

## RESEARCH ARTICLE

# Learning-Based Online Tracking Algorithms for Marine Litter in Multibeam Water Column Images

PEDRO ALVES GUEDES<sup>1</sup>, HUGO MIGUEL SILVA<sup>1</sup>, AND SEN WANG<sup>2</sup>, (Member, IEEE)

<sup>1</sup>Institute for Systems and Computer Engineering, Technology and Science (INESC TEC), 4200-465 Porto, Portugal

<sup>2</sup>Imperial College London, Imperial College of Science, Technology and Medicine, South Kensington, SW7 2AZ London, U.K.

Corresponding author: Pedro Alves Guedes (pedro.e.guedes@inesctec.pt)

This work is financed by National Funds through the Portuguese funding agency, FCT—Fundação para a Ciência e a Tecnologia with the Studentship for Doctoral Research Funding Programme. This work is funded by national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., under the support UID/50014/2025 and by the European Union under the Horizon Europe Program, Grant No. 101112812 (NETTAG+).

**ABSTRACT** Marine litter is a growing environmental threat, with severe ecological and socio-economic impacts. Most monitoring strategies rely on optical sensors to detect surface pollution, however these approaches fail to capture submerged plastics dispersed throughout the water column. Multibeam acoustic imaging offers a complementary solution, but the scarcity of annotated sonar datasets and the high noise levels of acoustic imagery make automated detection and tracking particularly challenging. This study presents a comparative evaluation of deep learning based multi-object tracking (MOT) algorithms applied to water column acoustic data. Pre-trained YOLOv8 detectors were integrated with tracking-by-detection frameworks including BoT-SORT, OC-SORT, ByteTrack, and DeepOC-SORT. Performance was assessed across acoustic frequencies and preprocessing strategies using standard MOT metrics. Results show that adaptive Gaussian thresholding and opening morphology improved robustness at lower frequencies (950 kHz and 1200 kHz), while unprocessed inputs proved more resilient to severe clutter at 1400 kHz. BoostTrack and ByteTrack achieved the most consistent tracking, effectively managing intermittent detections to maximise MOTA and IDF1. In contrast, OC-SORT underperformed, struggling with fragmented sonar trajectories. Furthermore, while efficient Nano models dominated at lower frequencies, Medium models were required under higher noise. These findings demonstrate the feasibility of applying MOT methods to sonar-based litter monitoring. Future work will explore unsupervised learning approaches to leverage intrinsic sonar data structure, reduce annotation needs, and enable scalable marine litter tracking.

**INDEX TERMS** Acoustic imaging, acoustic frequencies, autonomous surface vehicle, detection algorithms, marine litter, multibeam echosounder, multi-object tracking, ocean sensing, sonar.

## I. INTRODUCTION

Marine litter is a growing threat to marine ecosystems and to global sustainability, it is estimated that 73 million tons of plastic will enter the oceans annually by 2030 [1]. New technologies are being introduced for more efficient and sustainable plastic degradation, namely enzyme engineering, however, this type of solution will be useful after the marine litter is detected [2]. As global attention to sustainability grows, industries are increasingly making environmental

The associate editor coordinating the review of this manuscript and approving it for publication was Li Zhang<sup>1</sup>.

claims that are not always substantiated. A large number of these claims constitute greenwashing, where vague or misleading language is used to promote a false image of sustainability [3]. In this context, the development of transparent, sensor-based monitoring systems becomes even more critical.

Most detection methods focus on surface-level litter, relying on satellites, drones, and aircraft remote sensing. Examples of said solutions use hyperspectral imaging for the detection of marine litter and its classification [4]. Regions such as the North Pacific Garbage Patch, are not only hotspots for plastic accumulation, but also critical

habitats for surface-dwelling organisms, such as the case of neustons, whose distribution and survival are related to plastic dynamics [5], [6]. The usage of underwater cameras can allow to capture detailed images of marine debris, suffering, however, from attenuation due to the underwater medium, which compromises the image field of view, range and quality. Acoustic waves, being mechanical waves, suffer from less attenuation and can spread over a wider range [7].

A recent review of 80 studies examined the application of artificial intelligence (AI) to marine macrolitter research [8], covering classification, detection, segmentation, and quantification tasks across aquatic environments, including beached, floating, and seafloor litter. The review identified a limited number of sonar-based datasets, particularly for floating litter in the water column. Sonar imagery represented only 3.7% of the dataset types, and just two studies used sonar for detecting floating or seafloor litter [9], [10], revealing a gap in current monitoring efforts, which rely mostly on optical data.

Beyond this review, some datasets have been published for underwater object detection. The Underwater Acoustic Target Detection (UATD) dataset includes over 9,000 multibeam forward-looking sonar images, annotated with 10 object categories such as cubes, cylinders, and tyres, either suspended in the water or resting on the lakebed [11]. Other datasets include sonar fish classification with eight classes [12], and a semantic segmentation dataset with eleven classes of household debris and marine objects [13].

In our previous work [14], we applied multibeam acoustic imaging to detect different types of marine litter in the water column, achieving material-level differentiation using multiple acoustic frequencies. Building on this, we developed learning based classification and multi-label detection models, which performed well in controlled tests [15].

The main contributions of this paper include:

- A domain-specific benchmarking protocol for acoustic Tracking-by-detection (TBD): We formulate the under-explored problem of tracking marine litter in multibeam water-column sonar into a reproducible tracking-by-detection setting, providing a structured evaluation protocol for underwater acoustic tracking.
- Acoustic frequency-aware methodology and transferable insights: We systematically analyse the impact of frequency selection and signal preprocessing on detection and tracking stability. We provide transferable engineering insights regarding optimal You Only Look Once (YOLO)v8 model capacities (nano vs. medium variants) and preprocessing strategies depending on the operational acoustic noise levels.
- Benchmarking state-of-the-art multi-object tracking (MOT) algorithms for sonar data: This study presents the first comparative evaluation of leading deep learning-based Simple Online and Real-time Tracking (SORT) algorithms (BoostTrack, Bag of Tricks SORT (BoT-SORT), ByteTrack, Deep Observation-centric SORT (DeepOC-SORT), and Observation-centric

SORT (OC-SORT)) applied specifically to multibeam acoustic data, providing practical selection guidance based on algorithm behaviour under severe acoustic clutter.

The remainder of this paper is organised as follows. Section II provides an overview of related work in marine litter detection and tracking methodologies, introducing the state-of-the-art tracking algorithms selected for this study. Section III presents the experimental setup that was used to extract the dataset, acoustic video generation and pre-processing. Section IV outlines the tracking pipeline developed for this work, including the detection phase, data association with tracker integration, along with the evaluation metrics used to assess their performance. Section V presents the experimental results across different combinations of detection models and tracking methods, and discusses which approaches yield the most effective results for marine litter tracking in multibeam acoustic data. Finally, Section VI summarises the main findings and outlines directions for future work.

## II. RELATED WORK

### A. CLASSICAL TRACKING BASED APPROACHES

Classical targeting methods for acoustic imaging have progressed from filter-based techniques to statistical and machine learning approaches.

Research on sonar-based tracking for underwater vehicles began with Kalman filtering, enabling simultaneous monitoring of moving and stationary objects [16]. Subsequent work incorporated temporal features for behavioural analysis [17], hybridised Kalman and particle filters [18], [19], and introduced dual-frequency sonar with nearest-neighbour search for improved accuracy [20].

Deterministic methods gave way to probabilistic frameworks that reduced false alarms and improved state estimation, including Bayesian data association and probability density filters [21], [22], [23]. Parallel approaches explored feature extraction and temporal modelling, from optical flow and hierarchical tracking trees [24] to classification-based feature measures [25], [26], recurrent feature analysis for segmentation [27], and Hidden Markov Models for diver classification and tracking [28].

### B. MULTI-OBJECT TRACKING DETECTION BASED APPROACHES

Object detection typically follows either a two-stage paradigm, which generates proposals before classification and refinement, or a one-stage paradigm that treats detection as a dense prediction problem from predefined anchors [29], [30], illustrated in Fig. 1. Both approaches rely on backbone networks for feature extraction [31].

MOT builds on detection, assigning identities to objects across frames while managing trajectory consistency [32], [33]. TBD pipelines, illustrated in Fig. 2, detect per frame and associate over time, while joint detection and tracking (JDT)

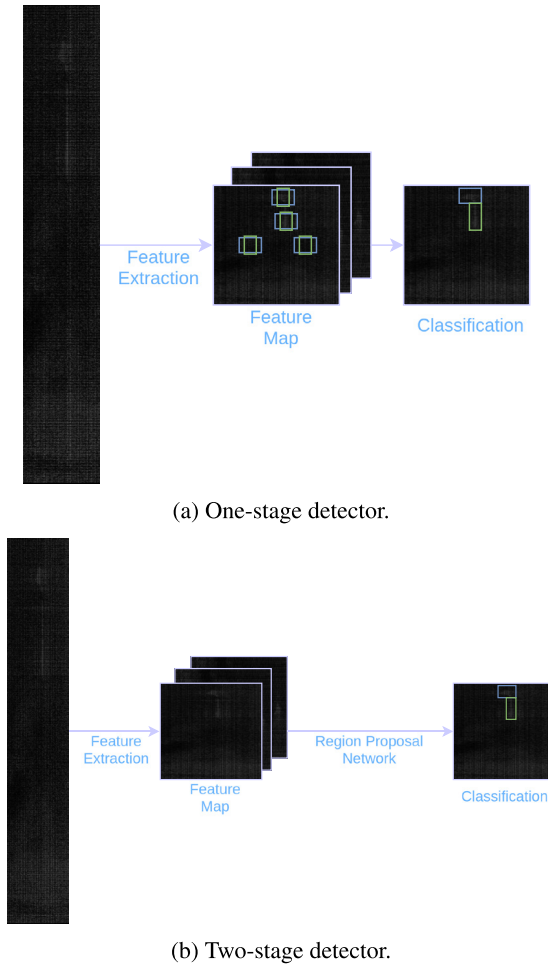


FIGURE 1. One and two stage detectors.

unifies both tasks within shared architectures [31], [34]. Examples include joint detection and embedding (JDE) [35], FairMOT [36], and CenterTrack [37]. Segmentation-based MOT offers pixel-level precision in crowded scenes [32].

Most MOT methods, like SORT [38], rely on the classical linear Kalman Filter (KF) [39] and the Hungarian algorithm [40] but often suffer from identity switches.

The KF assumes a linear dynamical system with Gaussian noise:

$$x_t = Fx_{t-1} + w_t, \quad w_t \sim \mathcal{N}(0, Q), \quad (1)$$

$$z_t = Hx_t + v_t, \quad v_t \sim \mathcal{N}(0, R), \quad (2)$$

where  $x_t$  is the state,  $z_t$  is a detection,  $F$  and  $H$  are the motion and observation matrices, and  $Q$  and  $R$  are the process and measurement noise covariances. The  $w_t$  and  $v_t$  are the process noise and measurement noise vectors.

### 1) PREDICTION STEP

The predictor equation:

$$x_{t|t-1} = Fx_{t-1|t-1}, \quad (3)$$

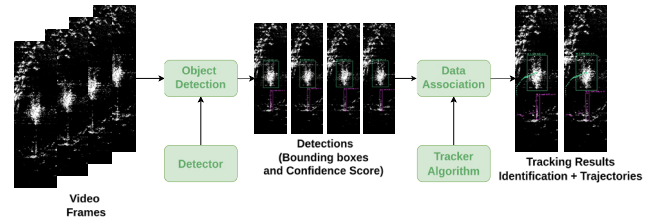


FIGURE 2. Tracking by detection (TBD).

The covariance extrapolation:

$$P_{t|t-1} = FP_{t-1|t-1}F^T + Q. \quad (4)$$

### 2) UPDATE STEP

The Kalman gain [39] is:

$$K_t = P_{t|t-1}H^T (HP_{t|t-1}H^T + R_t)^{-1}. \quad (5)$$

State correction:

$$x_{t|t} = x_{t|t-1} + K_t (z_t - Hx_{t|t-1}). \quad (6)$$

Covariance update, where  $I$  corresponds to the identity matrix:

$$P_{t|t} = (I - K_tH) P_{t|t-1}(I - K_tH)^T + K_tR_tK_t^T. \quad (7)$$

After prediction, a cost matrix  $C$  is built:

$$C_{ij} = \text{cost}(i, j). \quad (8)$$

This produces the optimal one-to-one matching between tracks and detections. All modern MOT trackers in BoxMOT use the Hungarian Algorithm, differing only in how  $C$  is constructed.

