Machine Learning with Applications 23 (2026) 100808

Contents lists available at ScienceDirect

Machine Learning with Applications

journal homepage: www.elsevier.com/locate/mlwa

Advancing marine mammal monitoring: Large-scale UAV delphinidae datasets and robust motion tracking for group size estimation

Leonardo Viegas Filipe [a,*], João Canelas [a], Mário Vieira [a], Francisco Correia da Fonseca [a], André Cid [b], Joana Castro [b,c], Inês Machado [a]

[a] WavEC Offshore Renewables, Edifício Diogo Cão, Doca de Alcântara Norte, 1350-352, Lisbon, Portugal
[b] AIMM - Associação para a Investigação do Meio Marinho, Rua Maestro Fred. Freitas N15-1, 1500-399, Lisbon, Portugal
[c] MARE - Marine and Environmental Sciences Centre/ARNET - Aquatic Research Network, Laboratório Marítimo da Guia, Faculdade de Ciências da Universidade de Lisboa, Av. Nossa Senhora do Cabo, 939, 2750-374, Cascais, Portugal

ARTICLE INFO

Keywords:
Artificial intelligence
Marine mammal surveys
UAV
Object tracking
Human-in-the-loop
Dataset

ABSTRACT

Reliable estimates of dolphin abundance are essential for conservation and impact assessment, yet manual analysis of aerial surveys is time-consuming and difficult to scale. This paper presents an end-to-end pipeline for automatic dolphin counting from unmanned aerial vehicle (UAV) video that combines modern object detection and multi-object tracking. We construct a large detection dataset of 64,705 images with 225,305 dolphin bounding boxes and a tracking dataset of 54,274 frames with 207,850 boxes and 603 unique tracks, derived from UAV line-transect surveys. Using these data, we train a YOLO11-based detector that achieves a precision of approximately 0.93 across a range of sea states. For tracking, we adopt BoT-SORT and tune its parameters with a genetic algorithm using a multi-metric objective, reducing ID fragmentation by about 29% relative to default settings. Recent YOLO-based cetacean detectors trained on UAV imagery of beluga whales report precision/recall around 0.92/0.92 for adults and 0.94/0.89 for calves, but rely on DeepSORT tracking whose MOTA remains below 0.5 and must be boosted to roughly 0.7 with post-hoc trajectory post-processing. In this context, our pipeline offers competitive detection performance, substantially larger and fully documented detection and tracking benchmarks, and GA-optimized tracking without manual post-processing. Applied to dolphin group counting, the full pipeline attains a mean absolute error of 1.24 on a held-out validation set, demonstrating that UAV-based automated counting can support robust, scalable monitoring of coastal dolphin populations.

1. Introduction

Marine mammal monitoring is a critical component of conservation and management strategies, as it provides essential information on population size, distribution, and behaviour. Traditional survey methods, such as boat-based counts or manned aerial observations, are resource-intensive, weather-dependent, and can disturb target species. In recent years, drones (unmanned aerial vehicles, UAVs) have emerged as a non-invasive and cost-effective alternative for acquiring high-resolution imagery over broad spatial and temporal scales (Álvarez-González et al., 2023).

In parallel, deep learning has transformed computer vision, enabling reliable detection and tracking of wildlife in large datasets. A dominant paradigm in multi-object tracking is tracking-by-detection, where each frame is first processed by an object detector and detections are then linked into trajectories (Adžemović et al., 2025; Guan et al., 2025; Li et al., 2024). The success of this strategy depends directly on detection quality, as tracking continuity is built upon accurate frame-level localization.

The YOLO (You Only Look Once) object detection model family has established itself as one of the most effective detectors for real-time applications, combining speed and accuracy in a unified, end-to-end model (Redmon et al., 2016). YOLO11 introduces architectural improvements that enhance efficiency and small-object detection (Jegham et al., 2025; Khanam & Hussain, 2024; Ultralytics, 2023), both crucial for identifying dolphins in drone footage. Complementing this, BoT-SORT is a state-of-the-art tracker that extends detection-based tracking by incorporating motion cues, appearance embeddings,

* Corresponding author.
E-mail addresses: leonardo.filipe@wavec.org (L. Viegas Filipe), joao.canelas@wavec.org (J. Canelas), mario.vieira@wavec.org (M. Vieira), francisco.fonseca@wavec.org (F. Correia da Fonseca), andre.cid@aimmportugal.org (A. Cid), joana.castro@aimmportugal.org (J. Castro), ines.machado@wavec.org (I. Machado).

https://doi.org/10.1016/j.mlwa.2025.100808
Received 27 October 2025; Received in revised form 25 November 2025; Accepted 1 December 2025
Available online 4 December 2025

and camera-motion compensation to achieve robust identity preservation (Aharon et al., 2022). The natural synergy between YOLO11 as a high-performing detector and BoT-SORT as an advanced detection-based tracker provides a powerful foundation for monitoring dolphins from UAV imagery.

Despite these advances, progress is still constrained by the limited availability of annotated datasets for model training and validation. Collecting UAV imagery of dolphins is logistically challenging and ecologically costly, and publicly available labelled datasets remain scarce. This creates barriers to developing generalizable, robust systems that can support long-term monitoring programmes rather than isolated case studies.

We advance computer vision and marine ecology with an end-to-end UAV pipeline for dolphin monitoring.

Our main contributions are as follows:

- We curate and release a large-scale dolphin detection dataset (64,705 images with 225,305 bounding boxes) and a tracking dataset (54,274 frames with 207,850 boxes and 603 unique tracks) derived from UAV line-transect surveys, capturing a wide range of sea states and group configurations.
- We train a YOLO11-based detector tailored to coastal dolphin imagery and show that it achieves high precision across variable environmental conditions, providing reliable frame-level localization for downstream tracking.
- We adopt BoT-SORT as our baseline tracker and use a genetic algorithm to tune its hyperparameters with a multi-metric objective combining HOTA, IDF1, and related measures, reducing identity fragmentation compared to default settings.
- We perform a detailed comparison between appearance-based and motion-only tracking variants, and we show that in this domain appearance re-identification does not improve performance over motion-only cues, informing the design of lightweight, field-deployable systems.
- We demonstrate that the resulting detection-and-tracking pipeline can be used to estimate dolphin group sizes from UAV video with low mean absolute error, illustrating the practical value of deep learning for scalable, non-invasive monitoring of coastal dolphin populations.

Together, these contributions provide both new resources and concrete guidance for deploying modern computer-vision pipelines in marine mammal monitoring, and they highlight opportunities and limitations that should be considered when integrating UAV-based automated counting into conservation practice.

