightweight YOLOv4-based detector. Their results highlight a recurring design trade-off in maritime detection: higher-capacity models can improve accuracy, while lightweight variants may provide a more practical balance between accuracy, inference speed, and model size.

Remote-sensing studies have also considered challenging sensing conditions such as thermal infrared and nighttime imagery. Li et al. [15] proposed a complete YOLO-based ship detection method for thermal infrared remote-sensing images, combining dataset

construction, contrast enhancement, and lightweight detector modification to improve robustness under low-texture conditions. Shao et al. [16] investigated nighttime vessel detection using VIIRS Day/Night Band imagery and proposed TASFF-YOLOv5, which incorporates a tiny-target detection layer, adaptive spatial feature fusion, and sea–land masking to reduce land-light false positives (FPs).

These studies demonstrate the effectiveness of YOLO-based detectors in specialized remote-sensing settings and provide valuable solutions for wide-area maritime monitoring. However, overhead SAR, thermal, and nighttime imagery differ markedly from fixed, ground-level marina surveillance. In remote-sensing imagery, vessels are typically observed from above, often with simplified object outlines and less persistent foreground clutter. In contrast, marina scenes contain dense infrastructure, masts, moored vessels, reflections, partial occlusions, and visually similar background structures. Consequently, high detection accuracy in remote-sensing benchmarks does not directly indicate tracking robustness or identity preservation in crowded ground-level marina environments.

2.2. UAV-Based Maritime Object Detection and Tracking

UAV imagery represents another important sensing modality for maritime object detection and tracking because it enables flexible coverage, rapid deployment, and observation from variable viewpoints. Vasilopoulos et al. [17] evaluated YOLOv5 detector variants on the SEAGULL dataset [18] and integrated the detections with DeepSORT for tracking. Their experiments on Jetson AGX Xavier hardware emphasized the trade-off between detection accuracy and real-time embedded performance, motivating the use of smaller YOLO variants in resource-constrained UAV applications.

Prayudi et al. [19] presented a UAV-based surveillance framework for illegal fishing prevention, comparing SSD MobileNetV2 for real-time embedded inference with Faster R-CNN ResNet50 for server-side processing. Their system also included hull plate identification using SIFT descriptors and K-nearest neighbor matching, illustrating how detection can be integrated with higher-level maritime monitoring tasks.

UAV-based studies are relevant because they often combine detection, tracking, and deployment considerations in end-to-end maritime perception pipelines. Nevertheless, UAV imagery introduces platform motion, rapidly changing viewpoints, and strong scale variations that differ from fixed-camera marina monitoring. In shore-based marina surveillance, the camera is static, but the visual scene is persistently cluttered; objects may remain partially occluded for long intervals, and identity preservation becomes important for continuous monitoring. Therefore, UAV-based tracking results provide useful methodological references but do not fully characterize the operational challenges of fixed-camera marina environments.

2.3. Ground-Level Maritime Surveillance

Ground-level camera systems provide a practical basis for near-shore maritime surveillance, port monitoring, and marina traffic analysis because they can often be integrated into existing infrastructure and provide continuous visual coverage of docks, berths, navigation channels, and vessel movements. Compared with overhead and UAV imagery, ground-level maritime scenes are characterized by low viewing angles, pronounced perspective effects, infrastructure clutter, water-surface reflections, and frequent inter-object overlap. These differences limit the direct transferability of results obtained from aerial or remote-sensing imagery and motivate the development of dedicated datasets, models, and evaluation protocols.

Several studies have adapted YOLO-based detectors to shore-based maritime imagery. Li et al. [20] proposed an enhanced YOLOv3-tiny network for real-time ship detection

on the SeaShips dataset [21]. Their method incorporates ship-specific anchor-box design, architectural modifications intended to improve sensitivity to small targets, and a Convolutional Block Attention Module (CBAM) to reduce interference from shoreline structures and water-surface backgrounds. Kim et al. [22] addressed the influence of dataset quality by introducing SMD-Plus, a revised version of the Singapore Maritime Dataset (SMD) [23]. Through tighter bounding-box annotations, corrected labels, consolidation of ambiguous classes, and task-specific augmentation, they demonstrated that annotation quality and dataset curation can substantially influence benchmark performance and reproducibility.

More recent studies have focused on architectural modifications for complex nearshore scenes. Wang et al. [24] proposed CDS-YOLO, a YOLOv10n-based detector designed to improve robustness under complex shoreline backgrounds, small targets, and occlusions. The model integrates Context Anchor Attention, Diverse Branch Blocks, and Slim-Neck components to strengthen feature representation while retaining computational efficiency. Sun et al. [25] modified YOLOv4 for unmanned surface vehicle (USV) perception by combining a weighted Cross Stage Partial Path Aggregation Network with a Feature Pyramid Transformer, specifically targeting long-distance small-object detection under adverse visibility conditions.

A closely related line of research is represented by the work of Petković et al., although the operational settings, objectives, and evaluation protocols of their studies differ from those of the present work. Their research primarily concerns vessel monitoring in passenger-port and collision-avoidance scenarios rather than identity-focused tracking in a crowded marina. Petković et al. [26] combined YOLOv3-based vessel detection with KF-based single-target tracking as a visual component of a collision-avoidance system, using footage captured by a camera mounted on a small vessel. Their qualitative analysis highlighted the difficulties associated with detecting distant vessels.

In subsequent work, Petković et al. [27] developed a fixed land-based camera system for vessel classification, multi-target tracking, and traffic counting at the entrance to the Port of Split. The system combined YOLOv4 with KF-based tracking, Hungarian assignment, and a track-stability assessment intended to suppress FP tracks. Its evaluation primarily addressed detection performance, vessel-counting accuracy and processing speed rather than identity-based MOT performance under systematically defined overlap and occlusion conditions.

Their later work extended the same fixed-camera monitoring system with monocular vessel-distance estimation based on the pinhole camera model [28]. The study compared Vessel-Focused Distance Estimation (VFDE), which uses YOLOv4 vessel detections and vessel-height information, with Vessel Object-Focused Distance Estimation (VOFDE), which uses YOLOv5 instance segmentation to identify standardized objects located on vessels. The methods were evaluated against instrument- and AIS-derived reference distances. The associated Split Port Ship Classification Dataset (SPSCD) [29] provides a dataset resource for vessel classification and monitoring in a realistic passenger-port environment.

More recently, Miličević et al. [30] developed a fixed shore-based monocular-vision framework that combines YOLOv9 and DeepSORT with geometry-based vessel-speed estimation validated against laser measurements. The estimated trajectories and speeds are intended to support an augmented-reality representation of collision risk for port-authority and vessel-traffic-service operators.

The principal distinction of the present study is its controlled evaluation of detector and motion-model selection within an identity-focused tracking-by-detection framework for a crowded marina environment. Four recent detector variants, YOLOv11s, YOLOv11m, YOLOv12s, and YOLOv12m, are evaluated on a newly collected fixed-camera marina dataset across multiple training seeds and combined with established KF and EKF motion

models using a common IoU-based association procedure and a shared tracking operating point. The tracking evaluation is conducted on dedicated sequences stratified according to predefined levels of vessel overlap and occlusion. In addition to COCO-style detection metrics, the study reports conventional MOT and HOTA-based metrics, including IDF1, HOTA, DetA, and AssA, thereby enabling a complementary assessment of detection quality, association performance, and identity preservation. The study additionally reports pipeline processing rate, mean latency, and peak GPU memory allocation under the evaluated setup.

Among the closely related studies considered here, previous evaluations primarily emphasize detection accuracy, vessel-counting performance, distance or speed estimation, and computational efficiency. The present work complements these objectives by examining how detector and motion-model choices affect downstream identity preservation under explicitly defined overlap and occlusion conditions. Table 1 summarizes the selected studies in terms of their primary objectives, detection-and-tracking configurations, and reported evaluation measures.

Table 1. Comparison of the present study with selected closely related maritime vessel-monitoring approaches.

Study	Primary Objective	Detection and Tracking Configuration	Reported Evaluation Metrics
[26]	Vessel detection and single-target tracking	YOLOv3 with KF	Processing time and qualitative evaluation
[27]	Vessel classification, multi-target tracking, and counting	YOLOv4 with KF	mAP, counting accuracy, and processing speed
[28]	Monocular vessel distance estimation	YOLOv4-based VFDE and YOLOv5-based VOFDE	Estimated distance and Percentage Deviation Index (PDI)
[30]	Vessel speed estimation and collision-risk support	YOLOv9 with DeepSORT and geometric speed estimation	Mean speed-estimation error, standard deviation, coefficient of variation, processing speed, and processing time
Present study	Vessel detection and multi-object tracking	YOLOv11/YOLOv12 with KF/EKF	mAP, HOTA, DetA, AssA, MOTA, MOTP, IDF1, pipeline processing rate, latency, and peak GPU memory allocation

2.4. Emerging Detector Architectures and Study Motivation

Recent YOLO generations increasingly differ in how they balance convolutional processing, multi-scale feature aggregation, computational efficiency, and attention-based contextual modeling. These architectural choices may influence not only frame-level detection accuracy but also the temporal consistency of detections. In tracking-by-detection pipelines, missed detections, localization fluctuations, and FPs propagate to the association stage and can therefore affect track continuity and identity preservation.

YOLOv11 and YOLOv12 provide a suitable basis for examining these effects because they represent recent detector families with different architectural emphases. YOLOv11 follows a predominantly efficiency-oriented convolutional design, whereas YOLOv12 places greater emphasis on attention-based feature aggregation and contextual representation. Such differences may be particularly relevant in crowded marina scenes containing small or partially visible vessels, reflections, infrastructure clutter, and recurrent occlusions. However, improvements in frame-level detection performance do not necessarily translate directly into more reliable downstream tracking.

Accordingly, the present study compares the small and medium variants of YOLOv11 and YOLOv12 under common training and evaluation conditions. Each detector is inte-

grated into the same tracking-by-detection framework using KF and EKF motion models, a common IoU-based association procedure, and a shared tracking operating point. Detection performance is evaluated across multiple training seeds, while tracking is assessed on dedicated fixed-camera marina sequences stratified according to predefined levels of vessel overlap and occlusion. This controlled design enables a direct assessment of how detector architecture and motion-model selection influence both frame-level detection performance and downstream association and identity-preservation quality.

3. Tracking-by-Detection Framework: Detectors and Motion Models

This section summarizes the main components of the tracking-by-detection framework evaluated in this work. The framework consists of two stages: frame-level vessel detection using YOLO-based object detectors and temporal association using KF-based motion models. Specifically, the study evaluates YOLOv11 and YOLOv12 detector variants and integrates their detections with KF- and EKF-based tracking components.

3.1. YOLO-Based Object Detection

The YOLO family consists of one-stage object detectors that predict object classes and bounding boxes in a single forward pass, providing a favorable balance between detection accuracy and computational efficiency. This property makes YOLO-based models suitable for monitoring applications in which timely frame-level detections are required. Recent YOLO generations introduce architectural refinements for multi-scale feature representation, feature aggregation, and robustness under visually complex conditions. In this work, YOLOv11 and YOLOv12 are evaluated as two recent YOLO detector families with different architectural emphases, particularly with respect to the stronger role of attention mechanisms in YOLOv12.

3.1.1. YOLOv11

YOLOv11 is provided within the Ultralytics framework and represents a recent evolution of the efficiency-oriented YOLO design. Compared with earlier YOLO variants, YOLOv11 introduces architectural refinements intended to improve feature representation while maintaining a compact and computationally efficient detection pipeline. One of its key components is the Cross Stage Partial with a kernel size of 2 (C3k2) block, which uses smaller-kernel convolutions to improve computational efficiency while preserving feature expressiveness. The architecture also retains the Spatial Pyramid Pooling–Fast (SPPF) module for low-latency multi-scale context aggregation. In addition, YOLOv11 incorporates the Cross Stage Partial with Spatial Attention (C2PSA) module after SPPF, allowing the network to emphasize spatially salient regions and improve feature representation across object scales. The network architecture used in this work follows the corresponding Ultralytics YAML specification [31].

Reported benchmark results indicate that YOLOv11 variants provide competitive accuracy–efficiency trade-offs on standard datasets such as MSCOCO. For example, YOLOv11m has been reported to achieve higher mAP than YOLOv8m with fewer parameters, indicating improved parameter efficiency [32]. Beyond object detection, the Ultralytics implementation supports related vision tasks such as instance segmentation, pose estimation, and oriented object detection, which further supports its use in downstream visual perception pipelines [32].

3.1.2. YOLOv12

YOLOv12, proposed by Tian et al. [33], extends the YOLO design by placing a stronger emphasis on attention-based feature extraction and fusion while retaining convolutional components. The resulting architecture can be viewed as an attention-centric YOLO de-

tector designed to improve contextual feature representation without abandoning the efficiency advantages of one-stage detection. Inspired by Vision Transformers (ViT) [34], YOLOv12 introduces an Area Attention module (A^2) to capture broader image context with reduced computational cost compared with dense self-attention. The architecture also incorporates FlashAttention to improve memory efficiency and throughput, addressing common limitations of attention-based visual models. Instead of relying on explicit positional encoding, YOLOv12 employs a 7×7 separable convolution position perceiver to encode spatial information within feature maps. These components are integrated into a YOLO-style backbone and neck through Residual Efficient Layer Aggregation Networks (R-ELAN). The model specifications used in this work follow the corresponding Ultralytics YAML implementation [31].

On the MSCOCO benchmark, YOLOv12 was reported to improve over YOLOv11 across several model scales under the evaluation setup of Tian et al. [33]. These reported gains suggest that the stronger use of attention mechanisms in YOLOv12 can improve detection performance while preserving computational practicality. Since marina scenes contain cluttered backgrounds, strong reflections, partial occlusions, and vessels at different scales, evaluating YOLOv12 together with YOLOv11 provides a useful basis for assessing how recent YOLO design choices affect both frame-level detection and downstream tracking performance in a fixed-camera maritime monitoring scenario.