The Hungarian algorithm solves:

$$\min_X \sum_{i=1}^N \sum_{j=1}^M C_{ij}X_{ij}, \quad (9)$$

With standard one-to-one assignment constraints:

$$X_{ij} \in \{0, 1\}, \quad \sum_i X_{ij} \leq 1, \quad \sum_j X_{ij} \leq 1. \quad (10)$$

This yields a binary assignment matrix  $X$ , where  $X_{ij}$  selects detection  $j$  for track  $i$ .

DeepSORT [46] reduces identification (ID) switching with appearance embeddings, though such cues are unreliable in sonar imagery, while ByteTrack [41] addresses occlusion and low-confidence detections through confidence-aware association, making it suitable for noisy environments. BoostTrack [45] enhances similarity scoring with confidence scaling and IoU or Mahalanobis metrics, whereas OC-SORT [42] and Deep OC-SORT [43] reformulate Kalman-based tracking with observation-centric recovery and adaptive appearance integration. BoT-SORT [44] combines motion- and appearance-based association in a two-stage matching

**TABLE 1. Summary of core association-cost equations and key differences between common MOT trackers. All implement the same KF motion model and Hungarian assignment; differences arise from association cost and update rules.**

Tracker	Association Cost $C$	Key Distinctions	Reference
<b>SORT</b>	$C = 1 - \text{IoU}(B_i, B_j)$	Baseline Intersection over Union (IoU) matching; no appearance or confidence fusion; standard KF.	[38]
<b>ByteTrack</b>	$C = 1 - \text{IoU}$ (two-stage)	Matches high-confidence detections first, then low-confidence ones to recover fragmented tracks.	[41]
<b>OC-SORT</b>	$C = 1 - \text{IoU}$ $x_{t t} = (1 - \alpha)x^p + \alpha z_t$	Adds observation-centric update for better recovery in jittery/noisy detections; improves stability.	[42]
<b>DeepOC-SORT</b>	$C = \beta(1 - \text{IoU}) + (1 - \beta)(1 - \cos(\phi_i, \phi_j))$	Adds Reidentification (ReID) appearance embeddings and smoothing; more effective in optical imagery.	[43]
<b>BoT-SORT</b>	$C = \alpha(1 - \text{IoU}) + (1 - \alpha)D_m D_m = (z - Hx)^\top S^{-1}(z - Hx)$	Fusion of IoU + Mahalanobis motion distance; uses FuseMotion + temporal smoothing; robust to noisy detections.	[44]
<b>BoostTrack</b>	$C = (1 - \lambda)(1 - \text{IoU}) + \lambda(1 - s_{\text{conf}})$	Confidence-weighted matching; enhances association when detector scores are unstable.	[45]

scheme, refining trajectories with temporal feature smoothing. Although all evaluated trackers originate from the classical SORT framework, their differences arise from the formulation of the association cost and the update rule. SORT relies exclusively on IoU for matching, making it sensitive to fragmented sonar detections. BoT-SORT improves robustness by integrating detection confidence and motion consistency (FuseMotion), while BoostTrack enhances the association score using confidence boosting. ByteTrack introduces high/low-score dual matching, which is advantageous in optical scenes but suboptimal in sonar due to inherently low-confidence detections. OC-SORT modifies the Kalman update to be observation-centric, improving recovery from rapid target shape changes typical in acoustic images. DeepOC-SORT further incorporates appearance embeddings, but their contribution is limited in sonar due to the absence of stable visual texture. These distinctions explain the performance variations observed across frequencies and pre-processing strategies and are summarised in Table 1.

### C. MULTI-OBJECT TRACKING SEGMENTATION AND TRANSFORMER BASED APPROACHES

Segmentation-based MOT methods improve the handling of occlusion by refining the target boundaries. Track-Region-based Convolutional Neural Network(RCNN) [33] extends Mask R-CNN with temporal convolutions and association heads, while MOTSNNet leverages optical flow for the annotation of the data set and the grouping of masks for the integration of features [47]. PointTrack++ treats pixels as point clouds for precise identity preservation in dense scenes [48].

Transformer-based MOT introduces an attention-driven association. TransTrack applies query-key matching [49], TrackFormer models crowded interactions with encoder-decoder attention [50], and MOTRv2 improves flexibility through deformable transformers [51]. Memory-augmented designs such as MeMOT [52] and MeMOTR [53] further

strengthen long-term associations by storing and reasoning over historical object features.

### D. MULTI-OBJECT TRACKERS APPLIED IN MULTIBEAM ECHOSOUNDER ACOUSTIC IMAGES

In sonar imagery, a YOLOv3–particle filter framework demonstrated 3D target tracking in tank experiments with Autonomous Underwater Vehicles (AUVs) and a turtle replica [54]. Another study combined YOLOv5 with DeepSORT, improving detection via image enhancement and DCGAN-based augmentation [55]. More recently, a Swin Transformer–YOLOv5 architecture expanded bounding boxes and receptive fields, reducing ID switches and trajectory interruptions [7].

### E. RELATED WORK SUMMARY

This work focuses on detecting and tracking marine litter in the water column using a multibeam echosounder. Most MOT research has centred on pedestrians in optical imagery, the acoustic domain has unique challenges.

The evolution from classical filter-based methods to modern deep learning frameworks shows advantages and limitations in underwater applications. Early KF approaches [16], [17] enabled motion prediction suitable for drift-prone litter influenced by currents but struggled with complex non-linear dynamics. Particle filters [18] improved non-linear motion at the cost of computational efficiency, restricting the use in real-time. Probabilistic data association methods [21], [22] offered robust uncertainty management, addressing the high false alarm rates common in sonar imagery.

TBD paradigms with online tracking capabilities, as suggested by [54], can be used for sonar-based marine litter tracking. These approaches allow the integration of detection with tracking models, useful when dealing with sparse and irregular acoustic object appearances.

Trackers like DeepSORT, which rely on re-identification networks trained on textured and colour-rich optical data, encounter limitations in sonar imagery lacking such cues.

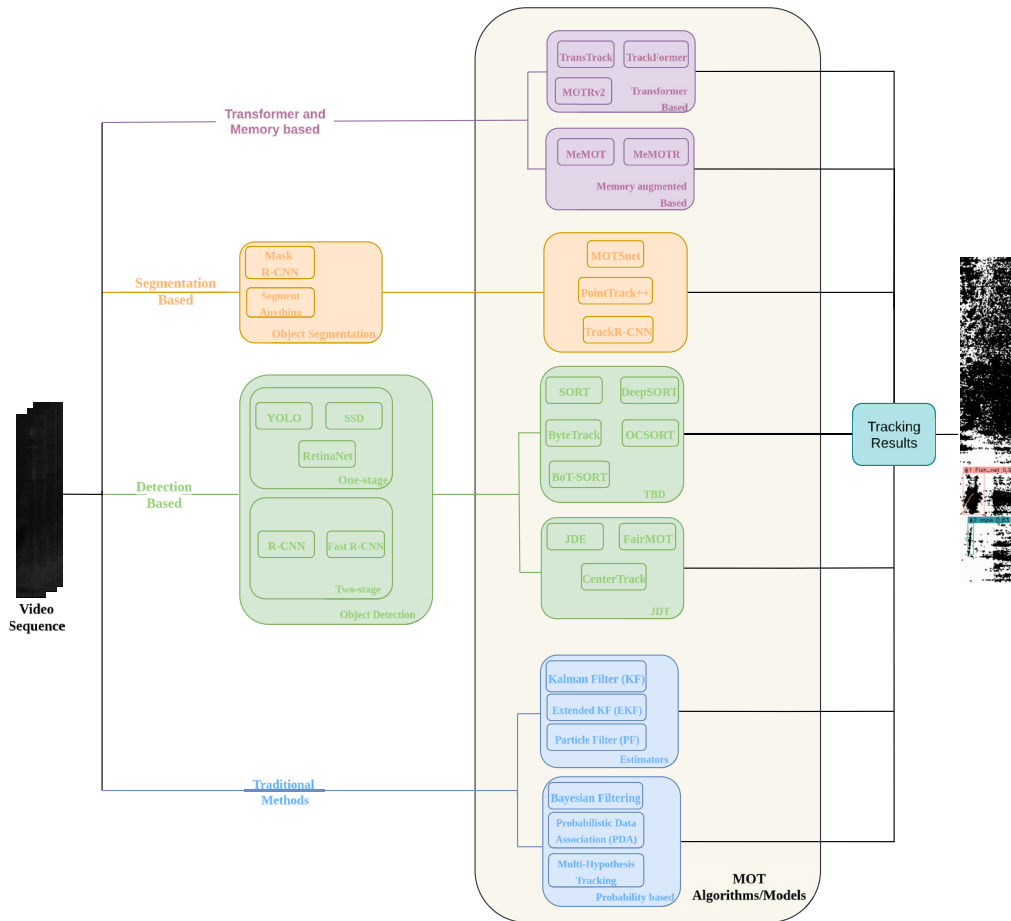


FIGURE 3. Tracking methodologies summary.

Motion-driven methods such as BoT-SORT and OC-SORT may provide more consistent performance. Confidence-aware association strategies, such as ByteTrack and BoostTrack, can be relevant in sonar contexts where detection confidence varies with acoustic noise, partial visibility, and target orientation. These methods extend the probabilistic principles of classical tracking [21] for more modern frameworks. Observation-centric designs (OC-SORT and Deep OC-SORT) can address occlusions caused by marine life, water column variability, or acoustic shadows.

Preliminary experiments with DeepSORT [56] although with satisfactory results, had common issues found in this tracker such as occlusion and ID switching. Literature indicates that approaches like BoT-Sort and BoostTrack, can offer enhanced robustness under such conditions. A summary of these methodologies is presented in Fig. 3.

The limited adoption of MOT techniques in acoustic imagery highlights a significant research gap—offering opportunities to develop specialised tracking strategies that integrate the strengths of classical statistical models and contemporary deep learning paradigms for robust marine litter tracking. Target detection and tracking can allow for the application of efficient manoeuvrability for the in-situ

extraction of marine litter, minimising the impact that this threat introduces in the environment [57].

### III. DATA EXTRACTION EXPERIMENTAL SETUP

Data was collected in a next-to-real-life scenario using PORTUS, an electric Autonomous Surface Vessel (ASV) developed by the Centre of Robotics and Autonomous Systems (CRAS) at INESC TEC, illustrated in Fig. 4b.



(a) Kongsberg M3 high-frequency model installed in PORTUS.



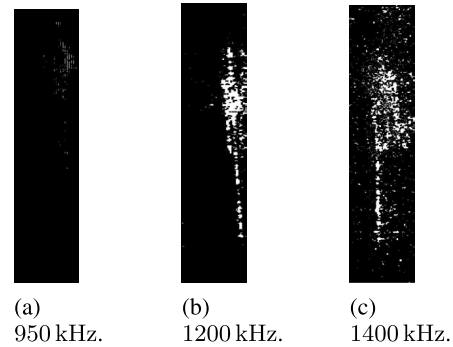
(b) PORTUS surface vessel.

FIGURE 4. Multibeam data acquisition mounted on PORTUS ASV during field trials in Leixões harbour.

PORTUS is equipped with a broad suite of sensors that support navigation and perception tasks in maritime envi-

**TABLE 2.** Sonar specifications for different operating frequencies. Resulting acoustic images can be checked in Fig. 5.

Specifications	Operating frequency (kHz)		
	950	1200	1400
Angular Resolution	140°×27°	75°×21°	45°×18°
Maximum Range (m)	8	8	8
Beams	256		
Reflections per beam	1573		
Pulse Type	Continuous Wave (CW) and Chirp Modulation		
TVG_A	20		
TVG_B	100		
TVG_C	-8		
TVG_L	100		
Image gain	10		
Gain threshold	140		



**FIGURE 5.** Wooden deck captured at different acoustic frequencies with the M3 MBES high frequency model.

ronments. PORTUS offers low acoustic and environmental impact as a fully electric platform, making it ideal for operations that use acoustic sensors. The ASV has a telescopic keel extending up to 1.5 meters in depth. It enhances the vessel's hydrodynamic stability during operations and serves as the mounting point for underwater sensors. By positioning these sensors below the waterline, the influence of surface disturbances—such as waves, turbulence, and air bubbles—is significantly reduced, thereby improving data quality and sensor reliability. The M3 high frequency multibeam echosounder (MBES) model from Kongsberg [58], Fig. 4a, was installed in the PORTUS telescopic keel, as illustrated in Fig. 6.

The mounting structure for the sonar was designed to support an Forward Looking Sonar (FLS) configuration and has a set of perforated discs. The MBES can be oriented at known angles by selecting specific hole alignments. For this data extraction, the sonar was fixed at a 25-degree tilt, providing forward visibility while avoiding reflections from the vessel's hull, the water surface, and the backscatter of the seabed.