This introduction is followed by a section dedicated to review related work done in this field (Section 2). We provide an overview of the proposed method in Section 3, detailing aspects of the different modules and overall architecture of the pipeline. After, we describe all the datasets used (Section 4) regarding acquisition and data characteristics. Next, we describe the developed annotation tool (Section 5). Model implementations are described in Section 6, starting with the training and evaluation of the detector and finishing with the multiple tracking configurations. The most relevant results are presented and discussed in Section 7. Section 8 describes limitation of our pipeline and proposes future work. Finally, Section 9 rounds up this paper presenting final conclusions.

2. Related work

Deep learning has transformed computer vision, enabling reliable detection and tracking of objects in large, heterogeneous datasets. In wildlife monitoring, it has been increasingly adopted to automate tasks such as species detection, individual identification, and movement analysis. A dominant paradigm in multi-object tracking is tracking-by-detection, in which each frame is first processed by an object detector and detections are then linked into trajectories (Adžemović et al., 2025; Guan et al., 2025; Li et al., 2024). The success of this strategy depends directly on detection quality, as tracking continuity is built upon accurate frame-level localization.

Several recent studies have reviewed the rapidly growing literature on deep learning for ecological and marine applications, highlighting both the potential and the limitations of current approaches (Adžemović et al., 2025; Guan et al., 2025; Li et al., 2024). Most existing work either focuses solely on detection in still images or relies on relatively small, proprietary datasets, making it difficult to assess how well models generalize across locations, environmental conditions, and survey protocols. Explicit treatment of the full pipeline from detection and tracking to biologically meaningful quantities such as group size or abundance remains comparatively rare. Recent YOLOv7-based systems have demonstrated high-precision detection of beluga whales from UAV video, achieving precision/recall around 0.92/0.92 for adults and 0.94/0.89 for calves, but rely on DeepSORT tracking whose MOTA must be boosted from 27%–48% to about 70% via post-hoc trajectory post-processing (Alsaidi et al., 2024).

For marine mammals observed from UAVs, additional challenges arise. Animals are often small in the field of view, partially submerged, and embedded in a dynamic sea surface whose appearance changes with sea state, lighting, and viewing geometry. These factors complicate both detection and tracking, and they interact with survey design choices (flight altitude, camera configuration, flight paths). At the same time, UAVs offer clear operational advantages over traditional boat-based or crewed-aircraft surveys, including reduced cost and risk, higher revisit frequency, and the ability to capture stabilized video suitable for automated analysis (Álvarez-González et al., 2023).

A key bottleneck across this literature is the scarcity of large, openly-described datasets that reflect the variability of real monitoring campaigns. Collecting UAV imagery of dolphins is logistically demanding and may require specialized vessels, pilots, and permits. Publicly available labelled datasets are therefore limited. The dataset introduced by Canelas et al. (2025) represents an important step in this direction, providing UAV video of coastal dolphins with frame-level annotations suitable for training and evaluating detection and tracking models under realistic survey conditions. However, even in this case, challenges remain in terms of dataset size, diversity of sea states and group configurations, and the availability of identity-consistent tracks for benchmarking multi-object tracking.

Within this context, our work contributes by extending the available data resources for dolphin monitoring and by systematically evaluating a modern detection-and-tracking pipeline tailored to UAV footage of small delphinids. In contrast to studies that focus only on algorithmic performance, we explicitly link computer-vision metrics to ecological outcomes, using the pipeline to estimate dolphin group sizes from survey video and assessing errors in terms that are directly relevant to abundance estimation and conservation practice.

3. Proposed method and overall architecture

Our goal is to obtain reliable dolphin group counts from UAV survey video using an automated, scalable pipeline. We adopt a tracking-by-detection strategy in which a frame-level detector produces dolphin bounding boxes and a multi-object tracker links these detections into temporally consistent trajectories. Group-level counts are then derived from the resulting tracks.

Fig. 1 provides an overview of the proposed architecture. Starting from raw UAV video, we (i) extract frames and create large-scale detection and tracking datasets using a semi-supervised, human-in-the-loop annotation workflow, (ii) train a YOLO11-based detector on the detection dataset, (iii) apply this detector to new UAV frames to obtain dolphin detections, (iv) use BoT-SORT to track dolphins over time, tuning its hyperparameters with a genetic algorithm, and (v) aggregate track information into frame-level and group-level counts, which are evaluated both with standard computer-vision metrics and with ecologically meaningful error measures.

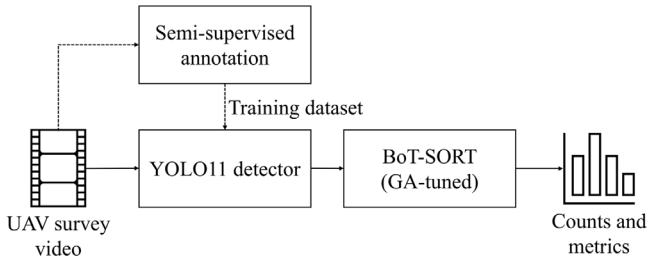


Fig. 1. Overview of the proposed UAV-based dolphin counting pipeline. Raw UAV survey video is first sampled into frames and annotated using a semi-supervised, human-in-the-loop workflow, producing large-scale detection and tracking datasets. A YOLO11-based detector is then trained on the detection dataset and applied to new UAV frames to produce dolphin bounding boxes. BoT-SORT, with hyperparameters tuned by a genetic algorithm, links detections across frames into trajectories. From these trajectories we obtain frame-level and group-level counts, as well as evaluation metrics such as mean absolute error (MAE), HOTA and IDF1, which quantify both counting accuracy and tracking quality.

3.1. Problem formulation

Let a UAV survey produce a video sequence $\{I_t\}_{t=1}^{T}$ of T frames. The primary computer-vision task is to estimate, for each frame I_t, the set of dolphin bounding boxes ($\mathcal{B}$)

$$\mathcal{B}_t = \{b_{t,k}\}_{k=1}^{K_t},$$

where K_t is the number of detected dolphins in frame t. The tracking task is to assign consistent identities across frames and produce a set of trajectories $\mathcal{T}$

$$\mathcal{T} = \{\tau_j\}_{j=1}^{J},$$

where each trajectory τ_j is a time-ordered sequence of detections corresponding to the same individual dolphin or tightly grouped individuals. J is the total number of trajectories in the evaluated video sequence.

These trajectories are then used to estimate dolphin group sizes and group-level abundance over predefined temporal windows or survey segments. In addition to standard detection and tracking metrics (e.g. precision, recall, mean Average Precision, HOTA, IDF1), we report ecological metrics such as mean absolute error (MAE) between predicted and manually derived group counts.