3.2. Multi-Object Tracking in a Tracking-by-Detection Pipeline

MOT aims to maintain consistent object identities and estimate object trajectories across consecutive video frames, despite missed detections, occlusions, localization noise, and object interactions [35]. In tracking-by-detection pipelines, frame-level object detections are first produced independently by an object detector and are then associated over time to form trajectories. The quality of the resulting tracks therefore depends not only on the motion model, but also on detector confidence, localization stability, FPs, false negatives (FNs), and the reliability of data association.

MOT methods range from classical Bayesian filtering approaches to deep learning-based trackers that incorporate appearance descriptors and learned association models [36]. In fixed-camera marina monitoring, lightweight motion-model-based tracking remains attractive because it is computationally efficient, interpretable, and suitable for scenarios in which objects often move slowly or approximately smoothly in the image plane. In this work, YOLO detections provide frame-wise measurements, while the KF and EKF are used as motion-model components for temporal prediction and track updating. This enables a controlled assessment of how the evaluated YOLO detectors interact with two KF-based tracking variants in the same marina surveillance setting.

3.2.1. Kalman Filter (KF)

The KF is a recursive Bayesian estimator for linear Gaussian state-space models. It combines a process model, which predicts the evolution of the system state, with noisy measurements to estimate the latent state of a dynamic system over time [37]. In MOT, the state typically represents object position and motion in the image plane, while the measurements are derived from detector outputs. The KF follows a predict–update cycle, as illustrated in Figure 1a.

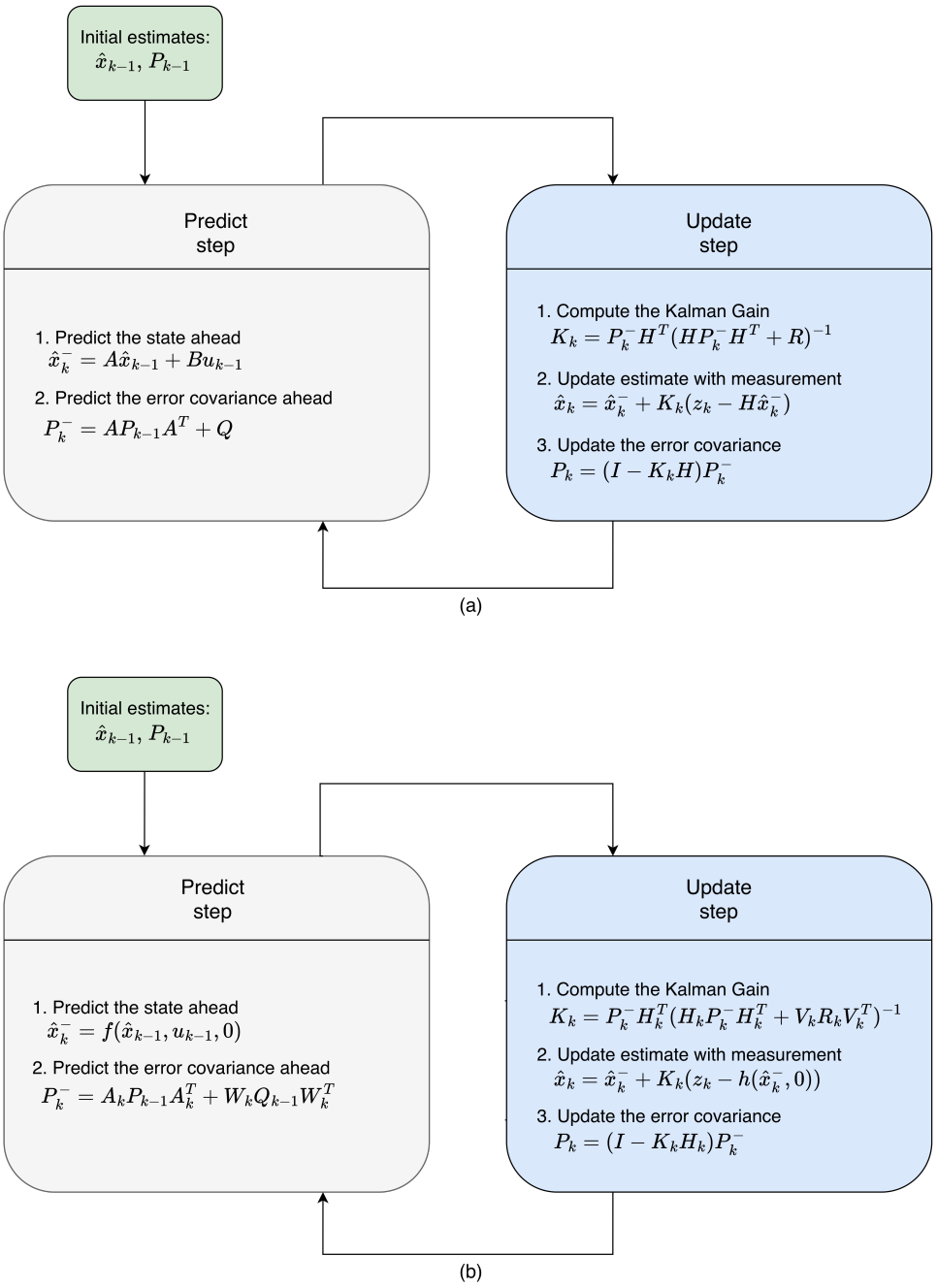


Figure 1. Predict–update structure of the tracking motion models: (**a**) Kalman filter (KF); (**b**) extended Kalman filter (EKF).

A linear discrete-time state-space model can be written as

$$x_k = Ax_{k-1} + Bu_{k-1} + w_{k-1}, \tag{1}$$

$$z_k = Hx_k + v_k, \tag{2}$$

where $x_k \in \mathbb{R}^n$ is the state vector, $z_k \in \mathbb{R}^m$ is the measurement vector, A is the state transition matrix, B is the control-input matrix, and H is the observation matrix. The process noise $w_{k-1} \sim \mathcal{N}(0, Q)$ and measurement noise $v_k \sim \mathcal{N}(0, R)$ are assumed to be zero-mean Gaussian and mutually independent.

The KF recursively alternates between prediction and update:

$$\hat{x}_k^- = A\hat{x}_{k-1} + Bu_{k-1}, \tag{3}$$
$$P_k^- = AP_{k-1}A^T + Q, \tag{4}$$
$$K_k = P_k^- H^T \left(HP_k^- H^T + R\right)^{-1}, \tag{5}$$
$$\hat{x}_k = \hat{x}_k^- + K_k\left(z_k - H\hat{x}_k^-\right), \tag{6}$$
$$P_k = (I - K_k H)P_k^-. \tag{7}$$

Here, $\hat{x}_k^-$ and P_k^- denote the predicted state estimate and covariance, respectively, while K_k is the Kalman gain.

This recursive procedure produces smoothed state estimates from noisy measurements and provides short-term predictions when detections are temporarily missing [37]. Since the KF assumes linear dynamics and a linear measurement model, its suitability depends on how well the object motion in the image plane follows the adopted linear approximation.

3.2.2. Extended Kalman Filter (EKF)

The EKF generalizes the KF to non-linear state-space models by linearizing the process and measurement functions around the current estimate using first-order Taylor approximations [37]. The EKF follows the same predict–update structure as the KF, but replaces the linear transition and observation matrices with non-linear functions and their local Jacobians, as illustrated in Figure 1b.

A non-linear discrete-time system can be modeled as

$$x_k = f(x_{k-1}, u_{k-1}, w_{k-1}), \tag{8}$$
$$z_k = h(x_k, v_k), \tag{9}$$

where $f(\cdot)$ is the non-linear state transition function and $h(\cdot)$ is the non-linear measurement function. The noise processes w_{k-1} and v_k are assumed to be zero-mean Gaussian with covariance matrices Q_k and R_k, respectively.

Given the previous estimate $(\hat{x}_{k-1}, P_{k-1})$, the EKF prediction step computes

$$\hat{x}_k^- = f(\hat{x}_{k-1}, u_{k-1}, 0), \tag{10}$$
$$P_k^- = A_k P_{k-1} A_k^T + W_k Q_{k-1} W_k^T, \tag{11}$$

where

$$A_k = \left.\frac{\partial f}{\partial x}\right|_{\hat{x}_{k-1}, u_{k-1}}, \tag{12}$$
$$W_k = \left.\frac{\partial f}{\partial w}\right|_{\hat{x}_{k-1}, u_{k-1}}. \tag{13}$$

When a measurement is available, the measurement model is linearized around the predicted state:

$$H_k = \left.\frac{\partial h}{\partial x}\right|_{\hat{x}_k^-}, \tag{14}$$
$$V_k = \left.\frac{\partial h}{\partial v}\right|_{\hat{x}_k^-}. \tag{15}$$

The Kalman gain and update step are then given by

$$K_k = P_k^- H_k^T \left(H_k P_k^- H_k^T + V_k R_k V_k^T\right)^{-1}, \tag{16}$$

$$\hat{x}_k = \hat{x}_k^- + K_k\left(z_k - h(\hat{x}_k^-, 0)\right), \tag{17}$$

$$P_k = (I - K_k H_k) P_k^-. \tag{18}$$

Here, A_k and H_k are local linear approximations of the non-linear process and measurement functions with respect to the state, while W_k and V_k describe how process and measurement noise enter the corresponding models.

The EKF is suitable when the system can be locally approximated by linearization and when non-linearities are mild to moderate [37]. In fixed-camera marina sequences, vessel motion in the image plane is often approximately smooth, but apparent motion can be affected by perspective, changes in direction, partial occlusion, and temporary missed detections. For this reason, both KF and EKF are evaluated as lightweight motion-model variants within the same tracking-by-detection framework. Their effect on identity preservation and tracking quality is assessed experimentally using MOT and HOTA metrics.

4. Methodologyand Experimental Setup

4.1. Dataset

Building on the preliminary study in [9], which introduced an initial ground-level marina dataset and YOLOv8-based experiments, the present work expands the data collection and annotation effort for fixed-camera vessel monitoring in a marina environment. The dataset was acquired using a single static ground-level camera positioned in a marina in the northern Adriatic, Croatia. This setup reflects a practical monitoring scenario in which vessels are observed from a fixed shore-based perspective under varying illumination, object scales, and background conditions.

The dataset was constructed from approximately 30 h of video footage recorded over four days at different times of day and under diverse lighting conditions. Frames were extracted from the original videos at fixed 3 s intervals to obtain temporally distributed samples while limiting excessive frame-to-frame similarity. To further reduce temporal redundancy, visually similar consecutive frames and near-duplicate samples were excluded. This step was used to increase visual diversity and avoid over-representation of nearly identical scenes rather than selectively retain or discard samples based on their semantic content. To ensure consistency in image geometry, only videos with a uniform spatial resolution of 2704×2028 pixels were included.

Particular care was taken to prevent data leakage between subsets. The training, validation, and test sets were formed without overlap so that identical scenes or video segments did not appear across different subsets. The final dataset contains 3546 images and was split into 70% training, 10% validation, and 20% testing. To improve background discrimination and reduce FPs, the training set was supplemented with unlabeled negative samples, corresponding to an additional 5% of frames without vessel annotations. Representative examples from the dataset are shown in Figure 2, illustrating variations in illumination, reflections, and vessel scale.

All images were manually annotated on a local workstation using LabelMe [38]. Initial annotations were prepared by one annotator and subsequently subjected to an independent secondary review. During this quality-control stage, missing, duplicated, or inaccurate annotations were identified, discussed, and corrected, and the final annotation set incorporated the agreed revisions. Annotation quality and consistency were therefore supported through secondary review and consensus-based correction of the identified discrepancies.



Figure 2. Representative frames from the proposed fixed-camera marina dataset under varying illumination, reflection, and vessel-scale conditions.

The annotation scheme comprised a single object class, *vessel*, represented by axis-aligned bounding boxes. Partially occluded vessels were annotated according to the visible extent, while vessels truncated by the image boundary were included with bounding boxes restricted to the visible image area. The single-class formulation is consistent with the objective of vessel detection and tracking in marina surveillance, where temporal association and identity preservation are evaluated at the object level rather than at the fine-grained vessel classification. Representative annotated examples are shown in Figure 3.

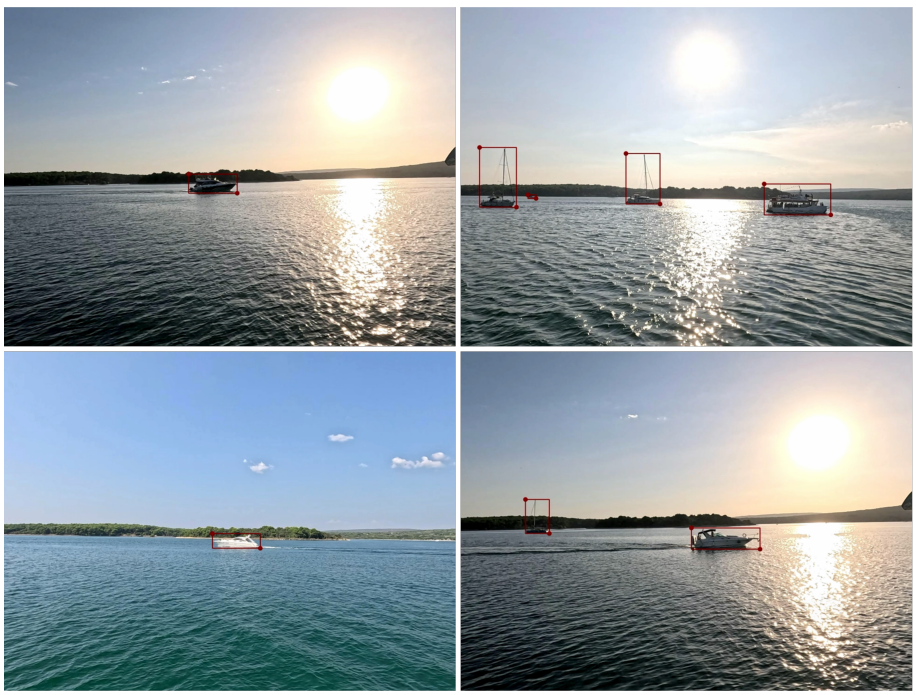


Figure 3. Representative annotated examples from the proposed dataset using a single object class, *vessel*, with axis-aligned bounding boxes.