This MBES relies on a development kit developed by Kongsberg, and can only operate using Windows Operating System. This software configures the sonar head and extracts the MBES data. An in-house Robot Operating System (ROS) driver was developed to extract the sonar data and reconfigure the M3 sonar head during the campaign, for example, changing the sonar maximum range. Having a solution based in a ROS allows to have access to the PORTUS navigation and synchronize it with the campaign sonar data, adding the possibility to repeat the mission with real data, as illustrated in Fig. 7b. The multiple Time Variant Gains (TVG) that are possible to configure were kept the same throughout the entirety of the data extraction. The set of configurations and the main sonar specifications are listed in Table 2.

Data was collected in the manoeuvring basin of Leixões harbour in Porto, Portugal (41°11'04.8"N 8°42'22.0"W), Fig. 7. The zone where data was collected is shown in Fig. 7a. These waters had a minimum depth of 2.5 meters and a maximum depth of 10 meters.

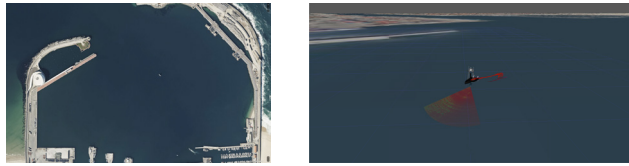


**FIGURE 6.** M3 HF MBES keel mounting point.

Targets were tethered to weights to simulate marine debris floating in the water column. While mounted in the FLS configuration it was possible to take advantage of the MBES azimuth's to generate acoustic footprints with depth. Different materials were used with similar shapes are detailed in Table 3, listing the characteristics of each target. The movement and changes in range and bearing to the targets provided multiple viewpoints, and their footprints changed significantly in the acquired acoustic images. The target depth in the water column was not altered during the experiment.

Polar acoustic images were generated based on the algorithm proposed in [15]. The acoustic images were generated with a maximum range of 8 m and a fixed field of view (FOV) of 45°. Each generated frame is composed of 450 × 4002 pixels. Different acoustic frequencies were used during the acquisition, namely: 950 kHz, 1200 kHz and 1400 kHz, as illustrated in Fig. 5. Different fields of view and, consequently, beam spacing leads to varying spatial resolutions among the acoustic configurations. The 1400 kHz acoustic images, as the one in Fig. 5c, achieves the highest resolution, characterised by closely spaced pixels and enhanced spatial detail, and opposite occurs with the 950 kHz, Fig. 5a, with lower resolution and larger beam spacing.

Acoustic images are generated based on the backscatter of the three-dimensional spatial environment. This type of image suffers from several drawbacks when compared to optical generated images:



(a) Mission aerial view. (b) Mission scenario with real-life navigation data and acoustic data in ROS.

**FIGURE 7. Mission scenario in the manoeuvring basin of Leixões harbour in Porto, Portugal (41°11'04.8"N 8°42'22.0"W).**

**TABLE 3. Dimensions and material composition of tracked objects.**

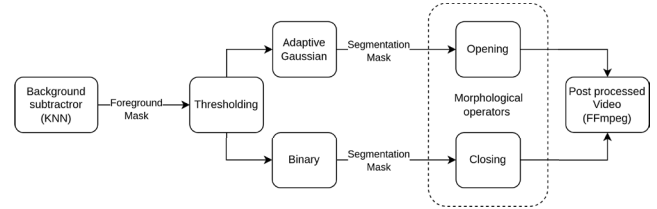
Object	Dimensions (cm)	Material
Aluminium tube	39 height 8 diameter	Aluminium
Fish net	Undisclosed	Nylon, foam and buoys
Floating wood tile	20 × 20	High Density Fibreboard
Perforated deck	56 × 56	Polyethylene
Plastic deck	30 × 30	Wood-plastic composite
Polyvinyl chloride (PVC) opaque square	50 × 50 × 0.3	PVC
Traffic cone	22 × 22 base 36 height	PVC
PVC transparent square	50 × 50 × 0.3	PVC
Vinyl sheet	49 × 62	Vinyl
Wooden deck	30 × 30 × 2	Wood

- Multipath propagation occurs when acoustic waves exhibit higher energy levels than those reflected from obstacles.
- Lower resolution in the produced images due to the transducer size and the quantity of transducers that can be incorporated into an array.
- A target backscatter can be affected by factors such as shape, material composition, and the distance to the sonar head. Furthermore, the angle at which acoustic waves strike the target may change as a result of the target’s movement, leading to the emergence of distinct regions within the acoustic image of the same target. These regions frequently appear as disjointed segments in acoustic imagery.

Video files were generated with the set of acquired acoustic frames, since MOT solutions rely on tracing foreground targets in a sequence of frames. This was done with the same procedure detailed in [56].

**A. ACOUSTIC VIDEO - GENERATION**

A set of acoustic images corresponding to a target acquired at a specific frequency is processed. The dimensions of these images are adjusted to ensure they are consistently divisible by 2. This requirement arises from video encoding formats, particularly those used by Ffmpeg, to enable efficient compression and encoding, especially when utilising the



**FIGURE 8. Acoustic video pre-processing [56].**

H.264 codec. If the images within the set vary in size, they are resized by adding padding to match the largest width and height while preserving the original resolution. Only padding is applied when necessary without altering the image content.

Each frame is embedded with a timestamp, generated from ROS message headers. The frame rate is computed based on the average time difference between consecutive frames. Ffmpeg is then employed to generate the video, using the H.264 codec with the calculated frame rate. The video is encoded in the MPEG-4/H.264 format, ensuring compatibility with the annotation tool and its supported video formats. Additionally, metadata for each video is stored in a JSON file.

Annotating the original videos proved challenging due to the high noise levels typically present in acoustic images and their low spatial resolution, which made target footprints difficult to discern. This issue became even more pronounced at lower acoustic signal frequencies. Pre-processing techniques were applied to enhance the videos to address these challenges.

**B. ACOUSTIC VIDEO - PRE-PROCESSING**

The video processing tasks involved background subtraction, thresholding, and morphological operations, as summarised in Fig. 8.

Background subtraction was performed using the K-Nearest Neighbors (KNN) method, which relies on temporal information from consecutive frames and accumulated history. The algorithm evaluates whether each pixel belongs to the background or the foreground: pixels that deviate from historical values are classified as foreground, while consistent pixels are assigned to the background. The background model is updated continuously to adapt to gradual scene variations [59]. Because this technique relies heavily on historical frames, the background model was initialized and updated strictly within isolated video sequences to prevent any temporal data leakage between the training and testing phases, as detailed in Section IV.

After background subtraction, thresholding was applied to the foreground mask to reduce background noise typical of acoustic imagery. Two thresholding strategies were used: binary thresholding and adaptive Gaussian thresholding. Binary thresholding applies a fixed global threshold to convert greyscale frames into binary form, whereas adaptive Gaussian thresholding computes local thresholds based on

neighbourhood intensity distributions. Both approaches produce a binary mask of the detected foreground objects [60].

Morphological filtering was then applied to refine the binary masks. Closing was performed after binary thresholding to fill small gaps within objects, while opening was used after adaptive Gaussian thresholding to suppress isolated noise while preserving object contours. These operations improved the consistency of the extracted foreground structures [59], [60].

With the pre-processing four different input video types can be considered. The Normal variant corresponds to polar images without any pre-processing, while the remaining three variants apply different binarization and filtering strategies prior to training. These transformations aim to enhance target visibility and suppress background noise, potentially improving the robustness of both detection and tracking models.

- **Normal:** Polar acoustic images directly used as input, without any additional processing. This serves as the baseline condition against which all other variants are compared.
- **Adaptive\_gaussian\_11\_opening\_KNN:** Images processed using adaptive Gaussian thresholding with a kernel size of 11, followed by morphological opening to remove small artifacts. A KNN filter is applied to refine the binary mask and reduce noise.
- **Binary\_50\_closing\_KNN:** Images binarized with a fixed threshold value of 50, followed by morphological closing to fill small gaps in detected regions. A KNN filter is again applied to improve contour consistency.
- **Binary\_250\_closing\_KNN:** Similar to the previous variant, but binarization is performed with a higher fixed threshold value of 250. The higher threshold reduces sensitivity to weak reflections, focusing on stronger structures. Morphological closing and KNN filtering are applied as in the 50-threshold case.

The processed frames were reconstructed into videos using the same procedure as the original generator, resulting in outputs that differ substantially from the raw acoustic sequences. The pre-processed videos provide clearer separation of foreground targets from background clutter. Pre-processing was necessary because target positions change in practice, influenced by currents, marine life, and the motion of the PORTUS surface vessel.

#### IV. MULTI-OBJECT TRACKING PIPELINE FOR MARINE LITTER DETECTION

The selected framework for this study was MOT based on TBD paradigm. Multi-object tracking is commonly formulated within this paradigm, where objects are first detected in each frame and then associated across time to form continuous trajectories. As specified in Section II the TBD pipeline consists of two primary stages: object detection and data association.

#### A. DETECTION PHASE

In the detection stage, a dedicated detector localises potential targets of interest in each frame. Since this work focuses specifically on the comparative evaluation of multi-object tracking algorithms rather than detection architectures, all tracking experiments were conducted using a single, high-performing YOLOv8 detector. This ensures that performance differences arise from the trackers themselves and not from variations in detector design.

Multiple YOLOv8-based models were trained according to acoustic frequency, input type, and pre-trained variant.

All models were initialised with pre-trained YOLOv8 weights and trained on an NVIDIA GeForce RTX 2080 SUPER GPU. Training was conducted for a maximum of 200 epochs, incorporating an early stopping mechanism with a patience of 50 epochs to prevent large training times without improvements. The input image size was  $640 \times 640$  pixels with a batch size of 12. To improve convergence and regularisation on the small dataset, the AdamW optimiser was applied with an initial learning rate of 0.001, weight decay of 0.0001, and momentum of 0.937.

Standard optical augmentation strategies such as mosaic, MixUp, vertical flipping, and hue/saturation colour shifts were disabled, as they violate the geometric constraints and single-channel intensity properties of FLS acoustic data. The augmentation pipeline that was applied to all models included horizontal flipping (0.5 probability), random rotations ( $\pm 10^\circ$ ), scaling ( $\pm 20\%$ ), and spatial translation ( $\pm 10\%$ ). Additionally, random intensity shifting (0.4) was applied to simulate varying acoustic gain and depth attenuation, alongside random erasing (0.4) to mimic acoustic shadowing and signal occlusion in the water column.

The performance was evaluated using mean Average Precision (mAP). Two measures were considered: mAP50, with an IoU threshold of 0.5, and mAP50:95, which averages results over thresholds from 0.5 to 0.95 in steps of 0.05.

Because this study evaluates a tracking-by-detection paradigm, the dataset splitting strategy must preserve the temporal continuity of the test sequences. A standard random train-test split at the frame level would fragment trajectories, rendering tracking evaluation impossible. Therefore, a contiguous sequence comprising 20% of each video was isolated exclusively for the test set. To prevent the selection of empty sequences where tracking cannot occur, a class-aware sliding window algorithm was implemented. This algorithm traversed the entire video to identify the continuous block of frames containing the highest density of tracking targets. Furthermore, to prevent the test set from depleting the training data, a constraint was applied to ensure that a minimum of 40% of the target instances remained outside the test window. Once the sequential test block was isolated and exported as a continuous video, the remaining frames were randomly shuffled and split into training (90%) and validation (10%) sets, providing the detector with diverse, uncorrelated data during the learning phase. All frames were annotated and

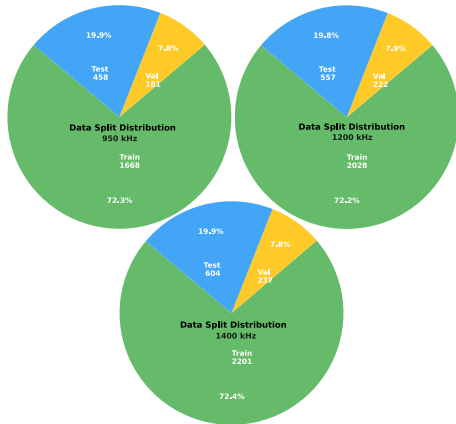


FIGURE 9. Sequential dataset split for detection and tracking models.

converted to the YOLOv8 labelling format. The resulting splitting for this dataset can be visualised in Fig. 9. To prevent temporal data leakage through the historical frame buffer of the KNN background subtractor, this sequence-level isolation was performed prior to applying any temporally dependent pre-processing. The KNN background model was initialized and run independently on the training/validation sets and the test sequences, ensuring the test set’s temporal history remained completely unseen during the training phase.

TABLE 4. Detector performance on the test set for well-represented targets (Criteria: Annotations  $\geq 100$  and  $mAP@50 \geq 0.65$ ).