3.2. Detection module

The detection module processes each frame I_t independently and outputs a set of dolphin detections $\mathcal{B}_t$. We use YOLO11L as our base detector, chosen for its balance between speed and accuracy on small objects in high-resolution imagery. The network predicts bounding boxes and confidence scores, which are then filtered by non-maximum suppression to produce the final set of detections per frame.

The detector is trained on the curated detection dataset described in Section 4, using a train/validation split that reflects the diversity of sea states, viewing geometries, and group configurations. Training details, including data augmentation, optimization settings, and selection of decision thresholds, are presented in Section 6.

3.3. Tracking module

To obtain temporally consistent trajectories from frame-level detections, we employ BoT-SORT as our tracking backbone. BoT-SORT follows a tracking-by-detection paradigm in which detections in consecutive frames are associated using a combination of motion modelling and similarity scoring.

In our configuration, the tracker maintains a set of active tracks and at each frame associates detections to existing tracks based on predicted motion and, optionally, appearance cues. Unmatched detections initialize new tracks, while tracks without recent matches are terminated. This yields a set of trajectories $\mathcal{T}$ that capture the movement of dolphins through the field of view.

BoT-SORT performance depends on several hyperparameters governing, for example, association thresholds, the lifespan of unmatched tracks, and matching criteria. Rather than setting these heuristically, we perform a two-stage optimization: a coarse grid search followed by a genetic algorithm that refines hyperparameters using a multi-metric objective combining HOTA, IDF1 and related tracking metrics. The optimization procedure and search space are detailed in Section 6.

Optionally, BoT-SORT can include a deep-learning model to aid on object Re-Identification (ReID) based on appearance cues, allowing the same individual to be recognized across frames. We performed some experiments to evaluate the benefit of appearance-based ReID, and the details of this implementation are presented in Section 6.2.3

3.4. Group counting and evaluation

Once trajectories have been estimated, we derive counts at two levels:

- frame-level counts, obtained by counting the number of active tracks or detections per frame after basic filtering (e.g. removal of very short tracks or low-confidence detections).
- group-level counts, obtained by aggregating trajectories over survey segments corresponding to dolphin groups, following the same protocol used for manual annotations.

For the computer-vision evaluation we use a dedicated tracking benchmark and report standard detection metrics: precision, recall, mean Average Precision at different Intersection over Union (IoU) thresholds; and tracking metrics: HOTA, IDF1, and related measures. For the ecological evaluation we compare predicted group sizes to expert-derived counts and report mean absolute error and related statistics, providing a direct link between algorithmic performance and monitoring accuracy.

This separation between the method-level description in the present section and the implementation details in Section 6 is intended to clarify the overall architecture of the proposed solution while preserving full transparency about training protocols and hyperparameter choices.

4. Datasets

4.1. Data acquisition

We used a dataset compiled by Canelas et al. (2025). Because in-situ ocean surveys are costly and logistically complex, and public aerial datasets for widespread species are scarce, they focused on Delphinidae and scraped 62 videos from YouTube, Pexels, and Dailymotion across unrestricted locations, times, and conditions to build a variable dataset for training generalizable detection/tracking models. Fig. 2 shows some examples of frames extracted from these videos, illustrating scene variability.

Additionally, Associação para a Investigação do Meio Marinho (AIMM) provided one video resultant from a marine mammal observation expedition.

4.2. Semi-supervised detection dataset

We began by manually annotating a seed set of a few hundred images, drawing bounding boxes around dolphins. This initial dataset

Fig. 2. Example of images in the dataset, illustrating scene variability.

was used to train a YOLO11 detector, which was then applied to the remaining unlabelled imagery. The automatically generated predictions were reviewed and corrected by human annotators, producing an expanded, higher-quality training set. The detector was then re-trained using this refined dataset.

This was repeated iteratively: YOLO11 was applied to the full dataset, its predictions corrected, and the model re-trained until the detection error stabilized at an acceptably low level, i.e., every annotator agreed with the annotations of all images. Through this cycle, the annotation burden was significantly reduced compared to fully manual labelling, while maintaining high dataset quality. Similar iterative labelling frameworks have been applied in UAV wildlife monitoring and remote sensing, confirming their suitability in data-scarce contexts (Kellenberger et al., 2018; Zhao et al., 2025). Fig. 3 illustrates this labelling method in the top blue box.

This process resulted in a dataset with a total of 64 705 images with 225 305 annotated boxes, giving an average of approximately 3.5 boxes per image. This dataset was partitioned in training and validation sets, using a 70–30 ratio, using, for each video, the last 30% for validation, keeping the first 70% for training, decreasing frame similarity between these two sets.

4.3. Tracking dataset

The creation of high-quality tracking datasets presents additional challenges due to data scarcity. Unlike established pedestrian or vehicle tracking benchmarks, there are no large-scale, publicly available datasets for dolphin tracking from UAV imagery. This lack of annotated video data constrains the training and evaluation of multi-object trackers. To address this limitation, we adopted a model-assisted annotation workflow. The YOLO11 detector, trained on the semi-supervised dataset, was applied frame by frame, and detections were linked into tracks using the BoT-SORT algorithm. This produced initial identity assignments across sequences. Because automatic tracking can introduce errors, such as identity switches, fragmented trajectories, or false positives, a manual correction stage was performed. Annotators reviewed the results, reassigning IDs where necessary, removing false positives, and refining trajectories. This pipeline enabled the production of a reliable dolphin tracking dataset that would have been infeasible to assemble through manual annotation alone. This process is shown in Fig. 3, specifically in the bottom green box. The resultant dataset consisted of 54 274 annotated frames from 18 different videos. 207 850 boxes made 603 unique tracks. There was an average of approximately 3.8 boxes per frame.

Table 1
Number of frames and total dolphin count (IDs_{GT}) on the AIMM dataset.

Seq.	Frames	IDs_{GT}
0	727	14
1	407	9
2	1411	12
3	681	16
4	380	11
5	511	2
	4117	**64**

4.4. AIMM dataset

AIMM's video was annotated for tracking using the process described in Section 4.3, and segmented into six Sequences (Seqs.) with varying number of frames, dolphin total count (Table 1), and drone height.

4.5. Sample taxonomy and background variability

Taxonomically, all positive annotations in the current release correspond to delphinids; no other marine mammal or fish species are explicitly labelled. The aerial footage was collected from various internet sources, containing footage taken on nearshore and offshore waters under a range of locations, sea states and illumination conditions, with frequent whitecaps, sun glint and vessel wakes. In addition to dolphin annotations, the raw imagery contains a variety of non-dolphin objects (e.g. surfers, kayakers, fish schools, birds, watercraft, algae and rocks), which are retained as unlabelled negatives and are illustrated in Fig. 4. Implications for domain shift are discussed in Section 8.