4.2. YOLO Detector Training

To balance model capacity, dataset scale, and computational requirements, this study evaluates the small- and medium-scale variants of YOLOv11 and YOLOv12, namely YOLOv11s, YOLOv11m, YOLOv12s, and YOLOv12m. These variants provide a controlled comparison across two model scales while keeping the evaluation focused on detector configurations that are practically relevant for fixed-camera marina monitoring. The model definitions follow the corresponding Ultralytics YAML configurations [31]. A structural summary of the evaluated models is provided in Table 2.

Table 2. Structural complexity of the evaluated YOLO detector variants.

YOLO Variant	Layers	Params	GFLOPs
YOLOv11s	181	9.46 M	21.7
YOLOv11m	231	20.1 M	68.5
YOLOv12s	272	9.28 M	21.7
YOLOv12m	292	20.2 M	68.1

As shown in Table 2, the YOLOv12 variants use a larger number of layers than the corresponding YOLOv11 variants while maintaining similar parameter counts and comparable computational cost. For example, YOLOv12m is deeper than YOLOv11m, but has nearly the same number of parameters and slightly lower GFLOPs. A similar relationship is observed for the small-scale variants.

All models were trained on the proposed marina dataset using transfer learning. Specifically, publicly available pre-trained weights from the Ultralytics framework were used for initialization, followed by fine-tuning on the proposed dataset. Hyperparameter tuning was performed independently for each model variant using the built-in Ultralytics tuning procedure. To ensure a fair tuning procedure, the same hyperparameter search space was used for all evaluated YOLO detector variants. The search space covered optimization parameters, loss coefficients, and data augmentation settings, as summarized in Table 3.

The tuning process was configured for 120 training epochs, with 80 tuning iterations for the small-scale variants and 100 tuning iterations for the medium-scale variants. Early stopping was applied after 10 consecutive epochs without improvement. All tuning runs used a batch size of 32, the AdamW optimizer, and an input image size of 640×640 pixels. To ensure comparability across model variants during hyperparameter search, the tuning seed was fixed to 0, and the selection criterion was mAP@50–95 on the validation split of the proposed dataset.

After selecting the best hyperparameter configuration for each model variant, the corresponding architecture was retrained five times using different random seeds. This multi-seed retraining protocol was used to quantify the stability of detector performance and reduce dependence on a single initialization. The final retraining stage used the selected hyperparameters for each model, with training conducted for 220 epochs, a patience value of 40 epochs, and a warm-up period of 3 epochs. The resulting detection performance is reported in Section 5 as mean $\pm$ standard deviation across the five runs.

All experiments were conducted using the Ultralytics v8.3.160 Python package with Python 3.13.2 on a workstation equipped with an NVIDIA RTX A5000 GPU, an AMD Ryzen Threadripper 7960X processor with 24 cores, and 128 GB RAM. The same software and hardware environment was used for training and evaluation to ensure consistency across all detector variants.

Table 3. Common hyperparameter search space used for tuning all YOLO detector variants.

Parameter	Range
Optimization parameters	
lr0	$(2 \times 10^{-4}, 2 \times 10^{-3})$
lrf	(0.1, 0.5)
weight_decay	$(2 \times 10^{-5}, 1 \times 10^{-3})$
warmup_epochs	(1.0, 3.0)
warmup_momentum	(0.6, 0.9)
momentum	(0.1, 0.5)
Loss coefficients	
box	(6.0, 12.0)
cls	(0.3, 1.2)
dfl	(0.5, 1.5)
Data augmentation	
hsv_h	(0.0, 0.05)
hsv_s	(0.0, 0.4)
hsv_v	(0.0, 0.3)
degrees	(0.0, 7.0)
translate	(0.0, 0.2)
scale	(0.05, 0.2)
shear	(0, 3)
perspective	(0.0, 0.0005)
fliplr	(0.4, 0.6)
mosaic	(0.0, 1.0)
mixup	(0.0, 0.1)
close_mosaic	(5, 10)

4.3. Tracking-by-Detection Setup

To evaluate vessel tracking in marina video sequences, a tracking-by-detection pipeline was adopted. Frame-level detections produced by each YOLO detector are used as measurements within a motion-model-based tracking framework. For a given detector variant, the KF and EKF configurations receive identical detections and use the same data-association and track-management procedures, enabling a controlled comparison of the two motion models. The overall processing flow is summarized in Figure 4.

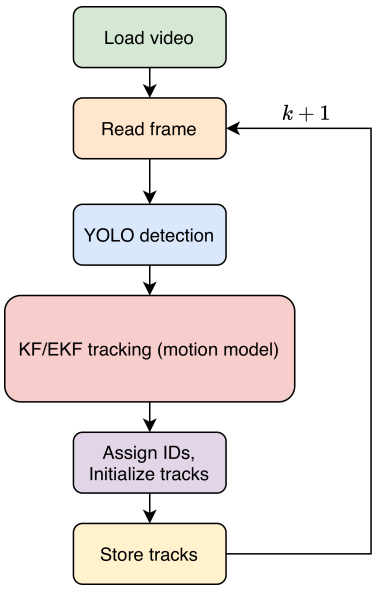


Figure 4. Tracking-by-detection pipeline combining YOLO detections with KF/EKF motion models.

For each video, frames are processed sequentially. At every frame, the detector produces vessel bounding boxes, from which the center coordinates, area, and aspect ratio are used as tracker measurements. Each active track first predicts its next state and corresponding bounding box. Pairwise IoU similarities between predicted and detected bounding boxes are then used to construct the assignment cost matrix. The Hungarian algorithm provides a minimum-cost one-to-one assignment for the current frame, after which candidate matches below the selected IoU threshold are rejected. Matched tracks are updated using the corresponding detections, whereas unmatched detections initialize new tracks. Unmatched tracks are retained for a limited number of frames to accommodate temporary missed detections and short occlusions. Track retention and confirmation are controlled by the maximum-age and minimum-hits parameters, respectively. This procedure is repeated until the end of each video sequence.

The KF uses a constant-velocity state

$$\mathbf{x}_k^{\mathrm{KF}} = \begin{bmatrix} c_x & c_y & A & r & v_x & v_y & v_A \end{bmatrix}^{\mathrm{T}}, \tag{19}$$

comprising the bounding-box center coordinates c_x and c_y, area A, aspect ratio r, and the corresponding frame-to-frame velocities v_x, v_y, and v_A.

The EKF uses the augmented state

$$\mathbf{x}_k^{\mathrm{EKF}} = \begin{bmatrix} c_x & c_y & A & r & v_x & v_y & \omega & v_A \end{bmatrix}^{\mathrm{T}}, \tag{20}$$

identical to the KF state but with an added image-plane turn rate ω; acceleration and heading are not modeled explicitly.

This turn rate is the source of the EKF's non-linearity. With $\phi = \omega \Delta t$, the center and velocity are propagated by a coordinated-turn update,

$$c_x' = c_x + \frac{\sin\phi}{\omega} v_x - \frac{1-\cos\phi}{\omega} v_y, \qquad v_x' = \cos\phi\, v_x - \sin\phi\, v_y, \tag{21}$$

with c_y' and v_y' defined analogously, while area and aspect ratio continue to evolve linearly. When $|\omega| \leq 10^{-3}$, the corresponding zero-turn-rate limit is used to avoid numerical instability caused by division by ω.

The corresponding Jacobian $\mathbf{F}_k = \partial f / \partial \mathbf{x}$ is the 8×8 identity matrix with additional off-diagonal terms coupling center position, velocity, and turn rate, evaluated at each predicted state. The observation Jacobians $\mathbf{H}_{\mathrm{KF}}$ and $\mathbf{H}_{\mathrm{EKF}}$ are simple selection matrices, since the measurement function itself is linear in both filters:

$$\mathbf{H}_{\mathrm{EKF}} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}, \tag{22}$$

with $\mathbf{H}_{\mathrm{KF}}$ defined analogously over the 7-dimensional KF state.

Measurement noise is set to the same covariance matrix for both filters:

$$\mathbf{R} = 10\mathbf{I}_4 = \mathrm{diag}(10, 10, 10, 10), \tag{23}$$

where the four components correspond to the measured bounding-box center coordinates, area, and aspect ratio, i.e., $\mathbf{z}_k = [c_x, c_y, A, r]^T$.

The process-noise covariance matrix for the KF is set to

$$\mathbf{Q}_{\mathrm{KF}} = \mathrm{diag}(1, 1, 1, 1, 100, 100, 100), \tag{24}$$

where the larger values correspond to the velocity and area-rate components.

For the EKF, the process-noise covariance matrix is set to

$$\mathbf{Q}_{\mathrm{EKF}} = \mathrm{diag}(1, 1, 1, 1, 100, 100, 0.01, 100), \tag{25}$$

where the velocity and area-rate components are assigned higher uncertainty, while the turn-rate component is assigned lower process noise.

Frame-level ground-truth annotations are required for quantitative tracking evaluation. Therefore, 12 short video sequences were extracted from the original recordings. To prevent data leakage, these sequences were kept separate from the YOLO image dataset and were not included in the detector training, validation, or test splits. The sequences are approximately 10–30 s long and were categorized as *Easy*, *Moderate*, or *Challenging* according to the severity of vessel overlap and occlusion. *Easy* sequences contain no vessel overlap, *Moderate* sequences contain partial overlap, and *Challenging* sequences contain complete or near-complete occlusion events in which one vessel passes behind another. Nine sequences were reserved for the final tracking evaluation, while three additional sequences, one from each difficulty level, were used exclusively for calibration of the common tracking operating point. The final test set contains 3688 frame-level vessel bounding-box annotations, whereas the calibration subset contains 1318.

To define a common tracking operating point before the final evaluation, three held-out calibration sequences were used, with one sequence selected from each predefined difficulty level. The objective was to identify a single configuration that could be applied uniformly to all detector–motion-model combinations rather than to optimize each configuration separately. The calibration sequences were excluded from the final test set, thereby preventing parameter selection on the test data.

The calibrated parameters were the detector confidence threshold, matching IoU threshold, maximum age, and minimum hits. Maximum age specifies the number of frames for which an unmatched track is retained, whereas minimum hits specifies the number of successful associations required before a track is confirmed. The initial parameter values, evaluated search ranges, and final selected values are reported in Table 4. Before a parameter was evaluated, parameters not yet considered were held at their initial values; parameters selected during earlier stages were held at their selected values.

Table 4. Calibration settings used to select the common tracking operating point.

Parameter	Initial Value	Search Range/Values	Selected Value
Confidence threshold	0.3	[0.3, 0.8], step 0.05	0.75
Matching IoU threshold	0.30	{0.30, 0.40, 0.50}	0.40
Maximum age	15	{15, 30, 45}	15
Minimum hits	1	{1, 2, 3}	3

The parameters were evaluated sequentially using a coordinate-wise search. The detector confidence threshold was considered first because it determines which detections enter all subsequent tracking stages. The matching IoU threshold was evaluated next because it directly controls the association of retained detections with predicted tracks. The track-management parameters were then considered, with maximum age evaluated before minimum hits because it governs the temporal retention of unmatched tracks, whereas minimum hits controls the confirmation of newly initialized tracks. At each stage, the value yielding the highest mean IDF1 score on the complete calibration subset was retained before proceeding to the next stage.

The sequential procedure was selected as a computationally practical means of obtaining a shared operating point for comparative evaluation. Alternative search orders

and exhaustive joint optimization were not examined. The procedure therefore does not explicitly capture all interactions among the parameters and does not guarantee a globally optimal joint configuration. Nevertheless, applying the same selected operating point to all detector–motion-model combinations prevents configuration-specific tuning and supports an internally consistent comparison under common tracking settings. The selected parameters should consequently be interpreted as a shared operating point for the evaluated setting rather than as universally optimal values.

All tracking sequences were resampled to 25 FPS prior to annotation. Ground-truth bounding boxes were annotated in CVAT [39], which supports video annotation and export to MOT-compatible formats. Tracking performance was evaluated using the TrackEval framework [40]. We report standard MOT metrics [41], including IDF1, MOTA, and MOTP, together with HOTA metrics [42], including HOTA, DetA, and AssA. This metric set captures both detection-related tracking quality and association consistency, which are particularly relevant in marina scenes with partial occlusions and visually similar vessels.

5. Results and Discussion

This section reports the experimental results for vessel detection on the proposed image dataset and MOT in marina video sequences. The detection results are first analyzed using COCO-style metrics, multi-seed variability, object-size-stratified AP, and qualitative error examples. The tracking results are then evaluated using MOT and HOTA metrics within the common tracking-by-detection pipeline evaluated in this study.

5.1. Object Detection Results

After the hyperparameter tuning procedure described in Section 4, one configuration was selected for each detector variant, namely YOLOv11s, YOLOv11m, YOLOv12s, and YOLOv12m, using validation mAP@50–95 as the selection criterion. The selected hyperparameter configurations for YOLOv11 and YOLOv12 are provided in Tables A1 and A2, respectively, in Appendix A. These configurations were subsequently used in the multi-seed retraining stage to assess both detection performance and run-to-run stability.

The selected detector variants were retrained on the proposed marina dataset using transfer learning with five different random seeds. Each run was trained for 220 epochs with a patience value of 40 epochs and a warm-up period of 3 epochs, while the remaining training settings followed the configuration described in Section 4. Table 5 reports the resulting mean ± standard deviation over the five runs, including validation mAP@50–95 and test-set mAP@50 and mAP@50–95.

Table 5. Detection performance of YOLO detector variants, reported as mean ± standard deviation over five training seeds.

YOLO Variant	Validation mAP@50–95	Test mAP@50	Test mAP@50–95
YOLOv11s	0.6754 ± 0.0039	0.9441 ± 0.0026	0.7347 ± 0.0038
YOLOv11m	**0.6810 ± 0.0055**	**0.9467 ± 0.0011**	**0.7458 ± 0.0027**
YOLOv12s	0.6796 ± 0.0040	0.9430 ± 0.0013	0.7348 ± 0.0032
YOLOv12m	0.6743 ± 0.0086	0.9362 ± 0.0053	0.7334 ± 0.0079

Note: Bold values indicate the highest mean for each performance metric.