Pre-processing	YOLO Variant	Target Class	Freq. (kHz)	Train Count	$mAP@50$
Adaptive Gaussian	Medium	Fish Net	950	162	0.917
Normal	Medium	Fish Net	950	162	0.914
Binary 50 Closing	Nano	Fish Net	950	162	0.850
Adaptive Gaussian	Medium	PVC Blue Square	950	108	0.844
Binary 250 Closing	Nano	Fish Net	950	162	0.798
Binary 250 Closing	Medium	PVC Blue Square	1200	116	0.746
Binary 50 Closing	Medium	PVC Perforated Deck	1200	113	0.743
Normal	Nano	PVC Perforated Deck	1200	113	0.741
Binary 50 Closing	Medium	PVC Blue Square	1200	116	0.740
Adaptive Gaussian	Medium	PVC Perforated Deck	1200	113	0.710
Binary 250 Closing	Medium	PVC Perforated Deck	1200	113	0.689
Adaptive Gaussian	Nano	Fish Net	1200	194	0.676
Binary 250 Closing	Medium	Fish Net	1200	194	0.652

The sequential splitting strategy inherently reduced the availability of frames containing targets in the training set. This is illustrated by the number of objects that were available in the training set at each acoustic frequency, in Fig. 10. This limitation, although impacting the detector, was mitigated in many classes through hyperparameter tuning and since the detector already has pre-trained weights with proven performance. Evaluation on the contiguous test set demonstrated robust detection capabilities for several key targets, even more when a larger training volume (train count) was available, as summarised in Table 4. Furthermore, the physical characteristics of the objects had a crucial role in the detector’s performance. Targets possessing larger acoustic footprints and higher reflectivity provided distinct signatures in the sonar imagery, yielding superior results. For instance, the fish net agglomerate, supported by 162 training instances at 950 kHz, achieved a  $mAP@50$  of up to 0.917, even though at this acoustic frequency there is smaller spatial resolution. As detailed in Table 4, the medium YOLO variant

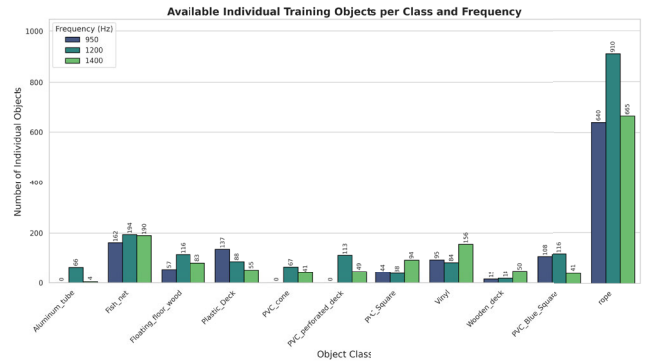


FIGURE 10. Number of objects available in the training set for each frequency.

had better performance with these constraints. Its higher parameter capacity allowed it to better learn the complex acoustic backscatter of these reflective materials compared to its lighter counterpart. Nevertheless, the nano variant still secured several top-performing spots (particularly in detecting the fish net at 950 kHz and 1200 kHz), proving that with the right thresholding, a lightweight architecture provides good results. These results were maximised by specific preprocessing techniques, particularly the Adaptive Gaussian configuration, which consistently anchored the top-performing setups for both model variants.

The test set comprises continuous video sequences that are pre-processed independently to prevent any data leakage. To ensure a comprehensive and rigorous evaluation of both the detection and tracking algorithms, a data augmentation pipeline was applied to these test videos to introduce realistic spatial and temporal variations. All test sequences underwent a mandatory horizontal flip to simulate alternate vehicle trajectories. Additionally, temporal reversal was applied with a 70% probability to guarantee that temporal information was not affecting how the detector performed, simulating targets moving in opposite directions in what was extracted in the original dataset. Random pixel intensity scaling (by a factor between 0.5 and 1.2) was introduced with a 50% probability to simulate natural fluctuations in acoustic backscatter strength and sonar gain settings, resulting in the augmentations represented in Fig. 11.

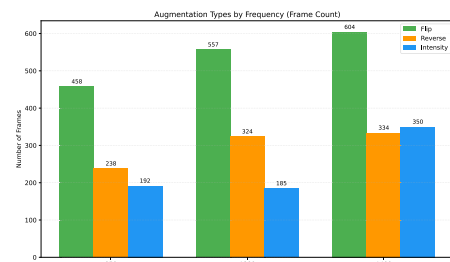


FIGURE 11. Augmentations performed to the test set by each frequency.

**TABLE 5. Pre-processing ranking comparison: Average mAP@50 for well-represented classes ( $\geq 100$  annotations) versus the overall dataset average.**

Pre-processing	Well-Represented mAP@50	Overall mAP@50
Adaptive Gaussian	0.611	0.502
Binary 50 Closing	0.551	0.462
Binary 250 Closing	0.530	0.446
Normal	0.525	0.493

Following this augmentation, the background subtraction, thresholding, and morphological operations pipeline was applied to each sequence individually, without relying on global dataset statistics. As detailed in Table 5, Adaptive Gaussian emerged as the highest-ranked pre-processing method, achieving an average mAP@50 of 0.611 on well-represented targets (those with  $\geq 100$  training annotations). Under these optimal training conditions, the intermediate binary thresholding approaches outperformed, confirming that appropriate pre-processing is crucial for effectively reducing acoustic clutter while preserving essential target signatures.

When evaluating the performance across the entire dataset (also containing the augmented video sequences), the overall averages naturally decrease due to the inclusion of severely under-represented and less reflective targets. As shown in Table 5, the overall mAP@50 for Adaptive Gaussian drops to 0.502, while unprocessed raw inputs (Normal) demonstrate relative resilience against acoustic clutter, maintaining an overall average of 0.493. This contrast between the well-represented targets and the global dataset explicitly demonstrates that higher training sample volumes directly yield substantially better detection results. The fact that the small dataset size diminishes overall average metrics does not seem to invalidate the tracking benchmarking, nor does it undermine the broader feasibility of automated marine litter detection.

Because the training dataset is inherently limited by these challenges of real-world data acquisition and requirement of sequential data with available targets, it was initially hypothesized that higher-capacity models might be prone to memorising the training data. However, as evidenced by the results on classes with sufficient annotations (Table 4), the medium YOLO variant demonstrated robust generalization on the unseen, augmented test videos. Its higher parameter count allowed it to better capture and learn the complex acoustic signatures of the targets when enough samples were provided.

This superiority of the medium variant, alongside the effects of acoustic frequency, is further corroborated when examining the comprehensive performance averaged across all targets, as detailed in Table 6. The medium variant paired with Adaptive Gaussian pre-processing consistently maintained the highest mAP@50 scores at lower acoustic frequencies (0.580 at 950 kHz and 0.550 at 1200 kHz). Furthermore, this comprehensive breakdown clearly illustrates how the optimal configuration shifts under extreme

acoustic clutter. At the 1400 kHz frequency, the performance dropped to a 0.424 mAP@50. Instead, the Medium model paired with unprocessed (Normal) inputs emerged as the most resilient configuration, achieving 0.526 mAP@50. This demonstrates that while higher model capacity is universally beneficial across the dataset, the pre-processing strategy must be explicitly tailored to the operational acoustic frequency and with enough data to generalise.

## B. TRACKING PHASE

Following detection, the association stage links current detections with previously tracked objects to maintain consistent identities across frames. For the application of marine litter detection, online tracking-by-detection pipelines are particularly relevant. The pipeline begins with the detection of potential litter items using trained acoustic image detectors, followed by online association that maintains trajectories even when objects undergo short-term occlusion or displacement due to currents. State-of-the-art association strategies offer a balance between robustness and computational efficiency, making them attractive for integration into autonomous monitoring platforms operating in dynamic underwater environments.

The selected state-of-the-art tracking algorithms were based on SORT as the likes of BoostTrack, BoT-SORT, ByteTrack, DeepOC-SORT, and OC-SORT. These methods were integrated into a TBD framework, relying on pre-trained detection models based on YOLOv8. All of these methods were applied through the use of BoXMot [61].

BoXMot has a modular framework for multi-object tracking that is well suited for integration into detection pipelines as is the case of the one applied to the extracted acoustic images. Its pluggable architecture allows state-of-the-art tracking modules to be inserted or replaced without major changes to the system, which facilitates experimentation with different association strategies. Because BoXMot is detector-agnostic, it can be used with any object detection, segmentation, or pose estimation model that outputs bounding boxes, including YOLO-based detectors, as the ones trained specifically on acoustic images of marine litter. This flexibility ensures that the framework can be directly applied to underwater monitoring without requiring custom structural modifications.

BoXMot's combination of modularity, detector-agnostic design and reproducible evaluation makes it a suitable framework for applying and benchmarking multi-object tracking approaches in marine environments [61].

During the tracking inference phase, the YOLOv8 detectors were configured with a minimum confidence threshold of 0.25 and an IoU threshold of 0.70, ensuring a balanced input of candidate bounding boxes for the downstream tracking algorithms without discarding possible targets.

## C. MULTI-OBJECT TRACKING METRICS

Evaluating MOT systems requires a comprehensive set of metrics that capture detection quality, identity

TABLE 6. Overall detector performance by configuration (Averaged Across All Targets).

Pre-processing	YOLO Variant	Freq. (kHz)	Precision	Recall	mAP@50	mAP@50:95
Adaptive Gaussian	Medium	950	0.643	0.539	0.580	0.183
Adaptive Gaussian	Nano	950	0.615	0.487	0.531	0.176
Adaptive Gaussian	Medium	1200	0.642	0.509	0.550	0.187
Adaptive Gaussian	Nano	1200	0.640	0.518	0.536	0.174
Adaptive Gaussian	Medium	1400	0.493	0.406	0.424	0.162
Adaptive Gaussian	Nano	1400	0.493	0.444	0.459	0.179
Binary 250 Closing	Medium	950	0.527	0.372	0.416	0.118
Binary 250 Closing	Nano	950	0.506	0.430	0.424	0.118
Binary 250 Closing	Medium	1200	0.633	0.488	0.536	0.177
Binary 250 Closing	Nano	1200	0.652	0.491	0.543	0.176
Binary 250 Closing	Medium	1400	0.511	0.339	0.359	0.125
Binary 250 Closing	Nano	1400	0.552	0.402	0.419	0.146
Binary 50 Closing	Medium	950	0.540	0.420	0.433	0.128
Binary 50 Closing	Nano	950	0.504	0.462	0.447	0.133
Binary 50 Closing	Medium	1200	0.600	0.470	0.499	0.167
Binary 50 Closing	Nano	1200	0.638	0.495	0.543	0.182
Binary 50 Closing	Medium	1400	0.543	0.364	0.406	0.140
Binary 50 Closing	Nano	1400	0.588	0.414	0.452	0.162
Normal	Medium	950	0.500	0.422	0.403	0.113
Normal	Nano	950	0.458	0.412	0.392	0.110
Normal	Medium	1200	0.625	0.513	0.547	0.185
Normal	Nano	1200	0.633	0.479	0.521	0.158
Normal	Medium	1400	0.585	0.519	0.526	0.196
Normal	Nano	1400	0.568	0.512	0.502	0.193

preservation, localisation precision, and computational efficiency. Classical measures such as precision and recall provide a baseline for detection performance, but they fail to capture the complexity of long-term tracking, where maintaining consistent identities and continuous trajectories is essential. To address this, standardised MOT metrics have been proposed [32], [62]. These metrics collectively enable fair benchmarking of tracking systems under diverse scenarios.

### 1) PRIMARY ACCURACY METRICS

The most widely adopted performance measure is the Multiple Object Tracking Accuracy (MOTA), which combines errors from false positives (FP), false negatives (FN), and identity switches (IDSW) into a single score. MOTA is defined as:

$$\text{MOTA} = 1 - \frac{FN + FP + \text{IDSW}}{GT} \quad (11)$$

where  $GT$  is the total number of ground-truth objects. MOTA penalizes three types of errors equally, offering a high-level overview of tracker accuracy. A value closer to 1 indicates stronger performance.

Complementing MOTA, the Multiple Object Tracking Precision (MOTP) focuses on localisation accuracy. MOTP measures the alignment between predicted and ground-truth bounding boxes:

$$\text{MOTP} = \frac{\sum_{t,i} d(t,i)}{\sum_t c_t} \quad (12)$$

where  $d(t,i)$  is the distance between detection and ground truth for object  $i$  at time  $t$ , and it is computed as:

$$d(t,i) = 1 - \text{IoU} \quad (13)$$

The  $c_t$  is the number of matches at time  $t$ . Unlike MOTA, which aggregates errors, MOTP quantifies the spatial alignment between predicted and ground-truth bounding boxes across all matched pairs. MOTP evaluates how precisely a tracker localizes objects once a correct association is made. Lower MOTP indicates higher localization precision under this distance definition. Since IoU is unitless, the resulting MOTP score is also unitless. Lower MOTP values indicate better localisation, with 0 representing perfect alignment. This metric complements MOTA by capturing localisation quality independently of detection count or identity consistency. Acoustic multibeam images in this work are generated in polar space, where the pixel spacing varies with range and beam geometry and the vertical and horizontal axes do not represent uniform physical distances (meters). Because of this using Euclidean distance in pixel or meter units would not be consistent across the image. IoU-based MOTP avoids this issue and is the standard MOT practice for non-uniform spatial grids.