4.6. Negative samples

In the current release of these datasets we only annotate dolphins; no other marine mammal or fish species are included in this version. In practice, our aerial footage for this study does not contain other cetacean species in appreciable numbers, so the main challenge is distinguishing dolphins from sea surface clutter (whitecaps, wave crests, wakes, sun glint) rather than from other animals. To provide sufficient negative examples, we explicitly keep frames and image regions without dolphins in the training sets, including scenes with surfers or kayakers, fish schools, birds, watercrafts, dolphin looking algae and rocks (Fig. 4). These act as hard negatives and help the detector learn to ignore non-dolphin objects with similar size and contrast. In future

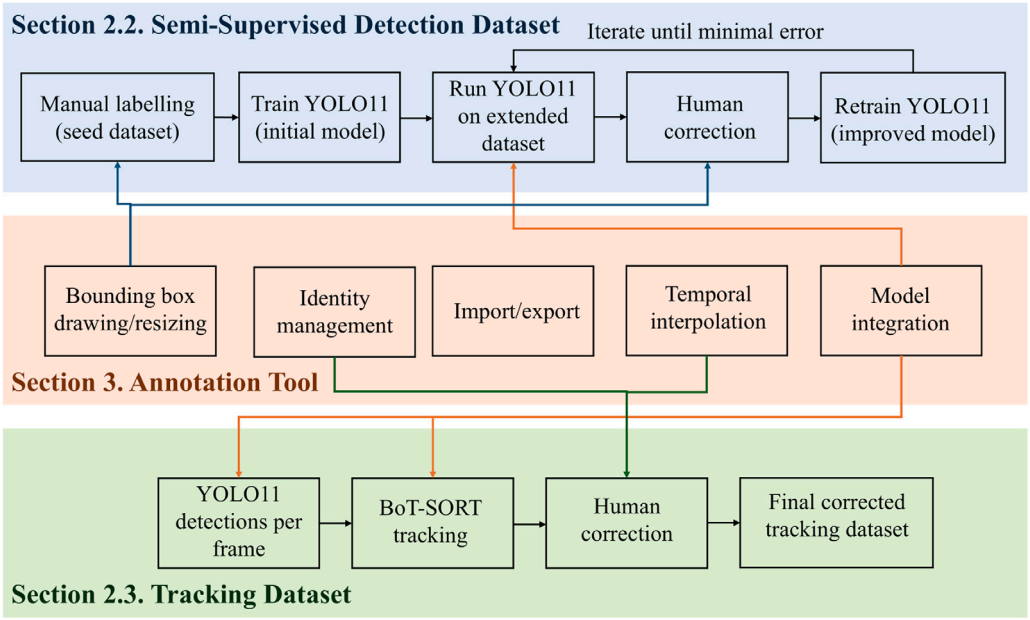


Fig. 3. Flowchart illustrating the process of dataset creation and how the annotation tool (Section 5) aids in the development. The top blue box illustrates the creation of the detection dataset using a human-in-the-loop procedure. The bottom green box represents the creation of the tracking dataset. The middle orange box summarizes the capabilities of the annotation tool, linking them to the tasks they aid in the creation of both datasets. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

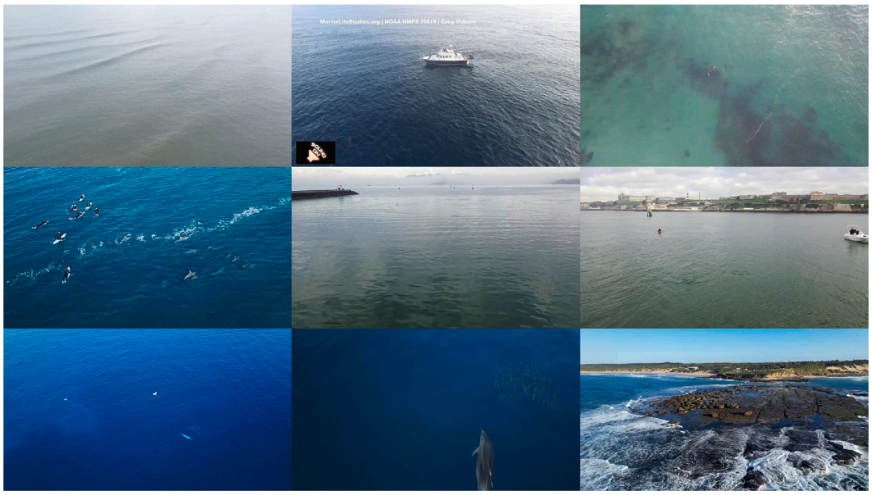


Fig. 4. Examples of images with negative samples, e.g., images with no dolphins, surfers or kayakers mixed with dolphins, fish schools, birds, watercrafts, dolphin looking algae and rocks.

work, as more multi-species data become available, we plan to extend the dataset and annotations to additional marine species to further stress-test cross-species generalization. Implications for domain shift are discussed in Section 8.

5. Annotation tool

To facilitate efficient correction of detections and tracks, we developed a custom annotation tool implemented in PyQt5 python framework. The tool integrates manual labelling, model-assisted refinement, and identity management. The main innovation of the tool lies in the integration of functionalities that were previously scattered across different annotation platforms. We combined the most useful features found in widely used tools such as CVAT (CVAT.ai Corporation, 2024), VIA (Dutta & Zisserman, 2019), and DarkLabel (Programmer, 2025) into a single environment designed to make the annotation process more effective. By consolidating these capabilities into one system, the created tool directly addresses the specific needs of object tracking annotation, significantly reducing manual effort while increasing flexibility in the annotation process. The middle orange box in Fig. 3 illustrates how the tool functionalities directly interact with the annotation processes. Fig. 5 shows the graphical interface of the tool: it consists of two main elements: the frame display and the side menu. The display shows the frame and respective imported annotations. The user annotates boxes and IDs in this area. The side menu consists of buttons with specific functions, e.g., import and export annotations ("Import/Generate Ground-Truth CSV", "Import/Generate predictions CSV"), and display toggles ("Display predictions/IDs").