YOLOv11m achieved the numerically highest mean detection performance, with the highest validation mAP@50–95 and test-set mAP@50 and mAP@50–95 values. On the test set, YOLOv11m reached mAP@50–95 of 0.7458 ± 0.0027, while YOLOv11s, YOLOv12s, and YOLOv12m formed a closely matched group, with mAP@50–95 values of 0.7347 ± 0.0038, 0.7348 ± 0.0032, and 0.7334 ± 0.0079, respectively. The relatively small standard deviations suggest stable training behavior across seeds for most evaluated variants, although

YOLOv12m shows the largest run-to-run variability. These trends are visualized in Figure 5, where the mean ± standard deviation are shown for each detection metric.

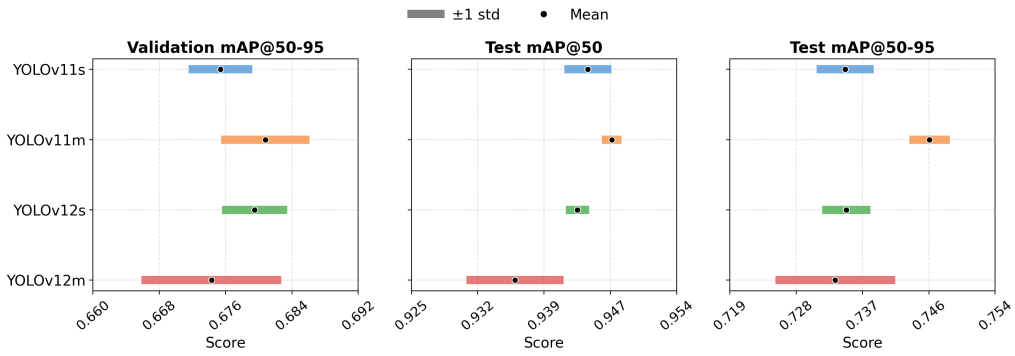


Figure 5. Mean ± standard deviation of validation and test mAP over five training seeds for YOLOv11 and YOLOv12 variants.

Because several differences among the detector variants were small, the test-set mAP@50–95 values obtained across the five training seeds were additionally compared using a repeated-measures ANOVA, with the training seed treated as the repeated-measures subject and the detector variant as the within-subject factor. The analysis revealed a statistically significant effect of detector variant on mAP@50–95 ($F(3, 12) = 6.196$, $p = 0.0087$). To identify which variants differed, Bonferroni-corrected pairwise tests were conducted. These revealed that YOLOv11m achieved a significantly higher mean test mAP@50–95 than YOLOv11s ($p = 0.043$) and YOLOv12s ($p = 0.037$). No other pairwise comparisons reached statistical significance after correction, including the comparison between YOLOv11m and YOLOv12m ($p = 0.129$). These results indicate that YOLOv11m provides a statistically meaningful improvement over YOLOv11s and YOLOv12s under the present experimental conditions, while differences among the remaining variants cannot be reliably distinguished given the available data.

To further examine the precision–recall behavior, precision–recall (PR) curves were computed at IoU thresholds 0.5 and 0.75 for each of the five training seeds and then averaged across seeds. Appendix B reports the per-seed PR curves in Figures A1 and A2, as well as the corresponding mean curves with ±1 standard deviation bands in Figures A3 and A4. The curves are tightly clustered for most configurations, indicating stable PR behavior across initializations. Variability becomes more visible at IoU = 0.75, particularly in the high-recall regime where localization precision becomes more demanding. This effect is most apparent for YOLOv12m, which is consistent with its larger standard deviation in Table 5.

Table 6 summarizes the test-set precision and recall scores averaged over five training seeds. YOLOv11m achieved the highest precision, while YOLOv11s achieved the highest recall. The differences among YOLOv11s, YOLOv11m, and YOLOv12s are small, whereas YOLOv12m produced the lowest precision and recall in this evaluation.

Table 6. Test-set precision and recall of the evaluated YOLO detector variants, averaged over five training seeds.

YOLO Variant	Precision	Recall
YOLOv11s	0.9450	**0.8635**
YOLOv11m	**0.9506**	0.8601
YOLOv12s	0.9486	0.8608
YOLOv12m	0.9423	0.8468

Note: Bold values indicate the highest mean for each performance metric.

Because fixed-camera marina scenes include both near-range vessels occupying larger image regions and far-range vessels represented by small bounding boxes, AP was additionally analyzed by object size. Following the standard COCO size thresholds implemented in pycocotools, objects were grouped into small (AP_s), medium (AP_m), and large (AP_l) categories, with areas of $<32^2$, 32^2–96^2, and $>96^2$ pixels^2, respectively. The test set contains 254 small, 434 medium, and 485 large object instances. The corresponding results are summarized in Table 7 and visualized in Figure 6.

Table 7. Test-set AP by object size for the evaluated YOLO detector variants, reported as mean ± standard deviation over five training seeds.

Metric	YOLOv11s	YOLOv11m	YOLOv12s	YOLOv12m
AP_s	0.3089 ± 0.0062	**0.3204 ± 0.0072**	0.3101 ± 0.0084	0.3004 ± 0.0169
AP_m	0.6135 ± 0.0072	**0.6289 ± 0.0053**	0.6178 ± 0.0038	0.6201 ± 0.0101
AP_l	**0.9126 ± 0.0028**	0.9111 ± 0.0032	0.9103 ± 0.0035	0.9070 ± 0.0016

Note: Bold values indicate the highest mean AP for each object-size category.

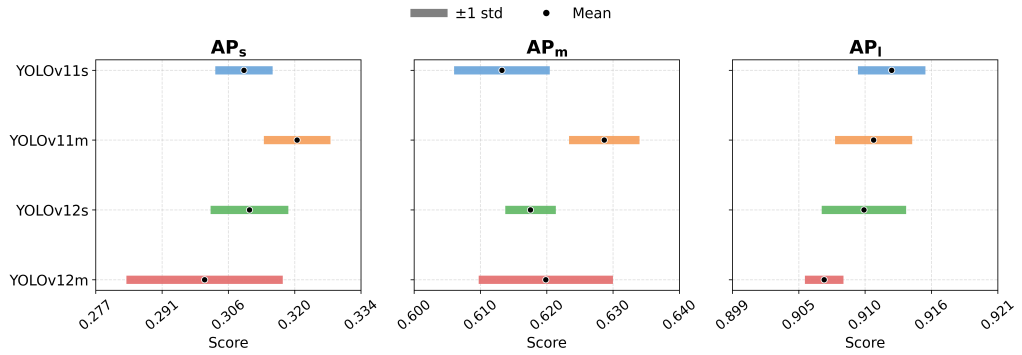


Figure 6. Mean ± standard deviation of test-set AP over five training seeds for YOLOv11 and YOLOv12 variants, stratified by object size.

The size-stratified results show that YOLOv11m achieved the highest mean AP for small and medium objects, with AP_s of 0.3204 ± 0.0072 and AP_m of 0.6289 ± 0.0053. For large objects, YOLOv11s obtained the highest mean AP_l, although the differences among detector variants are small and most standard deviation ranges overlap. This indicates that the evaluated detectors behave similarly for large vessels, while small and medium vessels remain more sensitive to model choice and training variability. YOLOv12m again shows the largest variability for small objects, consistent with the aggregate mAP results.

These results suggest that the attention-oriented design of YOLOv12 does not provide a uniform benefit across all object scales in the proposed dataset. Small vessels are represented by relatively few pixels and are frequently affected by distance, glare, low contrast, and background clutter; consequently, limited fine-grained spatial information remains available for the attention mechanisms to exploit. Under these conditions, small-object detection performance may depend more strongly on the preservation of high-resolution features and training stability, which may explain the lower AP_s and greater run-to-run variability of YOLOv12m. By contrast, the contextual feature aggregation of YOLOv12 may be more beneficial when vessels are sufficiently visible but surrounded by dense marina infrastructure, reflections, or overlapping objects, where contextual information can support object separation and more stable downstream association. For large vessels, all evaluated models receive sufficiently detailed visual information, which likely explains the narrow AP_l range and the limited measurable benefit of additional architectural mechanisms.

In addition to the fixed-IoU PR curves, Appendix B reports COCO-style PR curves aggregated over the standard IoU range from 0.5 to 0.95 in Figure A5. The narrow uncer-

tainty bands and strong overlap across seeds further support the stability of the detection evaluation across random initializations.

Qualitative examples are provided to illustrate typical detector behavior in the proposed marina dataset. Figure 7 shows detections on a representative test frame for all four detector variants. The examples indicate visually consistent localization of visible vessels under strong illumination and water reflections.

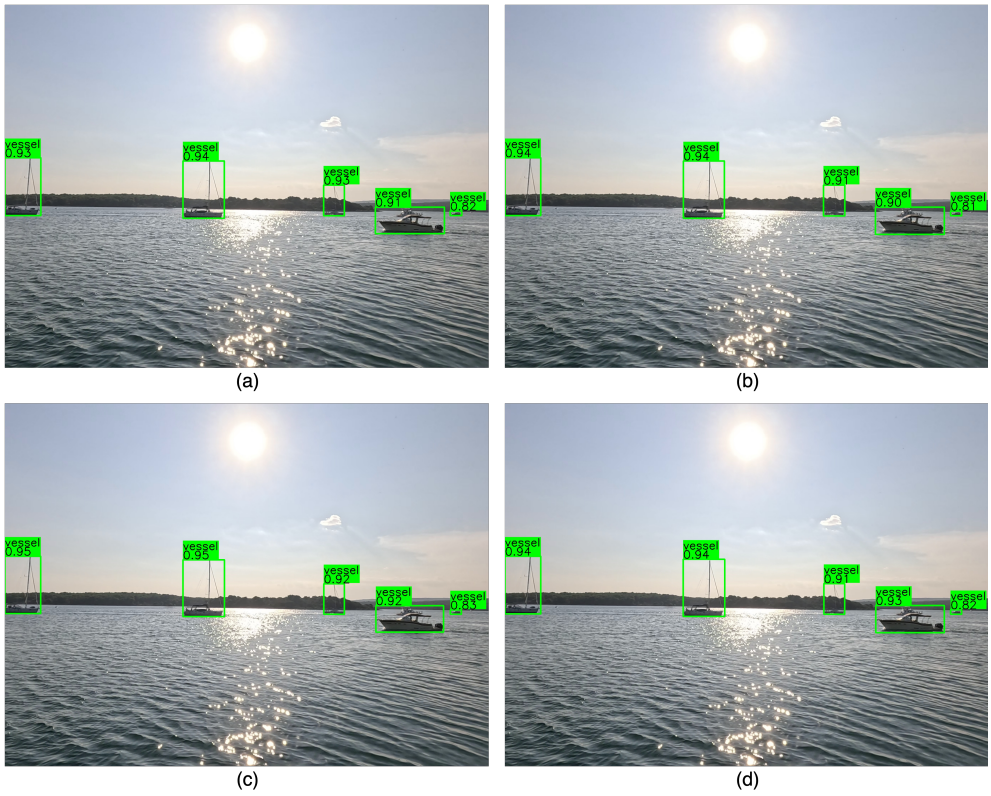


Figure 7. Qualitative detection comparison on a representative marina test frame with strong illumination and water reflections: (**a**) YOLOv11s; (**b**) YOLOv11m; (**c**) YOLOv12s; (**d**) YOLOv12m.

Representative FP and FN cases are shown in Figures 8 and 9, respectively. The FPs mainly occur near ambiguous distant structures, partial objects at image borders, or small background regions that visually resemble vessel components. FNs more often occur for distant vessels, weakly visible targets, or vessels affected by glare and low contrast. These qualitative cases are consistent with the quantitative results, where small-object detection is more challenging than large-object detection.

Overall, the detection results indicate that the evaluated YOLO variants provide reliable vessel localization within the evaluated fixed-camera marina setting, but they also show that architectural novelty alone does not guarantee superior performance in this application. Although YOLOv12 introduces a stronger attention-oriented design, YOLOv11m achieved the highest detection accuracy and the most favorable balance between precision and localization quality. This suggests that, for the present dataset scale and scene structure, detector capacity, training stability, and object-scale distribution may be more influential than the architectural family alone. From an operational perspective, the small performance differences among YOLOv11s, YOLOv11m, and YOLOv12s indicate that several detector variants may be suitable for marina monitoring, while small and distant vessels remain the most challenging cases.

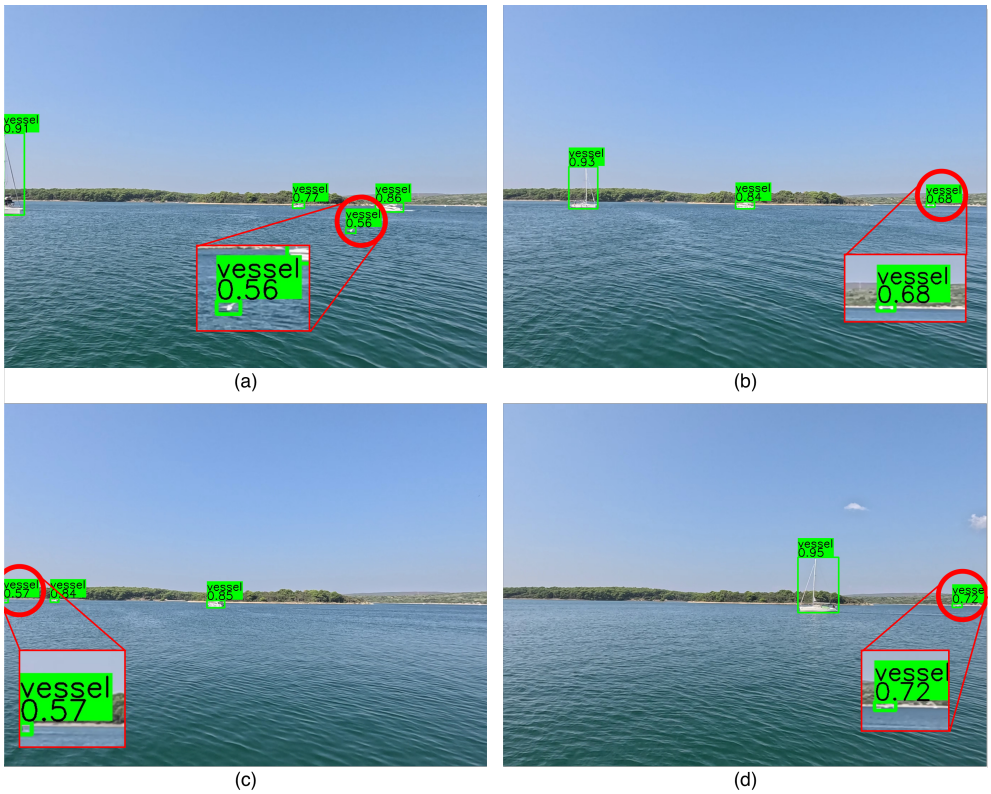


Figure 8. Representative false-positive (FP) detections in marina test frames, marked with red circles: (**a**) YOLOv11s; (**b**) YOLOv11m; (**c**) YOLOv12s; (**d**) YOLOv12m.