### 2) IDENTITY-FOCUSED METRICS

Identity preservation is critical in MOT since losing or switching identities can severely impact downstream analysis. The identification F1 Score (IDF1) evaluates this aspect by measuring the harmonic mean of identification precision (IDP) and recall (IDR):

$$\text{IDF1} = \frac{2 \times \text{IDTP}}{2 \times \text{IDTP} + \text{IDFP} + \text{IDFN}} \quad (14)$$

where  $\text{IDTP}$ ,  $\text{IDFP}$ , and  $\text{IDFN}$  denote true positive, false positive, and false negative identifications, respectively.

Supporting measures include:

$$IDP = \frac{IDTP}{IDTP + IDFP} \quad (15)$$

$$IDR = \frac{IDTP}{IDTP + IDFN} \quad (16)$$

High IDF1 scores indicate reliable identity continuity throughout a sequence. In practice, IDF1 has gained favour in recent benchmarks due to its ability to highlight fragmentation issues that may not be apparent in MOTA.

### 3) METRIC INTERPRETATION

The set of metrics presented above captures complementary aspects of tracker performance. For example, a tracker with high MOTA may still suffer from poor identity preservation, as revealed by a low IDF1 score. Conversely, strong IDF1 performance does not guarantee accurate localisation, which is reflected in MOTP. Comprehensive evaluation requires analysing all metrics jointly to understand the trade-offs between detection quality, identity preservation, trajectory continuity, and computational efficiency.

## V. EXPERIMENTAL RESULTS

This chapter presents the experimental evaluation of the proposed detection and tracking framework. The evaluation is structured to highlight the influence of different input pre-processing strategies, acoustic frequencies, and tracking algorithms on the overall system performance.

Multiple models were trained for the different input types and the three acoustic frequencies (950 kHz, 1200 kHz, and 1400 kHz), which led to the existence of 24 detection models. Each of these models was integrated into multiple MOT pipelines.

To evaluate the quality of the tracking association, five state-of-the-art trackers were considered:

- BoostTrack,
- BoT-SORT,
- ByteTrack,
- DeepOC-SORT,
- OC-SORT.

To ensure reproducibility and consistency across all evaluated MOT methods, all trackers were executed through the BoXMot framework, which standardises the internal KF implementation, motion model, and assignment logic. Each tracker introduces specific hyperparameters that influence the association strategy. Refer to Appendix for the tuned hyperparameters used for the tracking procedures.

This evaluation therefore encompasses 120 tracker–detector configurations, each assessed using the metrics introduced in Section IV-C. The results are reported per acoustic frequency and input pre-processing type, followed by comparative analyses across trackers.

Table 7 reports the overall distribution of metrics across all runs. The global averages show a MOTA of  $0.159 \pm 0.264$ , demonstrating moderate precision (0.604) but a lower recall (0.252). This table also makes reference to a set of figures that

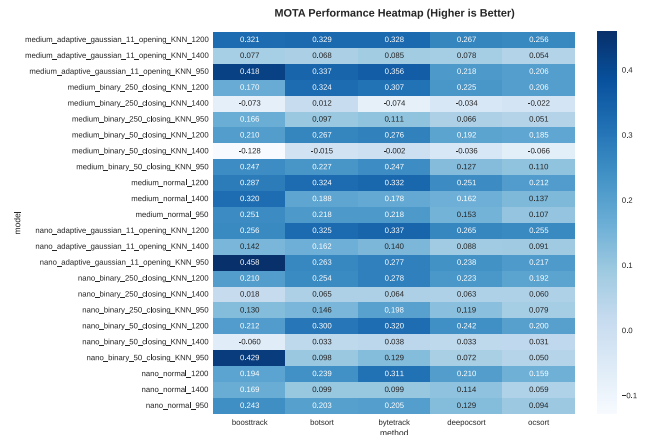


FIGURE 12. MOTA metric heat-map across multiple model and tracker configurations.

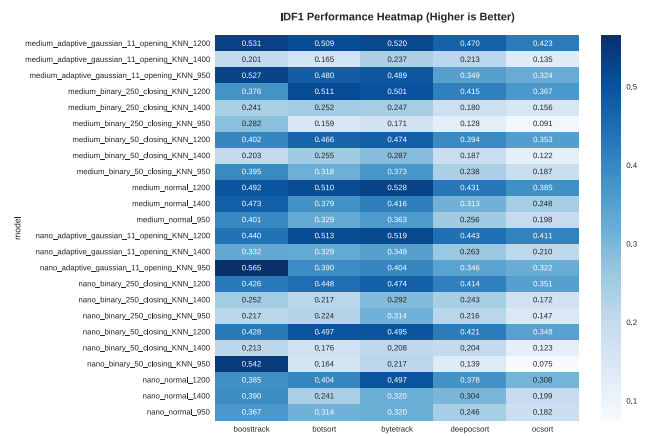


FIGURE 13. IDF1 metric heat-map across multiple model and tracker configurations.

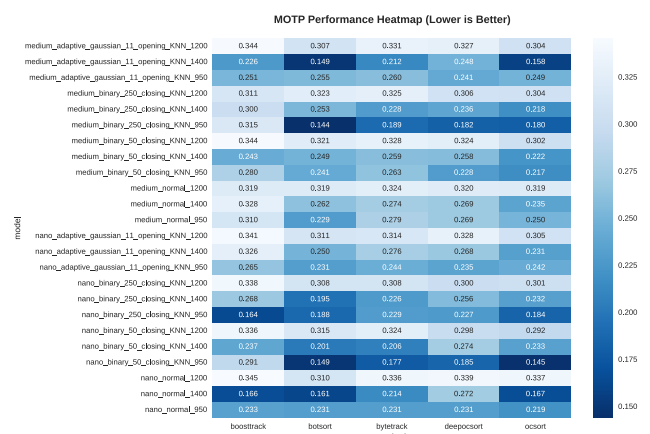


FIGURE 14. MOTP metric heat-map across multiple model and tracker configurations.

contain more detailed results per evaluated metric, dependent on each type of detector–tracker configuration for each type of video sequence that was evaluated.

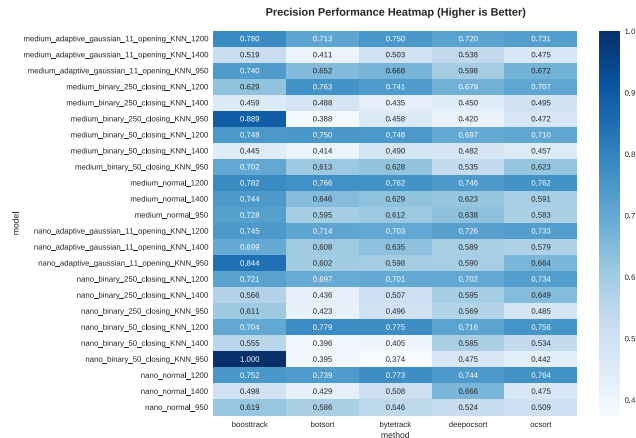


FIGURE 15. Precision metric heat-map across multiple model and tracker configurations.

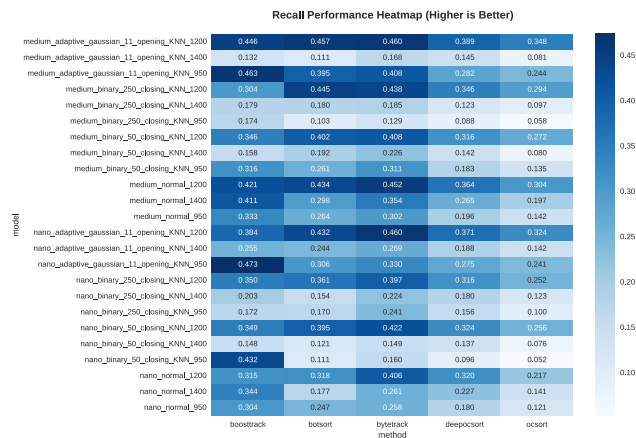


FIGURE 16. Recall metric heat-map across multiple model and tracker configurations.

TABLE 7. Overall performance statistics.

Metric	Mean ± SD	Range
MOTA, Fig. 12	0.159 ± 0.264	[−1.000, 1.000]
IDF1, Fig. 13	0.319 ± 0.257	[0.000, 1.000]
MOTP, Fig. 14	0.259 ± 0.149	[0.000, 0.490]
Precision, Fig. 15	0.604 ± 0.376	[0.000, 1.000]
Recall, Fig. 16	0.252 ± 0.224	[0.000, 1.000]

Table 8 compares the nano and medium YOLOv8 model variants. The medium architecture demonstrates a slight advantage in identity preservation (IDF1 of 0.325) and detection precision (0.606). Conversely, the nano model achieves a marginally higher overall tracking accuracy (MOTA of 0.160), proving to be an alternative that trades a slight drop in identity consistency for computational efficiency and faster training.

Table 9 quantifies the effect of the acoustic frequency on tracking performance. The results indicate that 1200 kHz yields the highest average MOTA (0.259) alongside the highest average number of processed frames (813),

TABLE 8. Performance by YOLOv8 architecture.

Model	MOTA	IDF1	Precision
Medium	0.157	0.325	0.606
Nano	0.160	0.312	0.602

TABLE 9. Frequency configuration impact.

Frequency	Avg MOTA	Avg Frames	Best MOTA
950	0.178	363	0.975
1200	0.259	813	0.750
1400	0.061	619	0.928

suggesting it provides the most stable and reliable conditions for detection and tracking. Conversely, while 950 kHz achieves the highest peak performance (Best MOTA of 0.975), its overall average MOTA drops to 0.178. Finally, the 1400 kHz frequency demonstrates the poorest average performance (0.061), highlighting significant diminishing returns at higher frequencies despite occasional high-performing runs.

Table 10 summarises tracker performance. ByteTrack and BoostTrack lead overall in terms of MOTA (0.188), with BoostTrack also achieving the highest identity preservation (IDF1 of 0.385). BoT-SORT follows closely behind them, while DeepOC-SORT and OC-SORT rank lowest across both metrics. This information is also complemented by Fig. 17.

TABLE 10. Tracking algorithm performance.

Algorithm	Avg MOTA	Avg IDF1
ByteTrack	0.188	0.372
BoostTrack	0.188	0.385
BoT-SORT	0.182	0.339
DeepOC-SORT	0.137	0.295
OC-SORT	0.116	0.240

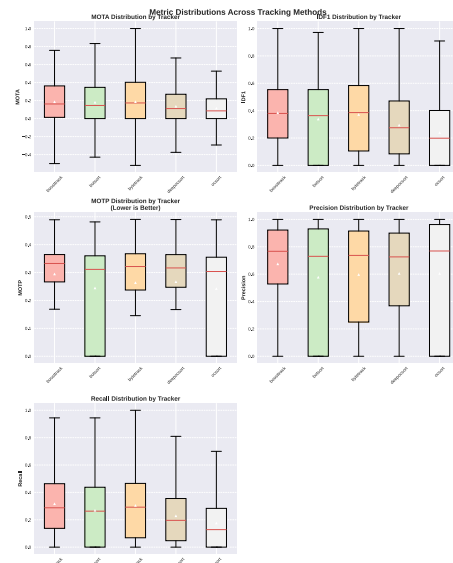


FIGURE 17. Box plot with the metric distributions across the different trackers.

**TABLE 11.** Overall average tracking performance of the best configurations by pre-processing method, YOLO variant, and frequency.

Pre-processing	Frequency (kHz)	MOTA	IDF1	Precision	Recall	Algorithm
Adaptive Gaussian	950	0.458	0.565	0.844	0.473	BoostTrack
Adaptive Gaussian	1200	0.337	0.519	0.703	0.460	ByteTrack
Adaptive Gaussian	1400	0.162	0.329	0.608	0.244	BoT-SORT
Binary 50 Closing	950	0.429	0.542	1.000	0.432	BoostTrack
Binary 50 Closing	1200	0.320	0.495	0.775	0.422	ByteTrack
Binary 50 Closing	1400	0.038	0.208	0.405	0.149	ByteTrack
Binary 250 Closing	950	0.198	0.314	0.496	0.241	ByteTrack
Binary 250 Closing	1200	0.324	0.511	0.763	0.445	BoT-SORT
Binary 250 Closing	1400	0.065	0.217	0.436	0.154	BoT-SORT
Normal	950	0.251	0.401	0.728	0.333	BoostTrack
Normal	1200	0.332	0.528	0.762	0.452	ByteTrack
Normal	1400	0.320	0.473	0.744	0.411	BoostTrack

Table 11 shows the joint effect of pre-processing method and frequency. Adaptive Gaussian thresholding and opening morphology pre-processing generally provides the strongest performance at lower frequencies, while unprocessed inputs (Normal) become the most reliable at 1400 kHz. Although the 950 kHz setting often shows the highest MOTA, it is important to note that some targets were not perceived at this frequency, resulting in missing detections and artificially lower error scores. This makes 1200 kHz and 1400 kHz more reliable indicators of real-world robustness.