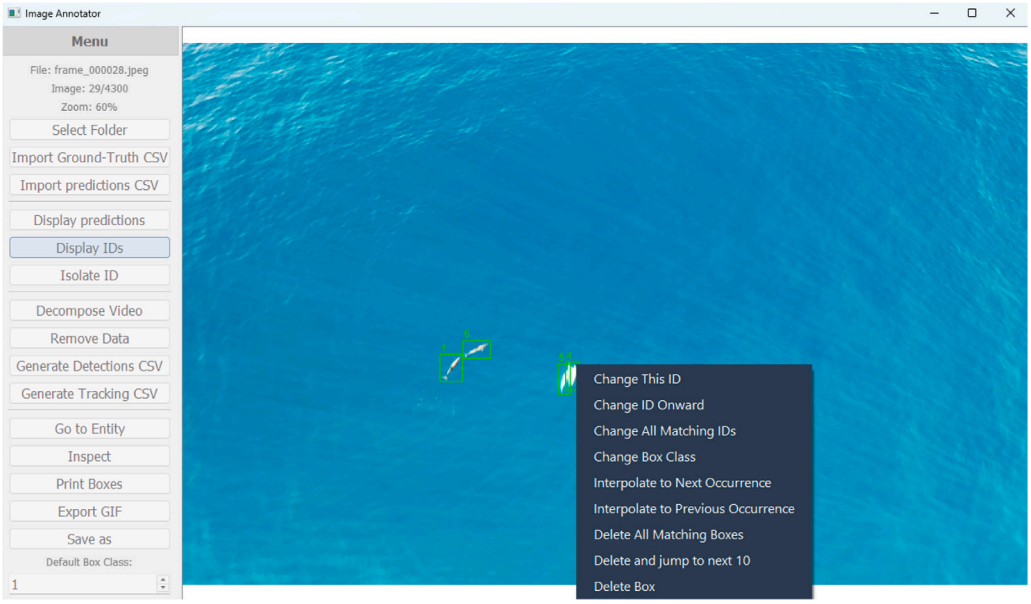


Fig. 5. Graphical interface of the annotation tool.

6. Implementation

6.1. Dolphin detector training and performance evaluation

We trained a YOLO11 large (YOLO11L) object detection model to detect dolphins in UAV imagery. A pretrained YOLO11L model was used as the starting point, exploiting transfer learning to accelerate convergence and improve performance compared to training from scratch, and chosen as a balanced compromise between detection accuracy on small objects and computational feasibility. Training was carried out using the Ultralytics YOLO implementation (Jocher et al., 2022) with default hyperparameters, except for the specific adjustments detailed below.

The training dataset followed the YOLO format and was specified in a custom YAML file containing image paths and class definitions. Images were resized to 640 × 640 pixels before being passed to the model. Training was performed for 17 epochs on a single NVIDIA GeForce RTX 3060 GPU. To ensure reproducibility, a fixed random seed (42) was used across all runs.

Data augmentation was enabled to improve model generalization. Specifically, we applied horizontal flips (fliplr = 0.1), vertical flips (flipud = 0.1), and slight adjustments to hue (hsv_h = 0.1) and brightness (hsv_v = 0.1). These augmentations are particularly relevant in marine environments, where dolphins may appear at varying orientations, under changing lighting conditions, and against heterogeneous sea backgrounds. Flipping augmentations are also important in UAV footage due to the rotational and translational freedom of the camera.

In preliminary experiments we found that standard mosaic augmentation decreased detection performance on our UAV dolphin dataset. A likely explanation is that combining frames into a single composite image reduces the effective resolution of already small dolphin targets and introduces sharp seams between sub-images with different sea states, illumination and flight altitudes. This produces unrealistic water textures and partially cropped dolphins or wakes, effectively increasing label noise. Since such artefacts do not appear in deployment imagery, the detector may overfit to mosaic-specific patterns rather than the fine-scale contrast cues that distinguish dolphins from wave crests and whitecaps, which in turn degrades performance on real UAV footage. Therefore, we kept YOLO defaults, avoiding mosaic; simple geometric/photometric augmentations sufficed.

Our simplified, ecologically realistic training preserved generalizability. We report mAP@50 and mAP@50:95, reflecting lenient versus stringent Intersection over Union (IoU) thresholds for localization and ranking (Solawetz, 2025).

To further investigate the regions that contribute most to dolphin detection, we generated importance maps using the approach proposed by Petsiuk et al. (2018). To better understand the limitations of the system, we performed a dedicated error analysis of the detector.

Table 2
Main hyperparameters used to train the YOLO11L detector.

Hyperparameter	Value
Input image size	640 × 640
Batch size	32
Number of epochs	17
Optimizer	AdamW
Initial learning rate	2×10^{-3} (cosine decay)
Weight decay	5×10^{-4}
Learning rate warm-up	3 epochs
Confidence threshold	0.5
NMS IoU threshold	0.3
Data augmentation	flips, colour jitter, brightness variation

Hyperparameters. The YOLO11L detector was trained on the detection split of our dataset with the hyperparameters summarized in Table 2. We used standard data augmentation (horizontal flips, colour jitter, random cropping) and selected the final confidence and NMS thresholds on the validation set to balance precision and recall for downstream tracking.

6.2. BoT-SORT optimization and ReID integration

6.2.1. Grid search (coarse exploration)

Before applying evolutionary optimization, we conducted an initial grid search to explore the sensitivity of BoT-SORT to key hyperparameters. This was carried out on the AIMM dataset (Section 4.4), chosen to reduce computational cost while still providing indicative results for parameter selection. The parameters explored were:

- track_high_thresh in {0.25, 0.5, 0.75},
- track_low_thresh in {0.1, 0.2},
- new_track_thresh in {0.25, 0.5},
- track_buffer in {30, 60} frames, and

- match_thresh in {0.8, 0.9}.

The other parameters were fixed to default values. All possible combinations of these values were enumerated, created and passed to BoT-SORT. Each configuration was executed on the AIMM dataset, and tracking outputs were stored for later comparison. This procedure allowed us to identify promising regions of the parameter space, which were subsequently injected as seeded individuals in the genetic algorithm optimization.

6.2.2. Genetic algorithm (fine exploration)

To refine the results of the grid search, we optimized BoT-SORT hyperparameters using a genetic algorithm (GA) implemented with the DEAP framework. The parameters targeted were the same thresholds and buffers but in a continuous and expanded interval of possible values:

- track_high_thresh in (0.2–0.8),
- track_low_thresh in (0.1–0.5),
- new_track_thresh in (0.2–0.8),
- track_buffer in (10–90 frames), and
- match_thresh in (0.1–0.8).