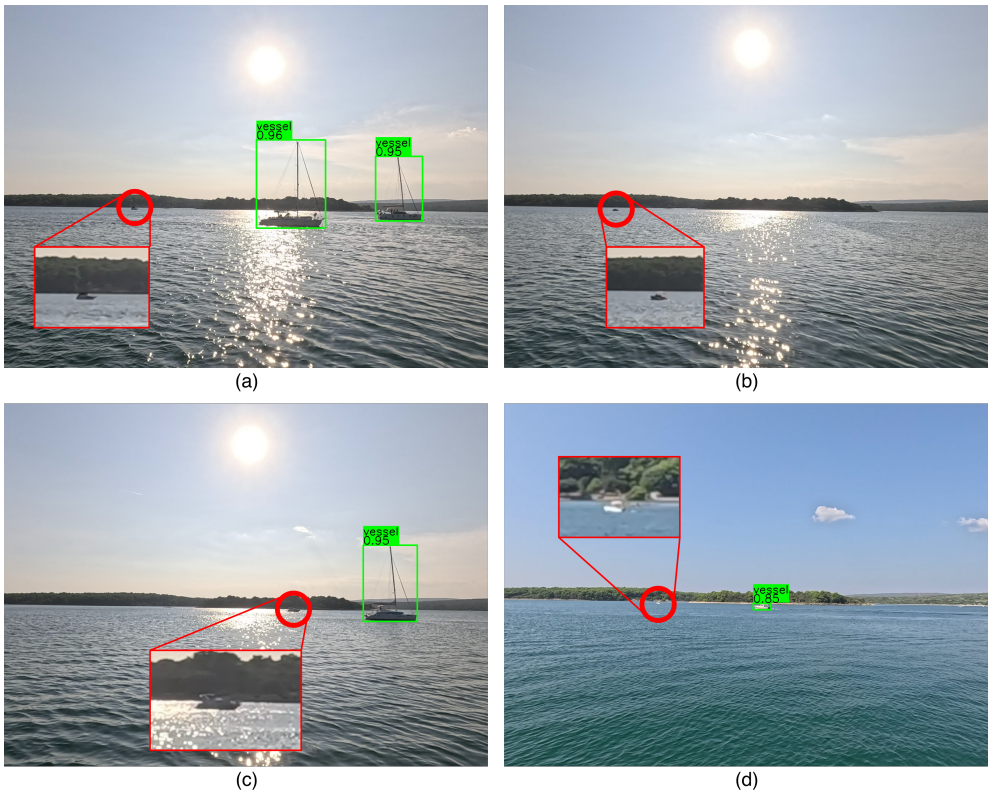


Figure 9. Representative false-negative (FN) cases in marina test frames, marked with red circles: (**a**) YOLOv11s; (**b**) YOLOv11m; (**c**) YOLOv12s; (**d**) YOLOv12m.

5.2. Multi-Object Tracking Results

Tracking experiments were conducted using the tracking-by-detection pipeline described in Section 4. Frame-level YOLO detections were used as measurements for KF- and EKF-based motion tracking, and the same data association and track-management procedure was applied in all experiments. The common tracking operating point was selected according to the calibration protocol in Section 4, using a confidence threshold of 0.75, a matching IoU threshold of 0.4, a maximum track age of 15 frames, and a minimum track confirmation length of 3 detections. An example tracking output with assigned object IDs is shown in Figure 10.

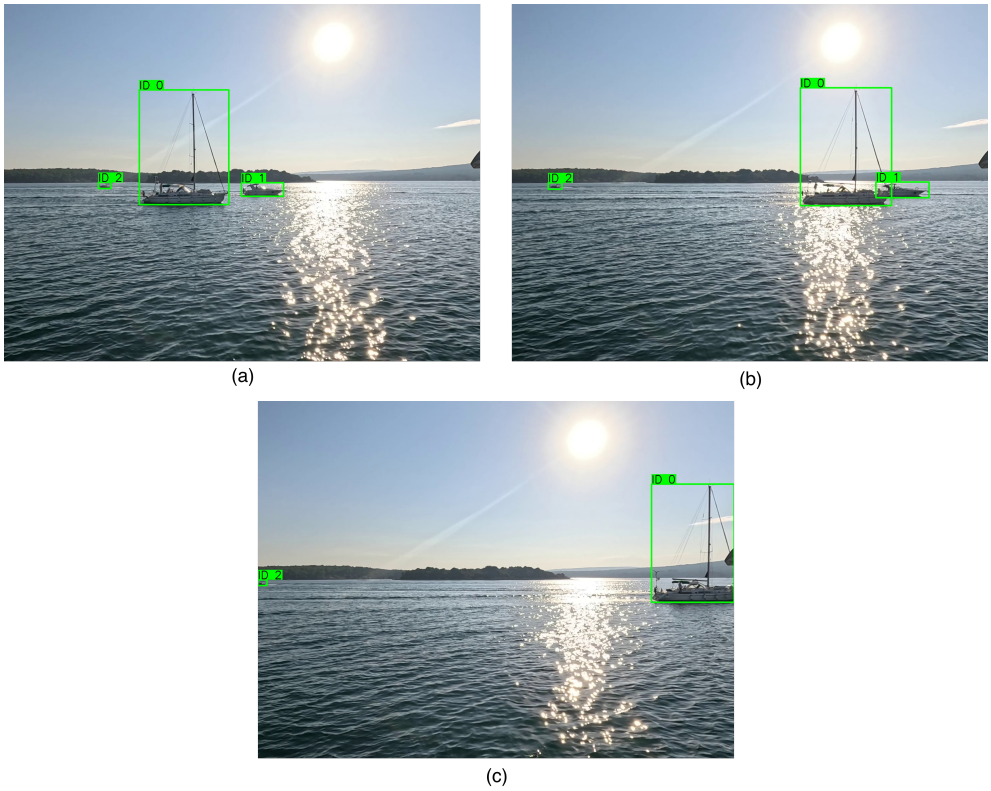


Figure 10. Example tracking output on selected test frames using YOLOv12m with KF-based motion modeling: (**a**) an earlier frame with three tracked vessels; (**b**) a subsequent frame showing maintained track IDs; (**c**) a later frame with vessels approaching the image boundaries. Assigned track IDs are shown in green.

For the final evaluation, nine held-out test sequences were used, with three sequences assigned to each difficulty level. The evaluation set covered *Easy*, *Moderate*, and *Challenging* cases, corresponding to no vessel overlap, partial vessel overlap, and complete or near-complete occlusion, respectively. In total, the tracking test set comprised 145 s of video and 3688 annotated vessel instances. Performance was evaluated using three standard MOT metrics, namely MOTA, MOTP, and IDF1, together with three HOTA metrics, namely HOTA, DetA, and AssA.

Table 8 reports the main detector-variant-level tracking summary by averaging each YOLO variant over both motion models and all difficulty levels. A more detailed supplementary analysis is provided in Appendix C: Figure A6 reports the tracking metrics separately by difficulty level for each YOLO–motion-model combination, Table A3 reports the tracing metric per each difficulty level separately, while Table A4 reports the corresponding KF/EKF results for each YOLO variant averaged over all nine test sequences. The aggregate comparison between KF and EKF motion models is then reported in Table 9.

Table 8. Tracking performance for each YOLO detector variant over nine test sequences, averaged over KF and EKF configurations, reported as mean ± standard deviation.

YOLO Variant	HOTA	DetA	AssA	MOTA	MOTP	IDF1
YOLOv11s	**0.6973 ± 0.1074**	**0.7315 ± 0.1015**	0.6774 ± 0.1396	**0.8284 ± 0.1472**	0.8582 ± 0.0371	0.7810 ± 0.1457
YOLOv11m	0.6784 ± 0.1117	0.7018 ± 0.1051	0.6686 ± 0.1460	0.7742 ± 0.1613	**0.8631 ± 0.0396**	0.7630 ± 0.1426
YOLOv12s	0.6921 ± 0.1120	0.7148 ± 0.1013	0.6799 ± 0.1373	0.7916 ± 0.1404	0.8625 ± 0.0411	0.7797 ± 0.1401
YOLOv12m	**0.6973 ± 0.1074**	0.7208 ± 0.1077	**0.6842 ± 0.1267**	0.8064 ± 0.1487	0.8584 ± 0.0400	**0.7880 ± 0.1241**

Note: Bold values indicate the highest mean for each tracking metric.

When averaged over the KF and EKF configurations, YOLOv12m and YOLOv11s achieved the joint-highest detector-variant-level HOTA score. YOLOv12m also achieved the highest AssA and IDF1 values, indicating stronger average association and identity-preservation performance in the evaluated tracking setting. In contrast, YOLOv11s achieved the highest DetA and MOTA, while YOLOv11m achieved the highest MOTP, indicating stronger localization precision. These results show that detection-sensitive tracking metrics and identity-preservation metrics do not necessarily rank detector variants in the same way. Therefore, frame-level detection performance alone is insufficient to fully characterize downstream MOT behavior in crowded marina scenes.

To assess the effect of the motion model, Table 9 reports the mean ± standard deviation performance of the KF- and EKF-based trackers, averaged over all YOLO variants and all nine test sequences. On average, the KF-based tracker achieved higher HOTA, DetA, AssA, MOTA, and IDF1 values, whereas the EKF-based tracker achieved only a marginally higher MOTP. The per-detector results reported in Appendix C further show that the relative KF/EKF performance is model- and metric-dependent; however, the aggregate results favor KF, particularly for association- and identity-related metrics.

Table 9. KF and EKF tracking performance averaged over all YOLO variants and nine test sequences, reported as mean ± standard deviation.

Metric	KF	EKF
HOTA	**0.7104 ± 0.0976**	0.6721 ± 0.1148
DetA	**0.7290 ± 0.1005**	0.7054 ± 0.1041
AssA	**0.7015 ± 0.1164**	0.6535 ± 0.1487
MOTA	**0.8208 ± 0.1411**	0.7796 ± 0.1533
MOTP	0.8595 ± 0.0393	**0.8616 ± 0.0386**
IDF1	**0.8035 ± 0.1213**	0.7524 ± 0.1460

Note: Bold values indicate the higher mean value for each tracking metric.

The tracking evaluation was intentionally limited to classical KF- and EKF-based configurations because the primary objective was to characterize the newly collected marina dataset using transparent and computationally lightweight baselines before introducing more complex MOT algorithms. This approach enables the effects of detector behavior, vessel overlap, occlusion, and motion-model selection to be examined without additional appearance models, re-identification components, or specialized association strategies. In contrast to modern trackers such as SORT, DeepSORT, ByteTrack, BoT-SORT, OC-SORT, or StrongSORT, which incorporate low-confidence detection recovery, enhanced association strategies, or appearance-based re-identification, the evaluated pipeline relies on motion prediction and IoU-based association to provide a simpler and more interpretable baseline. The results should therefore be interpreted as an initial dataset-oriented tracking analysis rather than a comprehensive comparison with state-of-the-art methods.

An examination of the tracking failures showed that identity switches were commonly observed when two or more vessels overlapped or passed in close proximity. During these events, the visible area of a vessel was temporarily reduced, while the bounding boxes

of adjacent vessels partially or substantially overlapped. This decreased the reliability of IoU-based association and increased the likelihood that a detection would be assigned to the trajectory of a neighboring vessel. The higher mean AssA and IDF1 obtained by YOLOv12m are consistent with more favorable average association and identity continuity in the evaluated tracking sequences; nevertheless, vessel overlap remained the principal source of identity-association failure across the evaluated combinations.

The finding that YOLOv11m achieved the highest mean frame-level mAP, whereas YOLOv12m achieved the highest mean AssA and IDF1, shows that detection accuracy and tracking suitability capture related but distinct properties. Frame-level mAP summarizes detection and localization quality across independently evaluated images, whereas AssA and IDF1 are additionally sensitive to the temporal continuity of detections and the preservation of object identities across consecutive frames. The higher mean association-related scores of YOLOv12m may reflect greater persistence or geometric consistency of its detections during vessel overlap, close-proximity passages, and partial occlusion. Short detection interruptions or small frame-to-frame variations in bounding-box position, dimensions, or confidence may have a limited effect on aggregate mAP but can reduce overlap with predicted tracks, fragment trajectories, or cause identity switches. From an operational perspective, detector selection should therefore follow the primary surveillance objective: frame-level accuracy remains important, but identity-sensitive tracking metrics should also be considered when continuous vessel identity is required.

The standard KF appeared to achieve a somewhat higher mean IDF1 than the EKF, with the mean decreasing from 0.8035 to 0.7524, an absolute difference of approximately 0.0511. To evaluate whether this difference was consistent across the nine test sequences and four detector variants, a two-way repeated-measures ANOVA statistical significance test was performed with tracker type and detector variant as within-sequence factors. The analysis showed a statistically significant main effect of the tracker type on IDF1 ($F(1,8) = 8.479$, $p = 0.0195$), whereas the main effect of detector variant was not statistically significant ($F(3,24) = 0.459$, $p = 0.714$). The tracker-type-by-detector interaction was also not statistically significant ($F(3,24) = 0.871$, $p = 0.470$), providing no statistical evidence that the effect of the motion model on IDF1 varied across the four YOLO variants.