At the 950 kHz frequency, the best-performing configuration was the nano model with Adaptive Gaussian thresholding and opening morphology pre-processing combined with BoostTrack, which achieved a MOTA of 0.458 and an IDF1 of 0.565. This configuration also exhibited strong precision (0.844) and recall (0.473), confirming its ability to consistently detect and track targets with stability. In contrast, the weakest performer at this frequency was the nano model with Binary 50 Closing and OC-SORT, which only reached a MOTA of 0.050 and an IDF1 of 0.075. Its critically low recall (0.052) highlighted severe difficulties in consistently detecting objects.

At the 1200 kHz frequency, the strongest performance was obtained by the nano model with Adaptive Gaussian pre-processing in combination with ByteTrack, which reached a MOTA of 0.337 and an IDF1 of 0.519. The configuration achieved good precision (0.703) and recall (0.460), indicating that it maintained the most reliable trajectories over these sequences. The weakest performer at this frequency employed the nano model variant with unprocessed (Normal) input paired with OC-SORT, which achieved only 0.159 MOTA. Its IDF1 dropped to 0.308, and recall fell to 0.217, exposing limitations in tracking targets without the aid of morphological pre-processing.

At the 1400 kHz frequency, the medium YOLOv8 model with unprocessed (Normal) input and BoostTrack stood out as the best performer, achieving a MOTA of 0.320 and an IDF1 of 0.473. The balance between precision (0.744) and recall (0.411) highlights the robustness of abandoning thresholding techniques at the highest acoustic frequency. By contrast, the weakest performance was observed with the medium variant model combined with Binary 50 Closing

and BoostTrack, which achieved an abysmal MOTA of  $-0.128$  and  $0.203$  IDF1. Recall was particularly low at  $0.158$ , making this configuration the weakest overall across all frequencies and pre-processing settings.

Contrary to initial assumptions, the nano YOLOv8 models dominate the best-performing configurations at 950 kHz and 1200 kHz when paired with strong pre-processing (Adaptive Gaussian) and trackers like BoostTrack and ByteTrack. However, as the acoustic signal degrades at 1400 kHz, the medium model paired with unprocessed raw data proved to be the most resilient tracking configuration overall.

Table 12 establishes a direct correlation between the most successful detection configurations (characterised by high training instance counts and high mAP@50 scores) and their corresponding trajectory tracking capabilities.

The results clearly demonstrate that supplying the tracking algorithms with better detections translates directly into robust trajectory management. For instance, the fish net class at 950 kHz under Adaptive Gaussian pre-processing achieved a detection of mAP@50 of 0.917 (backed by 162 training instances). When passed to the tracking algorithms, this sequence has a reliable average MOTA of 0.526, peaking at a value of 0.696 when paired with BoT-SORT.

Similarly, other classes with high training availability, such as PVC Blue Square (108 counts) and PVC perforated deck (113 counts), provided a solid foundation for the trackers. At the 1200 kHz acoustic frequency, combinations like Binary 250 Closing on PVC Blue square translated a  $0.746$  mAP@50 directly into a peak tracking MOTA of  $0.547$  with ByteTrack. Across all top-tier detector configurations, BoT-SORT and ByteTrack emerged as the superior trackers, efficiently leveraging the high-quality bounding boxes to maintain consistent object identities without fragmenting the trajectories.

## A. DISCUSSION

The evaluation of online MOT methods relies on several key metrics, with current challenges including distinguishing visually similar objects, handling scale variation, and managing complex motion while meeting real-time constraints. By integrating frequency, pre-processing, and tracker

TABLE 12. Tracking performance corresponding to the top detector configurations.

Pre-processing	Class	Freq (kHz)	Train Count	mAP@50	Avg MOTA	Best MOTA	Best Tracker
Adaptive Gaussian	Fish_net	950	162	0.917	0.526	0.696	BoT-SORT
Normal	Fish_net	950	162	0.914	0.359	0.509	BoT-SORT
Binary 50 Closing	Fish_net	950	162	0.850	0.362	0.519	ByteTrack
Adaptive Gaussian	PVC_Blue_Square	950	108	0.844	0.539	0.667	BoT-SORT
Binary 250 Closing	Fish_net	950	162	0.798	0.319	0.490	ByteTrack
Binary 250 Closing	PVC_Blue_Square	1200	116	0.746	0.307	0.547	ByteTrack
Binary 50 Closing	PVC_perforated_deck	1200	113	0.743	0.302	0.500	ByteTrack
Normal	PVC_perforated_deck	1200	113	0.741	0.190	0.333	ByteTrack
Binary 50 Closing	PVC_Blue_Square	1200	116	0.740	0.286	0.453	BoT-SORT
Adaptive Gaussian	PVC_perforated_deck	1200	113	0.710	0.272	0.462	BoT-SORT
Binary 250 Closing	PVC_perforated_deck	1200	113	0.689	0.275	0.490	BoT-SORT
Adaptive Gaussian	Fish_net	1200	194	0.676	0.389	0.669	BoT-SORT
Binary 250 Closing	Fish_net	1200	194	0.652	0.302	0.595	BoT-SORT

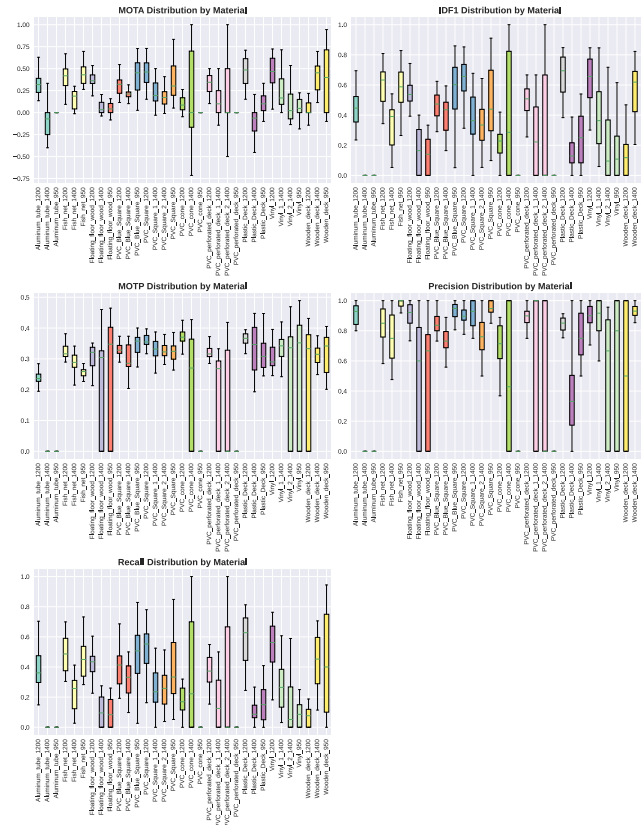


FIGURE 18. Box plot with the metric distributions across the different materials.

comparisons, the analysis highlights how specific design choices shape tracking reliability.

The distribution analysis across video sequences (Fig. 18) shows that MOTA is highly dependent on both material and acoustic frequency. While 950 kHz produced some of the highest peak performances for well-represented classes (as seen in the high detection accuracies in Table 12), its overall average MOTA was low (0.178 in Table 9) due to severe missed detections in underrepresented, reflective materials like PVC cones and aluminium tubes as illustrated in Fig. 10. More realistic and stable robustness across the entire dataset is achieved at 1200 kHz, where the average MOTA peaked

at 0.259. At this frequency, the nano model with Adaptive Gaussian thresholding and ByteTrack achieved the strongest stable result of 0.337 MOTA and 0.519 IDF1, with precision and recall of 0.703 and 0.460 respectively (Table 11). At 1400 kHz, although average performance dropped heavily (0.061 MOTA), the medium variant model with unprocessed (Normal) inputs and BoostTrack still achieved a competitive 0.320 MOTA and 0.473 IDF1, balancing precision (0.744) with moderate recall (0.411).

Pre-processing had a pronounced effect on robustness at lower frequencies. Adaptive Gaussian thresholding and opening morphology consistently improved MOTA for

well-represented targets, peaking at a 0.458 average with BoostTrack at 950 kHz, providing solid foundational bounding boxes for the trackers (Table 11). Binary thresholding methods showed intermediate behaviour. In contrast to initial expectations, unprocessed (Normal) inputs proved highly resilient at higher noise levels, emerging as the best configuration at 1400 kHz. The weakest overall configurations were often those applying rigid binary morphological operations in highly degraded acoustic conditions (e.g., Binary 50 Closing with ByteTrack at 1400 kHz only reached 0.038 MOTA and 0.208 IDF1).

IDF1 values generally mirrored MOTA across configurations, confirming that identity maintenance is highly sensitive to acoustic noise and target occlusion. BoostTrack and ByteTrack achieved the highest average MOTA (0.188 in Table 10), with BoostTrack showing the highest identity stability (average IDF1 of 0.385) (Fig. 17). BoT-SORT followed closely behind (0.182 MOTA, 0.339 IDF1), illustrating that it preserves identity well in favourable conditions. OC-SORT and DeepOC-SORT lagged significantly in both MOTA and IDF1, exposing their struggle to maintain identities when initial detections are fragmented.

MOTP distributions (Fig. 14) confirmed that most trackers maintained stable bounding box alignment, with only narrow materials showing higher error spreads. Pre-processing generally reduced misalignment at lower frequencies, especially under Adaptive Gaussian thresholding and opening morphology, reinforcing its stabilising role.

Precision was generally much higher than recall across configurations, confirming that false positives were relatively rare compared to missed detections. Recall varied widely, with values as low as 0.052 in weak configurations (nano with Binary 50 Closing and OC-SORT at 950 kHz), underscoring the dataset's main challenge of consistently detecting all targets. Stronger pipelines, such as the nano model with Adaptive Gaussian pre-processing and BoostTrack at 950 kHz, achieved a recall of 0.473, demonstrating the importance of pairing effective pre-processing with robust data association algorithms.

Tracker-wise, BoostTrack and ByteTrack emerged as the most consistent across materials and frequencies, achieving the highest average MOTA and tight distributions in IDF1 (Table 10). BoT-SORT followed closely, showing robustness but slightly lower stability across the more degraded sequences. DeepOC-SORT and OC-SORT consistently ranked lowest, with broader distributions and weaker recall, struggling to handle the intermittent detections inherent to the acoustic data.

Overall, the combination of nano YOLOv8 detectors, Adaptive Gaussian pre-processing, and trackers like BoostTrack or ByteTrack produced the most reliable results at lower frequencies (950 kHz and 1200 kHz). This success was heavily dependent on the volume of training annotations, as seen in the strong detector-to-tracker pipeline for well-represented classes like fish net and PVC blue square.

While the selected tracking-by-detection framework utilises architectures natively designed for online tracking, the primary objective was to evaluate tracking accuracy, identity preservation, and data association robustness under varying acoustic frequencies and preprocessing strategies. Because operational latency is highly hardware-dependent, strict real-time profiling is reserved for future deployment phases on embedded robotic systems.

## VI. CONCLUSION

This work presented a comprehensive evaluation of state-of-the-art online MOT algorithms applied to multibeam echosounder water column imagery for marine litter monitoring. To our knowledge, this is the first systematic benchmarking of SORT-family trackers integrated into a TBD pipeline for acoustic imaging data.

The study demonstrated that detection quality and the selected preprocessing strategy are the dominant factors governing tracking performance. The results established a direct correlation between the volume of class-specific training annotations, the resulting detector accuracy (mAP@50), and the ultimate stability of the tracking trajectory. Furthermore, optimal preprocessing was highly dependent on the acoustic frequency. Adaptive Gaussian thresholding and opening morphology consistently improved robustness at lower frequencies (950 kHz and 1200 kHz) by reducing clutter and strengthening object boundaries. At the 1400 kHz frequency, the unprocessed (raw) inputs emerged as the most resilient configuration.

Across the evaluated trackers, BoostTrack and ByteTrack emerged as the most consistent and reliable methods. BoostTrack provided the highest identity preservation (IDF1), benefiting from its robust data association mechanisms, while ByteTrack effectively leveraged the bounding boxes to maintain the highest overall tracking accuracy (MOTA). In contrast, OC-SORT and DeepOC-SORT underperformed across the dataset; their reliance on observation-centric updates struggled to handle the intermittent, highly fragmented, and fluctuating nature of acoustic blob targets.