The GA was initialized with a mixed population of randomly sampled individuals and seeded parameter sets from the grid search. Each individual represented a candidate configuration. The evolutionary process employed blend crossover ($\alpha = 0.4$), Gaussian mutation ($\sigma = 0.1$, mutation probability = 0.2), and tournament selection (tournament size = 3). The algorithm ran for 10 generations with a population size of 100 and a Hall of Fame archive of 8 individuals. Each candidate configuration was evaluated by running BoT-SORT on the AIMM dataset (Section 4.4) using YOLO11 detections as input. Performance was assessed with the TrackEval framework (Luiten & Hoffhues, 2020) using the Higher Order Tracking Accuracy (HOTA) metric (Luiten et al., 2020), as well as IDentification F1-score (IDF1). HOTA jointly evaluates detection, association, and localization to provide a single, unified tracker comparison metric. IDF1 evaluates how long the tracker correctly identifies an object and is based on the Identification Precision and Identification Recall. The overall fitness score was defined as:

$$\text{Fitness} = 0.6 \cdot \text{HOTA} + 0.4 \cdot \text{IDF1}. \quad (1)$$

In preliminary experiments, we used only HOTA as the fitness signal for the GA, since it jointly evaluates detection and association quality. However, we observed that some parameter configurations with similar (or slightly higher) HOTA still produced clearly inferior identity preservation in practice, with more fragmented tracks and ID switches. To address this, we extended the fitness to also include IDF1, a widely used identity-centric tracking metric, and combined the two after normalization as presented in Eq. (1). This choice reflects the fact that HOTA remains our primary objective (overall tracking quality), while IDF1 explicitly regularizes the search against solutions that achieve small HOTA gains at the cost of degraded identity consistency. In our experiments, moving from a HOTA-only objective to this two-term fitness systematically reduced severe identity fragmentation with negligible loss in HOTA, and better aligned optimization with our ecological goal of accurately counting unique individuals across the sequence.

Hyperparameters. Table 3 summarizes the main BoT-SORT hyperparameters before and after GA optimization. The GA search was run on the tracking validation set using a multi-metric fitness that combines HOTA and IDF1, as described earlier in this section. Unless otherwise stated, all tracking results in the paper use the GA-optimized motion-only configuration.

Table 3
Key BoT-SORT hyperparameters.

Association and lifespan	
track_high_thresh[a]	–
track_low_thresh[a]	–
new_track_thresh[a]	–
track_buffer (frames)[a]	–
match_thresh[a]	–
fuse_score	True
gmc_method	sparseOptFlow
proximity_threshold	0.5
GA configuration	
Population size	100
Number of generations	10
Crossover/mutation probabilities	0.4/0.1
Fitness metrics	HOTA, IDF1

[a] Values were obtained from the GA search.

Model naming and performance evaluation. Throughout this paper we named the default BoT-SORT coupled with the trained YOLO11 detector as BS. The GA-optimized BoT-SORT coupled with YOLO11 is referred to as BS_GA. All tracking configurations are evaluated on the AIMM dataset and BS as baseline comparison.

6.2.3. ReID integration (appearance cues)

To evaluate whether appearance cues could improve dolphin identity tracking, we trained ReID networks using our tracking dataset converted into ReID format. These networks are capable of discriminating visual features and re-identifying individuals based on these cues. This dataset consisted of cropped dolphin detections extracted from corrected tracks; each assigned an identity label. The data were divided into training and validation subsets. Prior to training, we applied a cleaning procedure to improve data quality. This involved (i) removing identities with fewer than 10 usable images, (ii) discarding images that were too small (short side <64 pixels) or excessively blurry (low Laplacian variance), and (iii) eliminating near-duplicate images within each identity using perceptual hashing, keeping only the highest-quality instance. We implemented a threshold-sweeping step to explore different parameter values for minimum image size, blur variance, identity size, and duplicate tolerance. By precomputing image metrics and simulating multiple cleaning configurations, we obtained summary statistics of images and identities retained under each scenario, which allowed us to select thresholds that balanced data quality with dataset coverage.

We trained and evaluated several ReID variants within the BoT-SORT framework using the deep-person-reid (Torchreid) framework (Zhou & Xiang, 2019). In all cases, detections from the YOLO11L detector were cropped with a small padding margin, resized to the input resolution of the ReID backbone, and normalized using standard ImageNet statistics before being passed through the network to obtain an embedding vector for each detection.

We considered two families of backbones commonly used in person re-identification and in existing BoT-SORT implementations, adapted to our single-class marine mammal setting:

- OSNet-based models, initialized with publicly available person-ReID weights.
- ResNet50mid, following the mid-level feature variant widely used in tracking-by-detection pipelines.

For both backbones, we replaced the final classification layer by a dolphin-specific classifier and fine-tuned the networks on cropped dolphin detections extracted from the training split of our dataset. Training followed a standard protocol combining cross-entropy loss on the identity labels with a batch-hard triplet loss on the embeddings, using random horizontal flips, colour jitter, and mild geometric perturbations as data augmentation. The resulting appearance embeddings

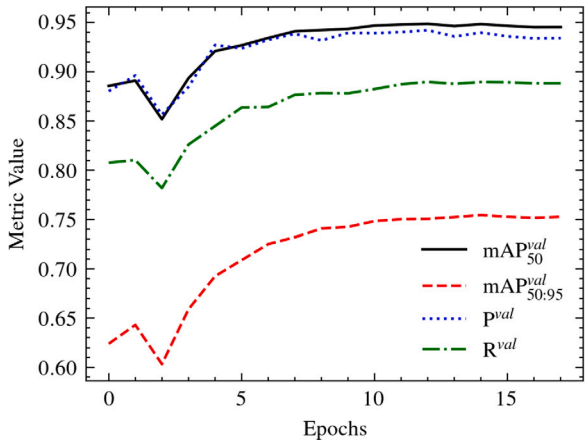


Fig. 6. YOLO11 performance on the validation set over the training epochs. Each metric value represents the best result obtained in each epoch, respectively. The notation mAP$^{val}_{t}$ represents the obtained mAP at a given IoU threshold t on the validation set.

were then integrated into BoT-SORT, where cosine similarity between embeddings was combined with motion cues during data association.

We evaluated these appearance models in two configurations: (i) a *full* BoT-SORT tracker that combines motion and appearance cues, and (ii) a *motion-only* baseline in which the appearance term is disabled and associations rely solely on Kalman filter predictions and IoU-based matching.

ReID training hyperparameters. Both OSNet and ResNet50mid ReID models were fine-tuned on cropped dolphin detections using a batch size of 32, an initial learning rate of 3×10^{-4} with single step scheduling with 20 stepsize, and 100 maximum training epochs. We optimized the networks with AdamW, using weight decay of 4×10^{-4} and a combined loss consisting of cross-entropy on identity labels and batch-hard triplet loss on the embeddings. All ReID experiments reported in this section use this configuration.