Overall, the tracking results show that the detector with the highest mean frame-level mAP is not necessarily the most suitable choice for downstream MOT. Although YOLOv11m achieved the highest mean test-set mAP@50–95, YOLOv12m obtained the highest mean AssA and IDF1 when averaged across the two motion-model configurations, while YOLOv12m and YOLOv11s shared the highest mean HOTA score. This result has a direct implication for detector selection: when continuous identity preservation is an operational priority, frame-level mAP should not be used as the sole selection criterion, and detector performance should be assessed jointly with identity-sensitive tracking metrics. The results also show that motion-model selection affected identity preservation, with the two-way repeated-measures ANOVA indicating a statistically significant effect on IDF1 in favor of KF under the evaluated conditions.

To characterize the computational performance of the combined detection-and-tracking pipeline, pipeline processing rate and per-frame latency were measured for all evaluated YOLO–motion-model combinations under the hardware and software setup described in Section 4.2. Frames were processed sequentially using a batch size of 1, an input resolution of 640 × 640 pixels, FP16 precision, and CUDA acceleration. The timed section included YOLO preprocessing, detector inference, post-processing with non-maximum suppression, and tracker association and state update. Video decoding was performed before the timed section and was therefore excluded from the reported latency. Accordingly, the reported values characterize the processing time of the detection-and-tracking pipeline

rather than end-to-end video-stream latency. Peak GPU memory allocated during the timed run was also recorded. The results are summarized in Table 10.

Table 10. Computational performance of the evaluated YOLO detector–motion-model combinations under the tested hardware and software setup.

Detector	Motion Model	Pipeline Processing Rate (FPS)	Peak GPU Memory Allocated (MB)	Mean Pipeline Latency (ms/frame)
YOLOv11s	KF	**188.93**	68.48	**5.30**
YOLOv11s	EKF	188.55	68.48	5.31
YOLOv11m	KF	169.76	107.53	5.89
YOLOv11m	EKF	167.90	105.19	5.96
YOLOv12s	KF	142.79	**67.81**	7.02
YOLOv12s	EKF	145.93	**67.81**	6.86
YOLOv12m	KF	140.30	101.54	7.13
YOLOv12m	EKF	140.90	101.42	7.10

Note: Bold values indicate the best result for each computational metric.

YOLOv11s achieved the highest pipeline processing rate and lowest mean latency with both motion models, reaching 188.55–188.93 FPS and 5.30–5.31 ms/frame. At the other end of the evaluated range, YOLOv12m produced the lowest processing rate and highest latency, with 140.30–140.90 FPS and 7.10–7.13 ms/frame. The differences across detector variants were generally larger than the corresponding differences between the two motion models.

Within each detector family, the medium variants required more peak GPU memory allocation than the corresponding small variants. Differences between KF and EKF for a fixed detector were small and showed no consistent motion-model trend across processing rate, latency, or peak memory allocation. Among the evaluated configurations, YOLOv11s therefore provided the highest processing rate and lowest latency with peak memory allocation close to the minimum observed value.

The unequal distribution of object sizes in the acquired dataset represents a limitation of the present study. The detection evaluation set contains 254 small, 434 medium, and 485 large instances, and the aggregate AP results are therefore influenced more strongly by medium and large targets. The smaller number of small instances also increases the uncertainty of the corresponding AP estimate and limits the generalization of the conclusions to scenarios dominated by small or distant vessels. Such targets typically provide weaker visual features and are more likely to produce intermittent detections and unstable bounding boxes, which can increase trajectory fragmentation and identity switches. Consequently, the reported detector ranking should not be assumed to remain unchanged on a dataset containing a larger and more balanced proportion of small targets.

An additional limitation concerns the geographical, temporal, and sensor diversity of the dataset. All data were collected from a single marina, using one fixed camera, over four recording days. Although the recordings include variations in illumination, reflections, vessel scale, and occlusion, they could be more representative in terms of the full range of camera heights and viewpoints, marina layouts, vessel populations, weather conditions and seasons, sea states, background structures, and imaging sensors that may occur in other operational environments. Changes in viewing geometry, vessel scale and density, occlusion frequency, background structure, and illumination patterns may affect both absolute performance and the relative ranking of the evaluated detector–motion-model combinations. The reported results should therefore primarily be interpreted within the context of the evaluated fixed-camera marina setting and should not be generalized directly to other ports, open-water environments, or moving-camera platforms without additional

evidence. Future work will address this limitation through cross-camera, cross-location, and cross-season validation using recordings from additional maritime environments.

6. Conclusions

This work presented a controlled, feasibility-focused comparative evaluation of YOLO-driven vessel detection and MOT in a fixed-camera marina environment. A custom ground-level dataset comprising 3546 annotated images was created from approximately 30 h of video recorded over four days at a marina in the northern Adriatic, with non-overlapping train, validation, and test splits. YOLOv11s, YOLOv11m, YOLOv12s, and YOLOv12m were evaluated for vessel detection using transfer learning across five independent training seeds. Their detections were subsequently integrated into a tracking-by-detection pipeline employing KF and EKF motion models.

The detection results showed that the evaluated YOLO variants achieved closely matched performance, with YOLOv11m obtaining the highest mean detection accuracy. On the test set, YOLOv11m achieved the highest mean mAP@50–95 of 0.7458 ± 0.0027, while YOLOv11s, YOLOv12s, and YOLOv12m formed a closely matched group. A repeated-measures ANOVA confirmed a statistically significant effect of detector variant on mAP@50–95 ($p = 0.0087$), with Bonferroni-corrected pairwise tests showing that YOLOv11m significantly outperformed YOLOv11s and YOLOv12s. The object-size analysis further showed that YOLOv11m achieved the highest AP for small and medium vessels, whereas YOLOv11s achieved the highest AP for large vessels, although the differences for large objects were small. These results indicate that, in the evaluated fixed-camera marina setting, detection performance depends not only on architectural design but also on model scale, training stability, and the object-scale distribution of the target scene.

The tracking results demonstrated that frame-level detection accuracy does not directly determine downstream MOT quality. Although YOLOv11m achieved the highest mean frame-level detection performance, YOLOv11s and YOLOv12m shared the highest mean HOTA score when averaged over the KF and EKF configurations. YOLOv11s obtained the highest mean DetA and MOTA scores, YOLOv12m achieved the highest mean AssA and IDF1 values, and YOLOv11m provided the highest mean localization precision according to MOTP. These findings indicate that detector selection should reflect the operational objective. When continuous identity preservation is required, particularly in crowded marina scenes affected by vessel overlap, temporary occlusion, and background clutter, frame-level mAP should not be used as the sole selection criterion; identity-sensitive metrics such as AssA and IDF1 should also be considered. The comparison between KF and EKF motion models showed differences in aggregate tracking performance. KF achieved higher mean HOTA, DetA, AssA, MOTA, and IDF1 in this evaluation setting, whereas EKF achieved only marginally higher MOTP. A two-way repeated-measures ANOVA confirmed a statistically significant effect of the motion model on IDF1 favoring KF, whereas the interaction between motion model and detector variant was not statistically significant.

Nevertheless, the evaluation was conducted using data collected from a single marina and a single fixed camera; therefore, the reported results do not yet establish that similar absolute performance or relative model rankings would be maintained across different locations, camera viewpoints, or environmental conditions. Consequently, the observed balance between frame-level detection accuracy and identity preservation should be interpreted within the evaluated setting until it is further examined through cross-location and cross-camera validation. Future work will focus on expanding the dataset to additional marinas and ports and incorporating multiple camera heights, viewing angles, seasonal and weather conditions, and vessel densities. This extension will enable cross-location evaluation and a more rigorous assessment of domain shift and model generalizability. In

parallel, appearance-based association, re-identification features, and adaptive tracking strategies will be investigated to improve identity preservation under heavy occlusion and challenging illumination.

Author Contributions: Conceptualization, N.L., K.S., N.G. and J.L.; methodology, N.L., K.S. and J.L.; software, N.L., K.S. and J.L.; validation, N.L., K.S. and J.L.; formal analysis, N.L., K.S. and J.L.; investigation, N.L., K.S. and J.L.; resources, N.L., N.G. and J.L.; data curation, N.L., K.S. and J.L.; writing—original draft preparation, N.L. and K.S.; writing—review and editing, N.L., K.S., N.G. and J.L.; visualization, K.S.; supervision, N.L., N.G. and J.L.; project administration, N.L., N.G. and J.L.; funding acquisition, N.L., N.G. and J.L. All authors have read and agreed to the published version of the manuscript.

Funding: Funded by the European Union—NextGeneration EU—SEALEARN–uniri-mz-25-49. The work was also supported by the European Regional Development Fund, grant no. PK.1.1.13.0083 (Collaborative Scientific Research Aimed at Increasing Coastal Traffic Safety Using Machine Learning) and the Operational Program NPOO under the project NPOO.C3.2.R3-I1.01.0124 funded by the European Union—NextGeneration EU.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data supporting the findings of this study are available from the corresponding author upon reasonable request, subject to applicable privacy, GDPR, and data-governance constraints related to potentially identifiable vessel imagery.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Supplementary YOLO Hyperparameter Configuration for Vessel Detection

Table A1. Tuned hyperparameters for YOLOv11 variants selected using validation mAP@50–95.

Parameter	YOLOv11s	YOLOv11m
lr0	0.00149	0.00131
lrf	0.10963	0.1
weight_decay	0.0006	0.00035
warmup_momentum	0.6	0.68642
momentum	0.5	0.49769
box	7.55948	8.45128
cls	0.45068	0.4754
dfl	1.5	1.13067
hsv_h	0.02478	0.01004
hsv_s	0.39947	0.33106
hsv_v	0.3	0.3
degrees	0.0	0.0
translate	0.08229	0.08958
scale	0.18003	0.14389
shear	0.0	0.0
perspective	0.0	0.0
fliplr	0.4	0.46985
mosaic	0.98165	0.87868
mixup	0.0	0.0
close_mosaic	9	10

Table A2. Tuned hyperparameters for YOLOv12 variants selected using validation mAP@50–95.

Parameter	YOLOv12s	YOLOv12m
lr0	0.00154	0.00126
lrf	0.1068	0.10982
weight_decay	0.00034	0.00047
warmup_momentum	0.6	0.75501
momentum	0.44293	0.16994
box	7.4187	7.71626
cls	0.34791	0.50478
dfl	1.07164	1.487
hsv_h	0.01394	0.01498
hsv_s	0.35864	0.36091
hsv_v	0.28643	0.2516
degrees	0.0	0.0
translate	0.05247	0.08029
scale	0.16893	0.17325
shear	0.0	0.0
perspective	0.0	0.0
fliplr	0.4	0.44398
mosaic	0.99634	0.99155
mixup	0.0	0.0
close_mosaic	9	9

Appendix B. Supplementary Precision–Recall Analysis for Vessel Detection

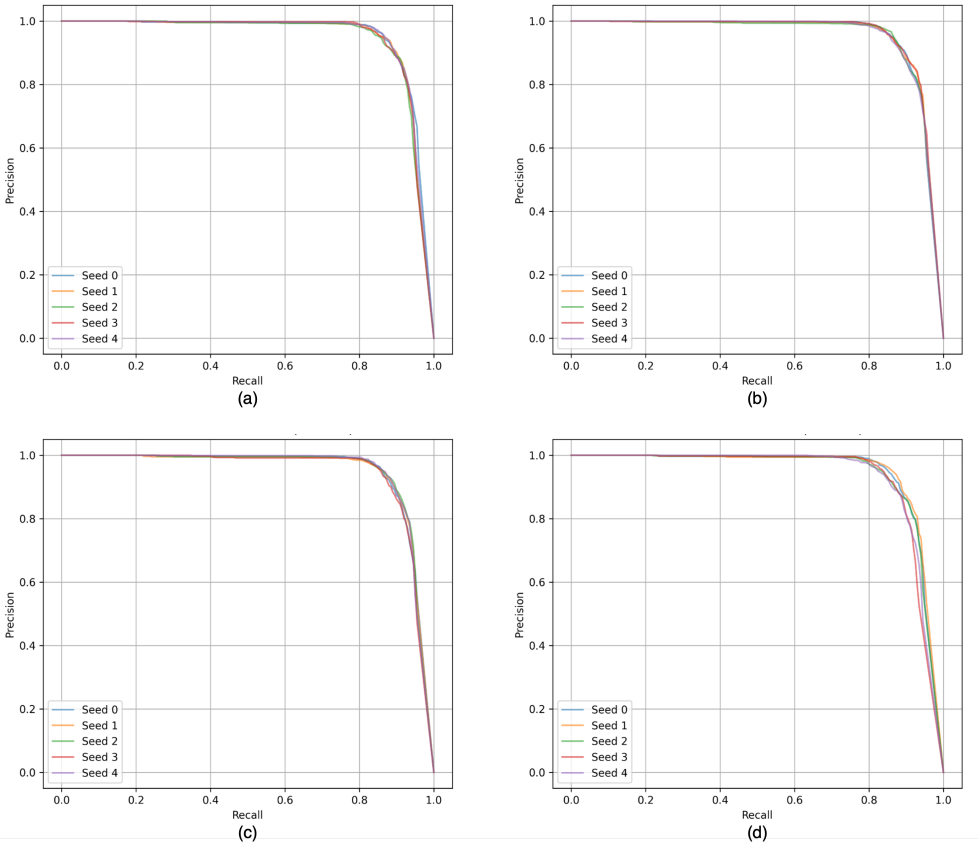


Figure A1. Per-seed precision–recall (PR) curves at IoU threshold 0.5 for the evaluated detectors: (**a**) YOLOv11s; (**b**) YOLOv11m; (**c**) YOLOv12s; (**d**) YOLOv12m.