The results also highlighted clear frequency-dependent behaviour and model size trade-offs. The 950 kHz frequency produced artificially high peak MOTA scores for well-represented objects, but suffered from severe missed detections on narrow materials. The 1200 kHz frequency provided the most reliable balance of resolution and noise, yielding the highest average MOTA and processing the most stable frames across the dataset. At both of these lower frequencies, the nano YOLOv8 models dominated the top-performing configurations. However, at 1400 kHz, where increased backscatter and acoustic clutter drastically reduced overall tracking stability, the medium YOLOv8 models were required to salvage tracking performance.

The current findings are tied to a specific MBES setup and acquisition scenario, generalising the optimal tracker-detector combinations or the observed frequency-dependent behaviours to vastly different sonar

hardware platforms, varying depths, or different open-water environments will require further empirical validation. Within this context, the study confirms that target recall remains the main bottleneck in sonar-based tracking pipelines. Missed detections caused by occlusions, variable incidence angles, weak target backscatter, or acoustic fragmentation propagate directly into lower MOTA scores and identity fragmentation. This reflects the intrinsic challenges of multibeam water-column imaging and indicates that tracking performance is bounded by detector robustness.

Overall, the findings confirm the feasibility of applying modern MOT techniques to underwater acoustic imagery and establish a foundation for more advanced sensing and learning strategies aimed at large-scale, automated marine litter monitoring.

Future work may explore multi-detector ensembles or transformer-based sonar detection models to further quantify how detector–tracker coupling affects MOT performance. Other research directions should focus on developing larger and more diverse annotated sonar datasets to overcome the recall bottlenecks identified here, extending to multi-sensor approaches, multi-modal tracking, and exploring 3D representations of marine litter distributions in the water column. A promising line of work involves the integration of unsupervised or self-supervised learning methodologies. These approaches could capture the intrinsic structure of sonar data and feed into the most effective tracker–detector combinations identified in this study, reducing the dependency on costly manual annotations while enabling more adaptive and scalable monitoring systems. Finally, while the algorithms evaluated here are architecturally suited for online processing, integrating the pipeline into the constrained hardware of AUVs/ASVs to evaluate strict end-to-end real-time performance remains a critical next step. Testing under closed-loop field conditions will provide necessary insights into hardware-dependent latency, tracking stability, and operational usability.

## APPENDIX: HYPERPARAMETERS FOR BoxMOT TRACKERS

This appendix contains the set of hyperparameters that were configured for each tracker that led to the best results. A short description is provided for each parameter. Some of the parameters, although with different names, share similarities between trackers.

### A. DeepOCSort

- **det\_thresh:** Detection confidence threshold to include boxes.
- **max\_age:** Frames a track can persist without detection.
- **min\_hits:** Minimum consecutive detections to confirm a track.
- **iou\_thresh:** IoU threshold for matching detections to tracks.
- **delta\_t:** Maximum frame gap for association.
- **asso\_func:** Choice of association metric, e.g. iou.

TABLE 13. DeepOCSort hyperparameters.

Parameter	Selected	Type / Range
det_thresh	0.5	uniform [0.3, 0.6]
max_age	30	randint [10, 60] step size of 10
min_hits	3	randint [1, 6]
iou_thresh	0.4	uniform [0.1, 0.4]
delta_t	3	randint [1, 6]
asso_func	iou	choice {iou, giou, diou, ciou, hmiou}
inertia	0.2	uniform [0.1, 0.4]
w_association_emb	0.75	uniform [0.5, 0.9]
alpha_fixed_emb	0.95	uniform [0.9, 0.999]
aw_param	0.5	uniform [0.3, 0.7]
embedding_off	false	choice {true, false}
cmc_off	false	choice {true, false}
aw_off	false	choice {true, false}
Q_xy_scaling	0.01	uniform [0.01, 1]
Q_s_scaling	0.0001	uniform [0.0001, 1]

TABLE 14. ByteTrack hyperparameters.

Parameter	Selected	Type / Range
min_conf	0.3	uniform [0.1, 0.3]
track_thresh	0.6	randint [0.4, 0.6]
track_buffer	30	randint [10, 61] step size of 10
match_thresh	0.9	uniform [0.7, 0.9]
frame_rate	sequence dependent	int > 0

- **inertia:** Weight of previous velocity for prediction.
- **w\_association\_emb / alpha\_fixed\_emb / aw\_param:** Appearance embedding weights and scaling.
- **embedding\_off / cmc\_off / aw\_off:** Flags to disable embeddings, camera motion compensation, or adaptive weighting.
- **Q\_xy\_scaling / Q\_s\_scaling:** Kalman filter process noise scaling for position/scale.

### B. ByteTrack

- **min\_conf:** Minimum detection confidence.
- **track\_thresh:** Threshold between high and low-confidence detections.
- **track\_buffer:** Frames to retain unmatched tracks.
- **match\_thresh:** Threshold for track-detection similarity.
- **frame\_rate:** Tracker frame rate.

### C. BoT-SORT

- **track\_high\_thresh / track\_low\_thresh:** Detection confidence thresholds.
- **new\_track\_thresh:** Confidence required to start a new track.
- **track\_buffer:** Frames to keep unmatched tracks.
- **match\_thresh / proximity\_thresh / appearance\_thresh:** Matching thresholds.
- **cmc\_method:** Camera motion compensation.

### D. BoostTrack

- **max\_age / min\_hits:** Track survival and confirmation.
- **det\_thresh / iou\_threshold:** Detection inclusion and matching.

TABLE 15. BoT-SORT hyperparameters.

Parameter	Selected	Type / Range
track_high_thresh	0.6	uniform [0.3,0.7]
track_low_thresh	0.3	uniform [0.1,0.3]
new_track_thresh	0.7	uniform [0.1,0.8]
track_buffer	30	randint [20,81]
match_thresh	0.8	uniform [0.1,0.9]
proximity_thresh	0.5	uniform [0.25,0.75]
appearance_thresh	0.25	uniform [0.1,0.8]
cmc_method	ecc	choice {sof,ecc}

- **use\_ecc**: Enable ECC-based camera compensation.
- **min\_box\_area / aspect\_ratio\_thresh**: Bounding box filters.
- **lambda\_iou / lambda\_mhd / lambda\_shape**: Matching cost weights.
- **use\_dlo\_boost / use\_duo\_boost / dlo\_boost\_coef / s\_sim\_corr / use\_rich\_s / use\_sb / use\_vt**: Motion boosting strategies.
- **with\_reid**: Enable ReID features.

TABLE 16. BoostTrack hyperparameters.

Parameter	Selected	Type / Range
max_age	30	uniform [15,90]
min_hits	3	uniform [1,5]
det_thresh	0.6	uniform [0.1,0.9]
iou_threshold	0.6	uniform [0.1,0.9]
use_ecc	false	choice {false,true}
min_box_area	10	uniform [5,100]
aspect_ratio_thresh	1.6	uniform [0.1,2.0]
lambda_iou	0.5	uniform [0.3,2.0]
lambda_mhd	0.25	uniform [0.5,2.0]
lambda_shape	0.25	uniform [0.5,2.0]
use_dlo_boost	true	choice {false,true}
use_duo_boost	true	choice {false,true}
dlo_boost_coef	0.65	uniform [0.3,2.0]
s_sim_corr	false	choice {false,true}
use_rich_s	true	choice {false,true}
use_sb	true	choice {false,true}
use_vt	true	choice {false,true}
with_reid	true	choice {false,true}

TABLE 17. OCSort hyperparameters.

Parameter	Selected	Type / Range
min_conf	0.3	uniform [0.1,0.3]
det_thresh	0.6	uniform [0,0.6]
max_age	30	grid_search [10,20,30,40,50,60]
min_hits	3	grid_search [1,2,3,4,5]
delta_t	3	grid_search [1,2,3,4,5]
asso_func	iou	choice {iou, giou, diou, ciou, hmiou}
use_byte	false	choice {true,false}
inertia	0.1	uniform [0.1,0.4]
Q_xy_scaling	0.01	loguniform [0.01,1]
Q_s_scaling	0.0001	loguniform [0.0001,1]

### E. OCSort

- **min\_conf / det\_thresh**: Detection thresholds.
- **max\_age / min\_hits / delta\_t**: Track lifespan and gating.

- **asso\_func**: Association metric.
- **use\_byte**: Enable ByteTrack-style association.
- **inertia / Q\_xy\_scaling / Q\_s\_scaling**: Kalman filter parameters.

### REFERENCES

- [1] K. Topouzelis, D. Papageorgiou, G. Suaria, and S. Aliani, "Floating marine litter detection algorithms and techniques using optical remote sensing data: A review," *Mar. Pollut. Bull.*, vol. 170, Sep. 2021, Art. no. 112675.
- [2] J. Yao, Y. Liu, Z. Gu, L. Zhang, and Z. Guo, "Deconstructing PET: Advances in enzyme engineering for sustainable plastic degradation," *Chem. Eng. J.*, vol. 497, Oct. 2024, Art. no. 154183.
- [3] L. Alizadeh, M. C. Liscio, and P. Sospiro, "The phenomenon of greenwashing in the fashion industry: A conceptual framework," *Sustain. Chem. Pharmacy*, vol. 37, 2024, Art. no. 101416. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2352554123004515>
- [4] S. Freitas, H. Silva, and E. Silva, "Hyperspectral imaging zero-shot learning for remote marine litter detection and classification," *Remote Sens.*, vol. 14, no. 21, p. 5516, Nov. 2022.
- [5] J. K. P. Edward, M. Jayanthi, H. A. Einarsson, R. Kannan, R. L. Laju, K. I. Jeyasanta, N. Sathish, and J. Patterson, "Assessment of beach litter, including abandoned, lost, or discarded fishing gear (ALDFG), along the coast of Tamil Nadu, India: Magnitude, sources, composition, pollution status, and management strategies," *Mar. Pollut. Bull.*, vol. 213, Apr. 2025, Art. no. 117700.
- [6] F. Galgani, A. L. Lusher, J. Strand, M. L. Haarr, M. Vinci, E. Molina Jack, R. Kagi, S. Aliani, D. Herzke, V. Nikiforov, S. Primpke, N. Schmidt, J. Fabres, B. De Witte, V. S. Solbakken, and B. van Bavel, "Revisiting the strategy for marine litter monitoring within the European marine strategy framework directive (MSFD)," *Ocean Coastal Manage.*, vol. 255, Sep. 2024, Art. no. 107254.
- [7] W. Zeng, R. Li, H. Zhou, and T. Zhang, "Underwater target tracking method based on forward-looking sonar data," *J. Mar. Sci. Eng.*, vol. 13, no. 3, p. 430, Feb. 2025.
- [8] D. V. Politikos, A. Adamopoulou, G. Petasis, and F. Galgani, "Using artificial intelligence to support marine macrolitter research: A content analysis and an online database," *Ocean Coastal Manage.*, vol. 233, Feb. 2023, Art. no. 106466.
- [9] M. Valdenegro-Toro, "Submerged marine debris detection with autonomous underwater vehicles," in *Proc. Int. Conf. Robot. Autom. Humanitarian Appl. (RAHA)*, Dec. 2016, pp. 1–7.
- [10] A. Aleem, S. Tehsin, S. Kausar, and A. Jameel, "Target classification of marine debris using deep learning," *Intell. Autom. Soft Comput.*, vol. 32, no. 1, pp. 73–85, 2022.
- [11] K. Xie, J. Yang, and K. Qiu, "A dataset with multibeam forward-looking sonar for underwater object detection," *Sci. Data*, vol. 9, no. 1, p. 739, Dec. 2022.
- [12] E. McCann, L. Li, K. Pangle, N. Johnson, and J. Eickholt, "An underwater observation dataset for fish classification and Fishery assessment," *Sci. Data*, vol. 5, no. 1, pp. 1–8, Oct. 2018.
- [13] D. Singh and M. Valdenegro-Toro, "The marine debris dataset for forward-looking sonar semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 3734–3742.
- [14] P. A. Guedes, H. Silva, S. Wang, A. Martins, J. M. Almeida, and E. Silva, "Multibeam multi-frequency characterization of water column litter," in *Proc. OCEANS*, Apr. 2024, pp. 1–6.
- [15] P. A. Guedes, H. M. Silva, S. Wang, A. Martins, J. Almeida, and E. Silva, "Acoustic imaging learning-based approaches for marine litter detection and classification," *J. Mar. Sci. Eng.*, vol. 12, no. 11, p. 1984, Nov. 2024.
- [16] G. N. Williams, G. E. Lagace, and A. Woodfin, "A collision avoidance controller for autonomous underwater vehicles," in *Proc. Symp. Auto. Underwater Vehicle Technol.*, 1990, pp. 206–212.
- [17] D. Dai, "A spatial-temporal approach for segmentation of moving and static objects in sector scan sonar image sequences," in *Proc. 5th Int. Conf. Image Process. its Appl.*, vol. 1995, 1995, pp. 163–167.
- [18] I. Quidu, L. Jaulin, A. Bertholom, and Y. Dupas, "Robust multitarget tracking in forward-looking sonar image sequences using navigational data," *IEEE J. Ocean. Eng.*, vol. 37, no. 3, pp. 417–430, Jul. 2012.
- [19] K. J. DeMarco, M. E. West, and A. M. Howard, "Sonar-based detection and tracking of a diver for underwater human-robot interaction scenarios," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, Oct. 2013, pp. 2378–2383.