7. Results and discussion

7.1. Detection (YOLO11L)

The training dynamics reveal a consistent improvement in detection performance across epochs, as summarized in Fig. 6. Validation precision increased quickly, indicating that the model rapidly achieved a high discrimination capacity. Recall followed a steady growth, highlighting enhanced sensitivity to positive detections. The mAP@50 metric exhibited a clear upward trajectory, confirming strong convergence and stable performance. In parallel, the loss functions (Fig. 7) decreased smoothly, with the training losses (Fig. 7(a)) remaining consistently low throughout and validation losses (Fig. 7(b)) stabilizing after the initial variance observed in the classification loss. These results demonstrate that the model reached convergence within the first ten epochs, with subsequent training epochs providing marginal refinements, leading to an overall balance of high precision, recall, and generalization.

Visual inspection of a few examples provides insight into how the model is functioning internally. Fig. 8 shows two selected cases that illustrate the generalization capabilities of the detection model. Fig. 8(a) demonstrates the ability of the model to identify small individuals with confidence scores above 0.7. The rightmost detection shows the model correctly detecting the individual, regardless of the glare that affects the area. Fig. 8(b) confirms this robustness, showing detection of dolphins under glare, while also illustrating a good detection where the individual is partially within the frame, further suggesting that the model is resilient to incomplete visual cues. In addition, detections with lower confidence values (around 0.5–0.6) observed in some examples indicate that the model is still capable of providing useful localization even under suboptimal visibility, though at a reduced certainty. Overall, these examples suggest that the model generalizes well across variations in illumination, scale, and framing, all of which are common in drone-based marine surveys.

The resulting importance map visualizations consistently highlighted the dorsal fin, rostrum, and caudal peduncle as the most influential features for the detector's predictions, regardless of body orientation (Fig. 9). In cases where the animals were partially occluded by water splashes, the model shifted its focus to these more distinct anatomical features, indicating a robustness to noise.

Fig. 10 illustrates the complementarity between the importance analysis and the detection output. The importance map (Fig. 10(a)) highlights that the model focuses predominantly on distinctive features of the dolphin body, particularly the dorsal fin and rostrum, which seem to be the most informative cues for identification. At the same time, parts of the background within the detection box are also activated, albeit with much lower weights. This suggests that some residual importance is given to the surrounding water, possibly due to reflections and local contrast patterns. The corresponding detection output (Fig. 10(b)) confirms that this importance is sufficient to support accurate localization, with high confidence score (above 0.9). Taken together, these results indicate that the model reliably detects dolphins while grounding its predictions in meaningful anatomical features, although some sensitivity to background artefacts remains.

Table 4
Tracking performance utilizing two distinct BoT-SORT parametrization: default and the resultant from GA optimization. The column IDs represents the total count of dolphins individuals within the video. The annotators agreed on a total of 62 dolphins. The detector was the trained YOLO11L (Section 6.1). BS stands for BoT-SORT.

Model	HOTA	MOTA	IDF1	IDs
BS (default)	0.631	0.720	0.734	101
BS_GA (ours)	**0.636**	**0.723**	**0.740**	**72**

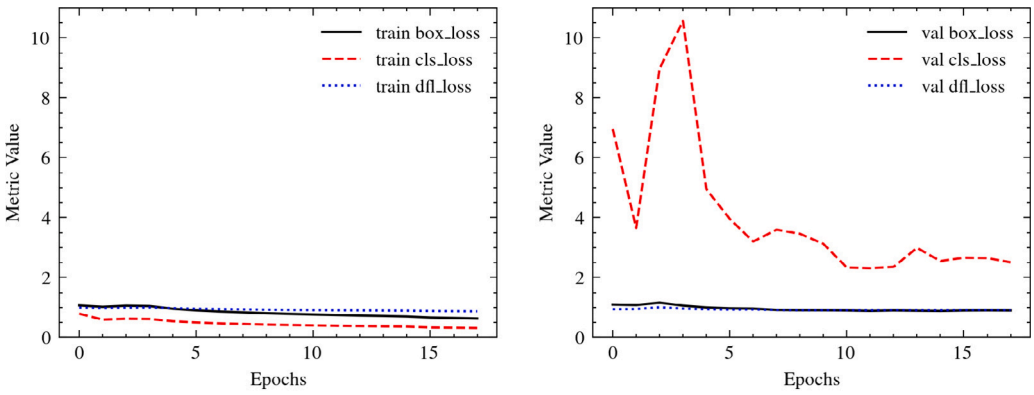
7.2. Tracking (GA optimized BoT-SORT without ReID)

Table 4 reports the performance of the baseline BS and the proposed BS_GA on the AIMM dataset. The proposed method shows consistent improvements across all metrics. HOTA and Multi-Object Tracking Accuracy (MOTA)[1] increased slightly. IDF1 also improved, reflecting better identity preservation. Regarding the total number of distinct IDs, the baseline produced 101 identities compared to the ground-truth 62, while the proposed method reduced this to 72, bringing it much closer to the reference. This reduction in over-fragmentation of tracks is the most significant improvement, indicating enhanced stability and continuity in identity assignment.

The evolutionary process showed a rapid improvement in the early generations, with the average fitness rising from 53.7 in generation 1 to nearly 60 by generation 5, as shown in Fig. 11. After this stage, the population converged quickly, with only minor improvements up to generation 10. The maximum fitness remained stable around 60 in the first six generations, only slightly increasing to 60.5 thereafter. This premature convergence can be attributed to the injection of the best individuals found in the gridsearch into the population, which accelerated the rise in average fitness.

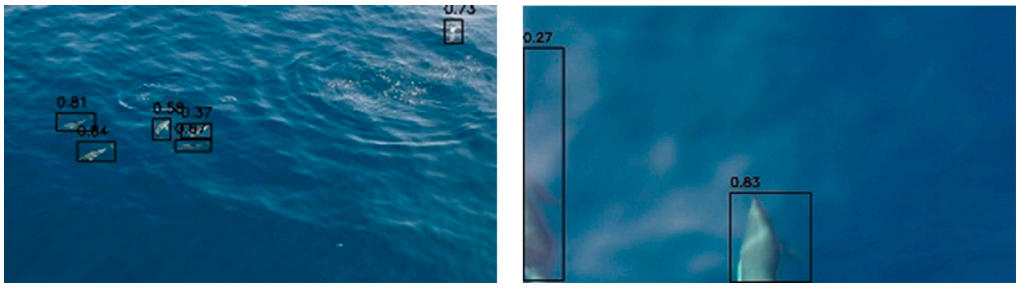
Although the absolute gains in standard tracking metrics appear modest (e.g. small increases in HOTA and IDF1), the GA search consistently reduced identity fragmentation and ID switches compared with

[1] MOTA is a summary metric that penalizes false positives, missed targets, and identity switches, aggregating these errors into a single higher-is-better score for multi-object tracking.