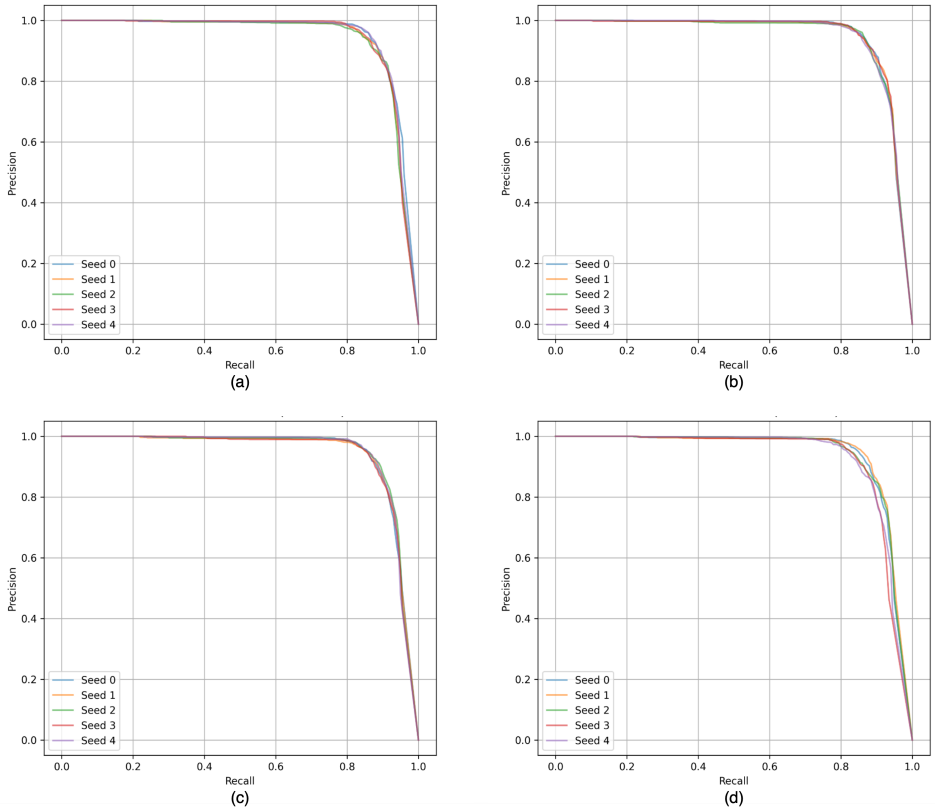


Figure A2. Per-seed PR curves at IoU threshold 0.75 for the evaluated detectors: (**a**) YOLOv11s; (**b**) YOLOv11m; (**c**) YOLOv12s; (**d**) YOLOv12m.

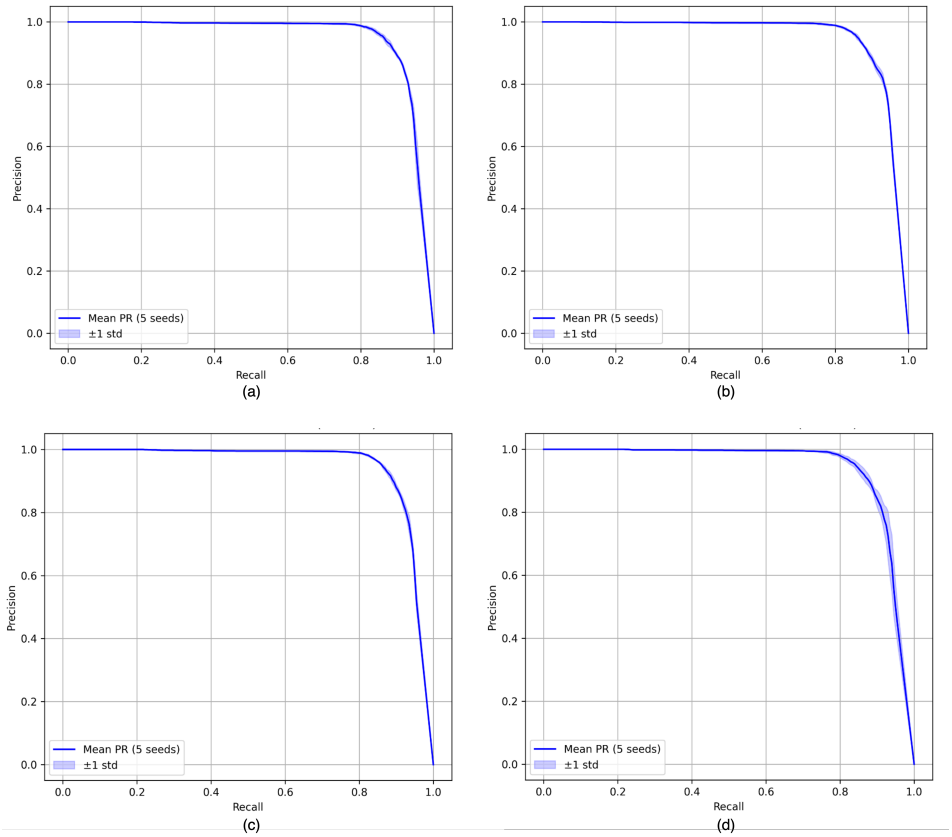


Figure A3. Mean PR curves at IoU threshold 0.5 with ±1 standard deviation over five training seeds: (**a**) YOLOv11s; (**b**) YOLOv11m; (**c**) YOLOv12s; (**d**) YOLOv12m.

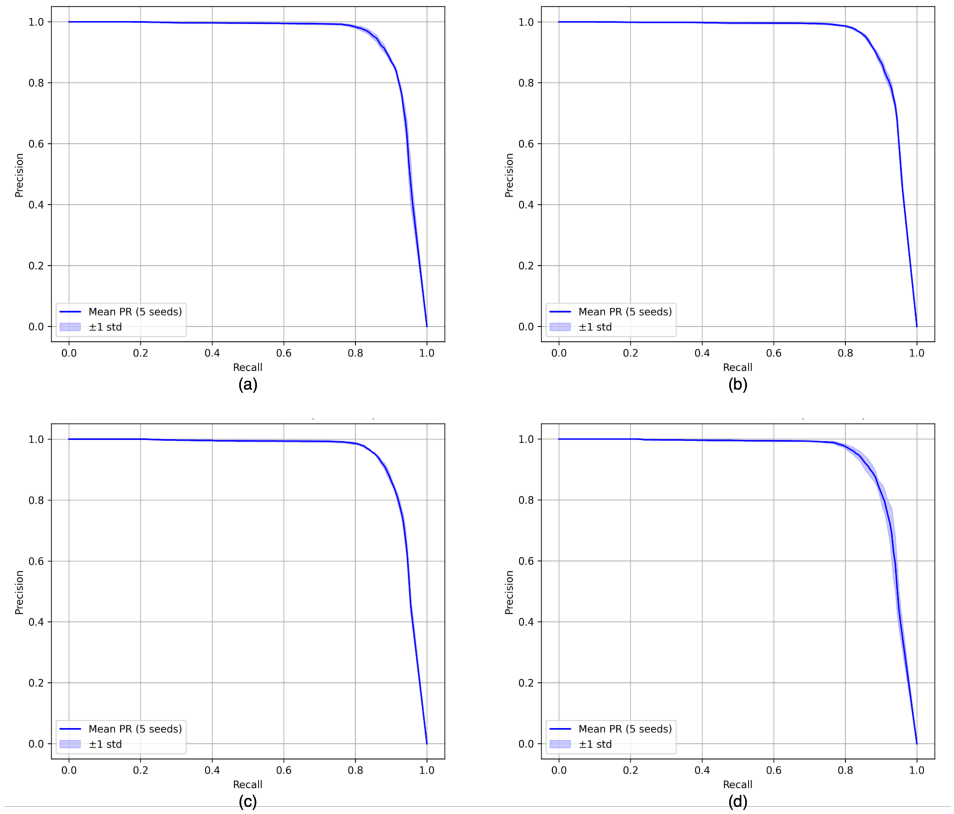


Figure A4. Mean PR curves at IoU threshold 0.75 with ±1 standard deviation over five training seeds: (**a**) YOLOv11s; (**b**) YOLOv11m; (**c**) YOLOv12s; (**d**) YOLOv12m.

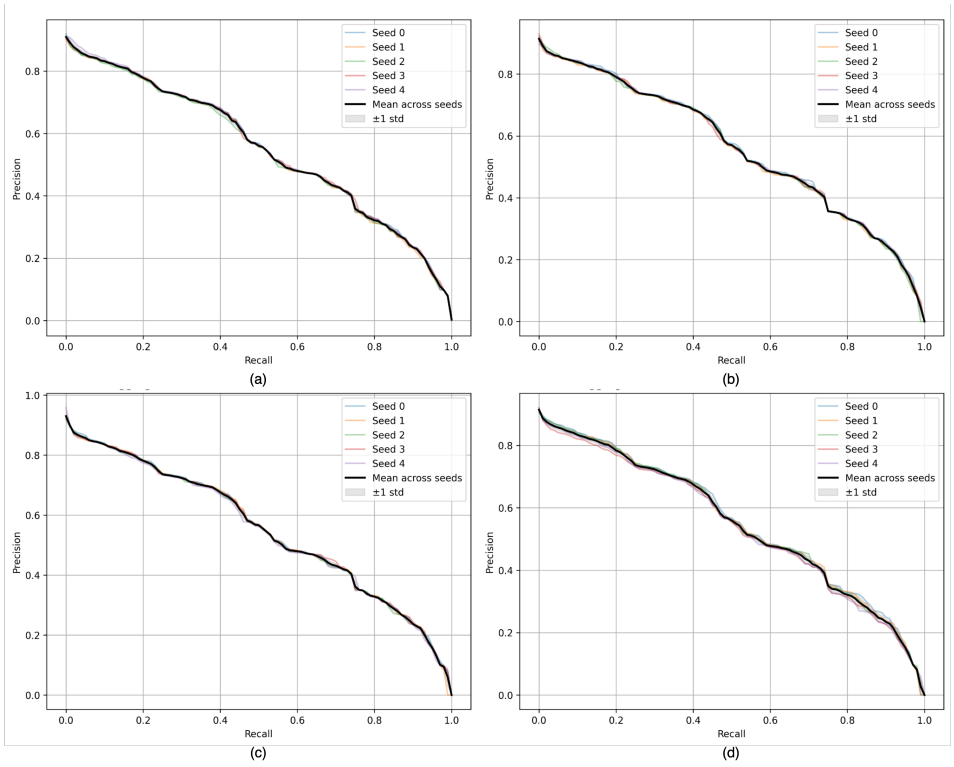


Figure A5. COCO-style PR curves aggregated over IoU thresholds from 0.5 to 0.95, shown across five training seeds with the mean curve and ±1 standard deviation: (**a**) YOLOv11s; (**b**) YOLOv11m; (**c**) YOLOv12s; (**d**) YOLOv12m.

Appendix C. Supplementary Tracking Analysis by Difficulty Level and Motion Model

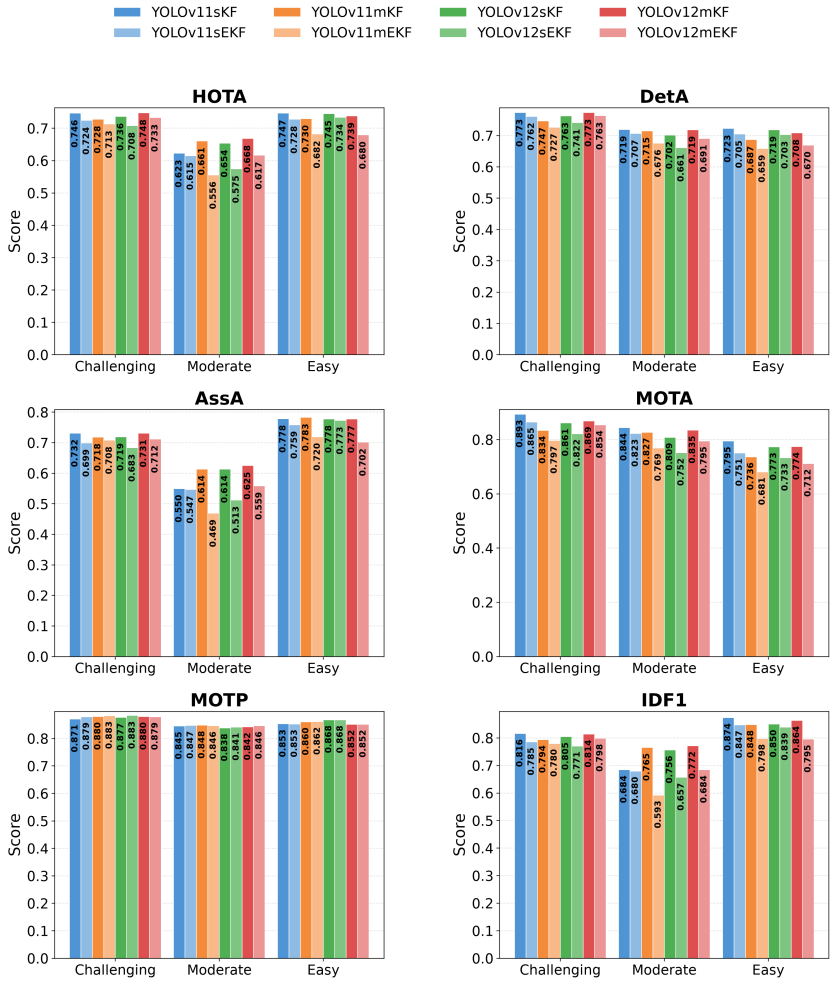


Figure A6. Tracking performance of KF and EKF motion models for each YOLO variant, averaged over the test sequences within each difficulty level.

Table A3. Tracking performance for each YOLO detector variant across *Easy*, *Moderate*, and *Challenging* test settings, reported as mean ± standard deviation.