- [20] X. Ye, Y. Sun, and C. Li, "FCN and Siamese network for small target tracking in forward-looking sonar images," in *Proc. OCEANS MTS/IEEE Charleston*, Oct. 2018, pp. 1–6.
- [21] D. E. Clark and J. Bell, "Bayesian multiple target tracking in forward scan sonar images using the PHD filter," *IEE Proc. - Radar, Sonar Navigat.*, vol. 152, no. 5, pp. 327–334, Oct. 2005.
- [22] D. Musicki, X. Wang, R. Ellem, and F. Fletcher, "Efficient active sonar multitarget tracking," in *Proc. OCEANS - Asia-Pacific*, May 2006, pp. 1–8.
- [23] N. Hurtós, N. Palomeras, A. Carrera, and M. Carreras, "Autonomous detection, following and mapping of an underwater chain using sonar," *Ocean Eng.*, vol. 130, pp. 336–350, Jan. 2017.
- [24] D. M. Lane, M. J. Chantler, and D. Dai, "Robust tracking of multiple objects in sector-scan sonar image sequences using optical flow motion estimation," *IEEE J. Ocean. Eng.*, vol. 23, no. 1, pp. 31–46, Jan. 1998.
- [25] M. J. Chantler and J. P. Stoner, "Automatic interpretation of sonar image sequences using temporal feature measures," *IEEE J. Ocean. Eng.*, vol. 22, no. 1, pp. 47–56, Jan. 1997.
- [26] I. Tena Ruiz, D. M. Lane, and M. J. Chantler, "A comparison of inter-frame feature measures for robust object classification in sector scan sonar image sequences," *IEEE J. Ocean. Eng.*, vol. 24, no. 4, pp. 458–469, Oct. 1999.
- [27] S. W. Perry and L. Guan, "A recurrent neural network for detecting objects in sequences of sector-scan sonar images," *IEEE J. Ocean. Eng.*, vol. 29, no. 3, pp. 857–871, Jul. 2004.
- [28] I. Karoui, I. Quidu, and M. Legrís, "Automatic sea-surface obstacle detection and tracking in forward-looking sonar image sequences," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 8, pp. 4661–4669, Aug. 2015.
- [29] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9626–9635.
- [30] L. Du, R. Zhang, and X. Wang, "Overview of two-stage object detection algorithms," *J. Phys., Conf. Ser.*, vol. 1544, no. 1, May 2020, Art. no. 012033.
- [31] S. Hassan, G. Mujtaba, A. Rajput, and N. Fatima, "Multi-object tracking: A systematic literature review," *Multimedia Tools Appl.*, vol. 83, no. 14, pp. 43439–43492, Oct. 2023.
- [32] S. Li, H. Ren, X. Xie, and Y. Cao, "A review of multi-object tracking in recent times," *IET Comput. Vis.*, vol. 19, no. 1, p. 70010, Jan. 2025.
- [33] P. Voigtlaender, M. Krause, A. Osep, J. Luiten, B. B. G. Sekar, A. Geiger, and B. Leibe, "MOTS: Multi-object tracking and segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7934–7943.
- [34] K. Huang, K. Lertniphonphan, F. Chen, J. Li, and Z. Wang, "Multi-object tracking by self-supervised learning appearance model," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2023, pp. 3163–3169.
- [35] Z. Wang, L. Zheng, Y. Liu, Y. Li, and S. Wang, "Towards real-time multi-object tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 107–122.
- [36] Y. Zhang, C. Wang, X. Wang, W. Zeng, and W. Liu, "FairMOT: On the fairness of detection and re-identification in multiple object tracking," *Int. J. Comput. Vis.*, vol. 129, no. 11, pp. 3069–3087, Nov. 2021.
- [37] X. Zhou, V. Koltun, and P. Krähenbühl, "Tracking objects as points," in *Proc. ECCV*, 2020, pp. 474–490.
- [38] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 3464–3468.
- [39] G. Welch and G. Bishop, "An introduction to the Kalman filter," Univ. North Carolina at Chapel Hill, 1995.
- [40] B. Sahbani and W. Adiprawita, "Kalman filter and iterative-hungarian algorithm implementation for low complexity point tracking as part of fast multiple object tracking system," in *Proc. 6th Int. Conf. Syst. Eng. Technol. (ICSET)*, Oct. 2016, pp. 109–115.
- [41] Y. Zhang, P. Sun, Y. Jiang, D. Yu, Z. Yuan, P. Luo, W. Liu, X. Wang, and W. Xinggang, "ByteTrack: Multi-object tracking by associating every detection box," *CoRR*, vol. abs/2110.06864, 2021. [Online]. Available: <https://arxiv.org/abs/2110.06864>
- [42] J. Cao, J. Pang, X. Weng, R. Khirodkar, and K. Kitani, "Observation-centric SORT: Rethinking SORT for robust multi-object tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Mar. 2023, pp. 9686–9696.
- [43] G. Maggolino, A. Ahmad, J. Cao, and K. Kitani, "Deep OC-SORT: Multi-pedestrian tracking by adaptive re-identification," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2023, pp. 3025–3029.
- [44] N. Aharon, R. Orfaig, and B.-Z. Bobrovsky, "BoT-SORT: Robust associations multi-pedestrian tracking," 2022, *arXiv:2206.14651*.
- [45] V. D. Stanojevic and B. Todorović, "BoostTrack: Boosting the similarity measure and detection confidence for improved multiple object tracking," *Mach. Vis. Appl.*, vol. 35, no. 3, p. 53, 2024.
- [46] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 3645–3649.
- [47] L. Porzi, M. Hofinger, I. Ruiz, J. Serrat, S. R. Bulo, and P. Kotschieder, "Learning multi-object tracking and segmentation from automatic annotations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6845–6854.
- [48] Z. Xu, W. Zhang, X. Tan, W. Yang, X. Su, Y. Yuan, H. Zhang, S. Wen, E. Ding, and L. Huang, "PointTrack++ for effective online multi-object tracking and segmentation," 2020, *arXiv:2007.01549*.
- [49] P. Sun, J. Cao, Y. Jiang, R. Zhang, E. Xie, Z. Yuan, C. Wang, and P. Luo, "TransTrack: Multiple object tracking with transformer," 2020, *arXiv:2012.15460*.
- [50] T. Meinhardt, A. Kirillov, L. Leal-Taixe, and C. Feichtenhofer, "TrackFormer: Multi-object tracking with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 8844–8854.
- [51] Y. Zhang, T. Wang, and X. Zhang, "MOTRv2: Bootstrapping end-to-end multi-object tracking by pretrained object detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 22056–22065.
- [52] J. Cai, M. Xu, W. Li, Y. Xiong, W. Xia, Z. Tu, and S. Soatto, "MeMOT: Multi-object tracking with memory," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 8090–8100.
- [53] R. Gao and L. Wang, "MeMOTR: Long-term memory-augmented transformer for multi-object tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 9867–9876.
- [54] S. Chun, C. Kawamura, K. Ohkuma, and T. Maki, "3D detection and tracking of a moving object by an autonomous underwater vehicle with a multibeam imaging sonar: Toward continuous observation of marine life," *IEEE Robot. Autom. Lett.*, vol. 9, no. 4, pp. 3037–3044, Apr. 2024.
- [55] B. Sun, W. Zhang, C. Xing, and Y. Li, "Underwater moving target detection and tracking based on enhanced you only look once and deep simple online and realtime tracking strategy," *Eng. Appl. Artif. Intell.*, vol. 143, Mar. 2025, Art. no. 109982.
- [56] P. A. Guedes, H. Silva, S. Wang, A. Martins, and J. M. Almeida, "Multibeam acoustic image based detection and tracking of marine litter in the water column," in *Proc. OCEANS Brest*, Jun. 2025, pp. 1–6.
- [57] B. Huang, Y. Song, H. Qin, J. Miao, and C. Zhu, "Safety-enhanced formation maneuver control for electric vehicle with edge-weighted topology and reinforcement learning strategy," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 61, no. 5, pp. 14716–14731, Oct. 2025.
- [58] *M3 Sonar HF-Kongsberg Discovery*. Accessed: Aug. 2025. [Online]. Available: <https://www.kongsberg.com/discovery/sea/floor-mapping/sonars/m3-sonar-hf/>
- [59] M. Nixon and A. Aguado, *Feature Extraction and Image Processing for Computer Vision*. New York, NY, USA: Academic, 2019.
- [60] R. Szeliski, *Computer Vision: Algorithms and Applications*. Cham, Switzerland: Springer, 2022.
- [61] *BoxMOT: Pluggable SOTA Tracking Modules for Object Detection, Segmentation and Pose Estimation Models*. Accessed: Sep. 2025. [Online]. Available: <https://github.com/mikel-brostrom/boxmot>
- [62] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: The CLEAR MOT metrics," *EURASIP J. Image Video Process.*, vol. 2008, pp. 1–10, 2008.

**PEDRO ALVES GUEDES** was born in Porto, Portugal, in 1995. He received the bachelor's degree in electrical and computer engineering from the School of Engineering of Porto, Polytechnic Institute of Porto, in 2016, and the master's degree in autonomous systems from the Polytechnic Institute of Porto, in 2019. He is currently pursuing the Ph.D. degree in electrical and computer engineering with the Faculty of Engineering, University of Porto. He earned a Ph.D. Scholarship from Portuguese Science Foundation (FCT), in 2021. Additionally, he is a Guest Assistant at the School of Engineering of Porto, Polytechnic Institute of Porto. He is actively involved as a Researcher in two projects: MYTAG during the master's degree and

NETTAG+ during the Ph.D. degree in INESC TEC-Instituto de Engenharia de Sistemas e Computadores, Tecnologia e Ciência. His doctoral research focuses on underwater multi-frequency multibeam echosounder acoustic image processing for the classification and detection of marine litter with machine learning.

**HUGO MIGUEL SILVA** received the Lic. degree in electrical engineering–electronics and computers from the Instituto Superior de Engenharia do Porto, in 2004, the M.Sc. and Ph.D. degree in electrical and computer engineering from the Instituto Superior Técnico, Universidade de Lisboa, in 2008 and 2014, respectively. He was a Principal Investigator in several national and European research projects and has contributed to numerous others in the fields of robotics and sensing systems. He is currently a Principal Investigator with INESC TEC and an Invited Adjunct Teacher with the Instituto Politécnico do Porto. He has authored more than 40 scientific publications, including journal articles, conference papers, and book chapters, and has supervised research at the doctoral and master's levels in underwater perception, hyperspectral imaging, and robotic sensing. His research interests include robotics, underwater perception, remote sensing, and AI-based methods for environmental monitoring.

**SEN WANG** (Member, IEEE) is currently an Associate Professor with the Robotics and Autonomous Systems and the Director of the Sense Robotics Laboratory with the Department of Electrical and Electronic Engineering and I-X, Imperial's cross-college flagship initiative in AI. He is the Founding Director of the MSc in artificial intelligence applications and innovation.

His research sits at the intersection of robotics, computer vision and machine learning, driving robots and intelligent machines to understand and operate autonomously in unstructured, dynamic environments through probabilistic and learning approaches. Through the UKRI ORCA Hub (EP/R026173/1 and EP/W001136/1), he led a research team to develop underwater sensing and robotic technologies for autonomous inspection of offshore energy infrastructure, and successfully carried out the first autonomous wind farm foundation inspection at EDF's Blyth Offshore Wind Farm. His main research interests include robot localization, autonomous navigation, SLAM, robot vision, robot learning and their application on real-world robot systems to help tackle the challenges we face in our society, from climate change to healthcare.

Prof. Wang was awarded the 2024 AI Most Influential Scholar Award Honorable Mention in Robotics. He has served as an Associate Editor of IEEE TRANSACTIONS ON ROBOTICS, IEEE TRANSACTIONS ON AUTOMATION SCIENCE AND ENGINEERING, IEEE ROBOTICS AND AUTOMATION LETTERS, ICRA, and IROS. A full list of his publications can be found on Google Scholar.

• • •