YOLO Variant	HOTA	DetA	AssA	MOTA	MOTP	IDF1
			Easy setting			
YOLOv11s	0.7371 ± 0.1274	0.7139 ± 0.1682	0.7685 ± 0.0765	0.7728 ± 0.2477	0.8531 ± 0.0437	0.8605 ± 0.1451
YOLOv11m	0.7060 ± 0.1168	0.6731 ± 0.1625	0.7513 ± 0.0880	0.7088 ± 0.2452	0.8610 ± 0.0367	0.8232 ± 0.1272
YOLOv12s	0.7394 ± 0.1335	0.7108 ± 0.1631	0.7752 ± 0.0954	0.7530 ± 0.2317	0.8679 ± 0.0483	0.8447 ± 0.1619
YOLOv12m	0.7091 ± 0.1319	0.6890 ± 0.1683	0.7395 ± 0.1067	0.7430 ± 0.2372	0.8518 ± 0.0448	0.8297 ± 0.1492
			Moderate setting			
YOLOv11s	0.6192 ± 0.1049	0.7132 ± 0.0651	0.5484 ± 0.1582	0.8334 ± 0.0560	0.8465 ± 0.0376	0.6820 ± 0.1539
YOLOv11m	0.6085 ± 0.1302	0.6955 ± 0.0778	0.5415 ± 0.1738	0.7983 ± 0.0597	0.8470 ± 0.0450	0.6790 ± 0.1618
YOLOv12s	0.6144 ± 0.1113	0.6816 ± 0.0681	0.5632 ± 0.1558	0.7803 ± 0.0604	0.8395 ± 0.0415	0.7065 ± 0.1442
YOLOv12m	0.6425 ± 0.1171	0.7050 ± 0.0795	0.5918 ± 0.1505	0.8148 ± 0.0860	0.8440 ± 0.0454	0.7281 ± 0.1198
			Challenging setting			
YOLOv11s	0.7355 ± 0.0319	0.7674 ± 0.0130	0.7154 ± 0.0611	0.8791 ± 0.0486	0.8751 ± 0.0286	0.8005 ± 0.0873
YOLOv11m	0.7207 ± 0.0540	0.7367 ± 0.0511	0.7131 ± 0.0678	0.8156 ± 0.1299	0.8814 ± 0.0353	0.7869 ± 0.1155
YOLOv12s	0.7224 ± 0.0369	0.7521 ± 0.0254	0.7011 ± 0.0542	0.8416 ± 0.0690	0.8800 ± 0.0252	0.7878 ± 0.0903
YOLOv12m	0.7403 ± 0.0459	0.7684 ± 0.0230	0.7213 ± 0.0706	0.8615 ± 0.0545	0.8795 ± 0.0234	0.8061 ± 0.0947

Table A4. Tracking performance of KF and EKF motion models for each YOLO variant, averaged over nine test sequences, reported as mean ± standard deviation.

YOLO Variant	HOTA	DetA	AssA	MOTA	MOTP	IDF1
KF motion model						
YOLOv11s	0.7053 ± 0.1114	0.7385 ± 0.1022	0.6867 ± 0.1483	0.8441 ± 0.1381	0.8566 ± 0.0379	0.7913 ± 0.1526
YOLOv11m	0.7064 ± 0.0960	0.7164 ± 0.1076	0.7050 ± 0.1088	0.7992 ± 0.1652	0.8626 ± 0.0404	0.8024 ± 0.1141
YOLOv12s	0.7120 ± 0.0978	0.7277 ± 0.0982	0.7035 ± 0.1115	0.8142 ± 0.1325	0.8609 ± 0.0427	0.8036 ± 0.1196
YOLOv12m	0.7181 ± 0.1020	0.7335 ± 0.1104	0.7120 ± 0.1122	0.8257 ± 0.1483	0.8578 ± 0.0428	0.8165 ± 0.1159
EKF motion model						
YOLOv11s	0.6892 ± 0.1093	0.7245 ± 0.1064	0.6682 ± 0.1387	0.8128 ± 0.1626	0.8599 ± 0.0384	0.7707 ± 0.1468
YOLOv11m	0.6504 ± 0.1246	0.6872 ± 0.1069	0.6323 ± 0.1746	0.7493 ± 0.1630	0.8636 ± 0.0412	0.7236 ± 0.1635
YOLOv12s	0.6722 ± 0.1272	0.7019 ± 0.1085	0.6562 ± 0.1624	0.7691 ± 0.1523	0.8641 ± 0.0419	0.7557 ± 0.1617
YOLOv12m	0.6765 ± 0.1146	0.7081 ± 0.1100	0.6574 ± 0.1411	0.7871 ± 0.1554	0.8590 ± 0.0397	0.7594 ± 0.1321

References

1. Voulodimos, A.; Doulamis, N.; Doulamis, A.; Protopapadakis, E. Deep Learning for Computer Vision: A Brief Review. *Comput. Intell. Neurosci.* **2018**, *2018*, 7068349. [CrossRef] [PubMed]
2. Munim, Z.H.; Dushenko, M.; Jimenez, V.J.; Shakil, M.H.; Imset, M. Big data and artificial intelligence in the maritime industry: A bibliometric review and future research directions. *Marit. Policy Manag.* **2020**, *47*, 577–597. [CrossRef]
3. Lopac, N.; Severinski, K.; Palaić, D.; Lerga, J. Machine learning for ship-motion-based sea-state estimation. *Ocean Eng.* **2026**, *362*, 126362. [CrossRef]
4. Yilmaz, B. Cybersecurity Applications in the Maritime Industry: A Scientometric Approach. *Sci. J. Marit. Res.-Pomor.* **2025**, *39*, 254–278. [CrossRef]
5. Galić, S.; Lušić, Z.; Mladenović, S. The prediction of ship groundings: A review. *Sci. J. Marit. Res.-Pomor.* **2025**, *39*, 324–339. [CrossRef]
6. Lorencin, I.; Frank, D.; Vusić, D. Zero-Shot Learning in Maritime Domain: Classification of Marine Objects using CLIP. *Sci. J. Marit. Res.-Pomor.* **2024**, *38*, 239–249. [CrossRef]
7. Zou, Z.; Chen, K.; Shi, Z.; Guo, Y.; Ye, J. Object Detection in 20 Years: A Survey. *Proc. IEEE* **2023**, *111*, 257–276. [CrossRef]
8. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 4–9 December 2017.
9. Lopac, N.; Severinski, K.; Krljan, T.; Lerga, J. Computer Vision-Based Real-Time Vessel Detection and Tracking in Marinas. In Proceedings of the 17th Baška GNSS Conference: Global Navigation Satellite Systems and Green Navigation and Smart Systems, Baška, Croatia, 11–15 May 2025; pp. 131–136.
10. Zhao, W.; Syafrudin, M.; Fitriyani, N.L. CRAS-YOLO: A Novel Multi-Category Vessel Detection and Classification Model Based on YOLOv5s Algorithm. *IEEE Access* **2023**, *11*, 11463–11478. [CrossRef]
11. Hou, X.; Ao, W.; Song, Q.; Lai, J.; Wang, H.; Xu, F. FUSAR-ship: Building a high-resolution SAR-AIS matchup dataset of Gaofen-3 for ship detection and recognition. *Sci. China Inf. Sci.* **2020**, *63*, 140303. [CrossRef]
12. Ghiasi, G.; Cui, Y.; Srinivas, A.; Qian, R.; Lin, T.Y.; Cubuk, E.D.; Le, Q.V.; Zoph, B. Simple Copy-Paste is a Strong Data Augmentation Method for Instance Segmentation. In *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021*; IEEE: New York, NY, USA, 2021; pp. 2917–2927. [CrossRef]
13. Li, J.; Qu, C.; Shao, J. Ship detection in SAR images based on an improved faster R-CNN. In *Proceedings of the 2017 SAR in Big Data Era: Models, Methods and Applications (BIGSARDATA), Beijing, China, 13–14 November 2017*; IEEE: New York, NY, USA, 2017; pp. 1–6. [CrossRef]
14. Jiang, J.; Fu, X.; Qin, R.; Wang, X.; Ma, Z. High-Speed Lightweight Ship Detection Algorithm Based on YOLO-V4 for Three-Channels RGB SAR Image. *Remote Sens.* **2021**, *13*, 1909. [CrossRef]
15. Li, L.; Jiang, L.; Zhang, J.; Wang, S.; Chen, F. A Complete YOLO-Based Ship Detection Method for Thermal Infrared Remote Sensing Images under Complex Backgrounds. *Remote Sens.* **2022**, *14*, 1534. [CrossRef]
16. Shao, J.; Yang, Q.; Luo, C.; Li, R.; Zhou, Y.; Zhang, F. Vessel Detection From Nighttime Remote Sensing Imagery Based on Deep Learning. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 12536–12544. [CrossRef]
17. Vasilopoulos, E.; Vosinakis, G.; Krommyda, M.; Karagiannidis, L.; Ouzounoglou, E.; Amditis, A. A Comparative Study of Autonomous Object Detection Algorithms in the Maritime Environment Using a UAV Platform. *Computation* **2022**, *10*, 42. [CrossRef]
18. Ribeiro, R.; Cruz, G.; Matos, J.; Bernardino, A. A Data Set for Airborne Maritime Surveillance Environments. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *29*, 2720–2732. [CrossRef]

19. Prayudi, A.; Sulistijono, I.A.; Risnumawan, A.; Darojah, Z. Surveillance System for Illegal Fishing Prevention on UAV Imagery Using Computer Vision. In *Proceedings of the 2020 International Electronics Symposium (IES), Surabaya, Indonesia, 29–30 September 2020*; IEEE: New York, NY, USA, 2020; pp. 385–391. [CrossRef]
20. Li, H.; Deng, L.; Yang, C.; Liu, J.; Gu, Z. Enhanced YOLO V3 Tiny Network for Real-Time Ship Detection From Visual Image. *IEEE Access* **2021**, *9*, 16692–16706. [CrossRef]
21. Shao, Z.; Wu, W.; Wang, Z.; Du, W.; Li, C. SeaShips: A Large-Scale Precisely Annotated Dataset for Ship Detection. *IEEE Trans. Multimed.* **2018**, *20*, 2593–2604. [CrossRef]
22. Kim, J.H.; Kim, N.; Park, Y.W.; Won, C.S. Object Detection and Classification Based on YOLO-V5 with Improved Maritime Dataset. *J. Mar. Sci. Eng.* **2022**, *10*, 377. [CrossRef]
23. Prasad, D.K.; Rajan, D.; Rachmawati, L.; Rajabally, E.; Quek, C. Video Processing From Electro-Optical Sensors for Object Detection and Tracking in a Maritime Environment: A Survey. *IEEE Trans. Intell. Transp. Syst.* **2017**, *18*, 1993–2016. [CrossRef]
24. Wang, Y.; Song, J.; Wang, Y.; Wang, R.; Chen, H. Object Detection Method of Inland Vessel Based on Improved YOLO. *J. Mar. Sci. Eng.* **2025**, *13*, 697. [CrossRef]
25. Sun, X.; Liu, T.; Yu, X.; Pang, B. Unmanned Surface Vessel Visual Object Detection Under All-Weather Conditions with Optimized Feature Fusion Network in YOLOv4. *J. Intell. Robot. Syst.* **2021**, *103*, 55. [CrossRef]
26. Petković, M.; Kezić, D.; Vujović, I.; Pavić, I. Target Detection For Visual Collision Avoidance System. *Pedagogika* **2021**, *93*, 159–166. [CrossRef]
27. Petković, M.; Vujović, I.; Kaštelan, N.; Šoda, J. Every Vessel Counts: Neural Network Based Maritime Traffic Counting System. *Sensors* **2023**, *23*, 6777. [CrossRef] [PubMed]
28. Petković, M.; Vujović, I. Distance Estimation Approach for Maritime Traffic Surveillance Using Instance Segmentation. *J. Mar. Sci. Eng.* **2023**, *12*, 78. [CrossRef]
29. Petković, M.; Vujović, I.; Lušić, Z.; Šoda, J. Image dataset for neural network performance estimation with application to maritime ports. *J. Mar. Sci. Eng.* **2023**, *11*, 578. [CrossRef]
30. Miličević, M.; Vujović, I.; Petković, M.; Kuzmanić Skelin, A. Toward an Augmented Reality Representation of Collision Risks in Harbors. *Appl. Sci.* **2025**, *15*, 9260. [CrossRef]
31. Jocher, G.; Qiu, J.; Chaurasia, A. Ultralytics YOLO; Ultralytics. 2023. Available online: https://github.com/ultralytics/ultralytics (accessed on 27 July 2026).
32. Khanam, R.; Hussain, M. YOLOv11: An Overview of the Key Architectural Enhancements. *arXiv* **2024**, arXiv:2410.17725. [CrossRef]
33. Tian, Y.; Ye, Q.; Doermann, D. YOLOv12: Attention-Centric Real-Time Object Detectors. In Proceedings of the 39th Conference on Neural Information Processing Systems (NeurIPS), San Diego, CA, USA, 2–7 December 2025; pp. 78433–78457
34. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929. [CrossRef]
35. Luo, W.; Xing, J.; Milan, A.; Zhang, X.; Liu, W.; Kim, T.K. Multiple object tracking: A literature review. *Artif. Intell.* **2021**, *293*, 103448. [CrossRef]
36. Hassan, S.; Mujtaba, G.; Rajput, A.; Fatima, N. Multi-object tracking: A systematic literature review. *Multimed. Tools Appl.* **2024**, *83*, 43439–43492. [CrossRef]
37. Welch, G.; Bishop, G. *An Introduction to the Kalman Filter*; Technical Report TR 95-041; Department of Computer Science, University of North Carolina at Chapel Hill: Chapel Hill, NC, USA, 1995.
38. Wada, K. Labelme: Image Polygonal Annotation with Python. 2021. Available online: https://zenodo.org/records/5711226 (accessed on 27 July 2026).
39. CVAT.ai Corporation. *Computer Vision Annotation Tool (CVAT)*; CVAT.ai Corporation: Wilmington, DE, USA, 2023. [CrossRef]
40. Luiten, J.; Hoffhues, A. TrackEval. 2020. Available online: https://github.com/JonathonLuiten/TrackEval (accessed on 27 July 2026).
41. Bernardin, K.; Stiefelhagen, R. Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics. *EURASIP J. Image Video Process.* **2008**, *2008*, 246309. [CrossRef]
42. Luiten, J.; Osep, A.; Dendorfer, P.; Torr, P.; Geiger, A.; Leal-Taixé, L.; Leibe, B. HOTA: A Higher Order Metric for Evaluating Multi-object Tracking. *Int. J. Comput. Vis.* **2020**, *129*, 548–578. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.