(a) Training losses over epochs. (b) Validation losses over epochs

Fig. 7. Evolution of YOLO11 training and validation losses over the epochs. YOLO11 minimizes three distinct loss functions: box_loss is a regression loss for bounding box position; cls_loss stands for object classification loss; and dfl_loss stands for distribution focal loss and is designed to handle class imbalance and enhance performance on challenging detection scenarios.



(a) Detection under glare. (b) Detection with partial dolphin cues.

Fig. 8. Examples of YOLO11 predictions illustrating good performance (a) under glare and (b) with partial dolphin cues.

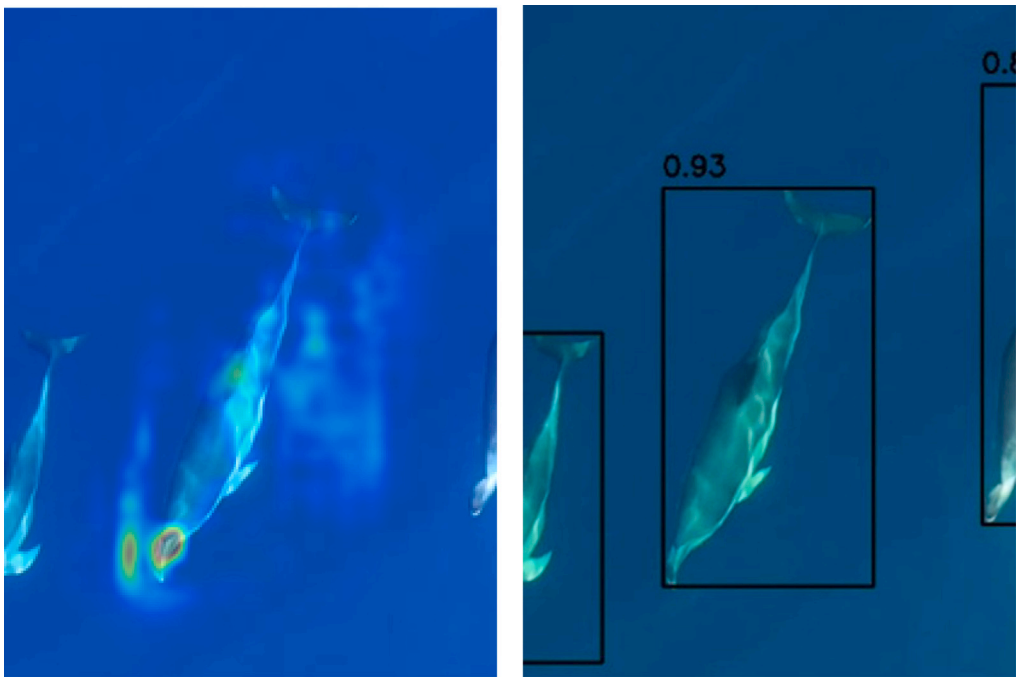
(a) Models sees dolphin dorsum. (b) Model sees caudal fin and rostrum.

Fig. 9. Two YOLO11 importance maps, generated using an occlusion-based approach, showing how the detector tends to prioritize biologically relevant features. Each map shows the importance only on the most confident detection.

the default BoT-SORT configuration, with around a 29% reduction in fragmented tracks in our experiments. This difference is important in our application, where ecological indicators (e.g. abundance estimates, re-sightings) depend on maintaining consistent identities over long sequences rather than on per-frame accuracy alone. From a practical standpoint, the GA optimization is a one-off, offline cost: once a good parameter set is learned for a given sensor and survey protocol, it can be reused across many campaigns with no further tuning. In contrast, manual hyperparameter search is labour-intensive and difficult to reproduce. We therefore consider the additional complexity of GA tuning justified in operational settings where identity preservation and robust, long-term counting are critical, while acknowledging that for quick exploratory studies or applications tolerant to moderate identity noise, the default tracker configuration may be sufficient.

7.3. Tracking (BoT-SORT with ReID)

Quantitative results indicated that appearance-assisted BoT-SORT did not improve tracking compared to the default configuration on the AIMM dataset. Metrics remained unchanged or slightly worsened,



(a) Importance map. (b) Bounding box.

Fig. 10. Importance map and bounding box pair, illustrating a case where the detector attributes relevance to the water surrounding the dolphin.

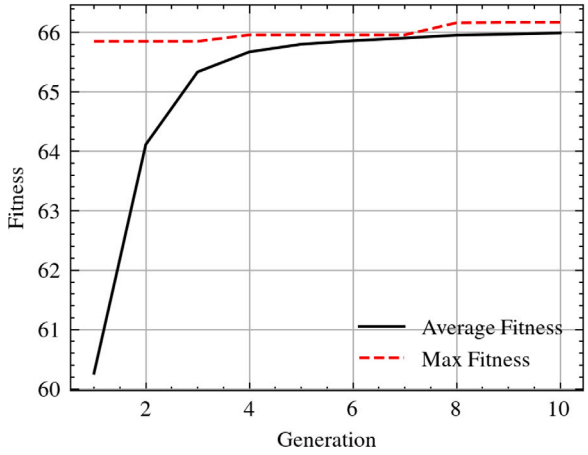


Fig. 11. Fitness evolution over the GA generations.

with HOTA and IDF1 dropping slightly. More critically, the number of predicted identities increased from 101 to 105–106, moving further away from the ground-truth value of 62. These results confirm that incorporating appearance cues introduced noise rather than improving discrimination, leading to more fragmented tracks and reduced identity consistency.

This outcome illustrates an important insight for wildlife monitoring: we show that appearance-based ReID does not seem suitable for dolphins in overhead UAV imagery, where dorsal views provide weak and unstable visual features. Unlike pedestrians or vehicles, dolphins lack strong inter-individual distinctiveness when viewed from above, and specular reflections further obscure appearance cues. Motion-based association thus remains more reliable in this ecological context, and our results caution against uncritical transfer of person-ReID methods to marine mammal tracking.

The ablation study revealed that metrics were not affected by appearance weight variations within 0.1 and 0.5.

Our negative result with ReID is consistent with the visual characteristics of dorsal dolphin imagery from UAVs. First, individuals are typically small in the frame and seen from above as uniformly grey, elongated shapes with limited visible pigmentation or scarring; inter-individual differences are therefore extremely subtle at the available resolution. In contrast, intra-individual appearance varies strongly with pose (roll, pitch and yaw), partial occlusions by waves and whitecaps, and specular highlights from sun glint. Second, the training crops inevitably include a large amount of water and wake, so the network tends to encode short-lived context (local texture of the sea surface) rather than stable animal-specific cues. Finally, group movement patterns mean that different dolphins within the same school often have very similar dorsal silhouettes and wake signatures at a given instant. Together, this produces a regime where inter-individual variability is low, intra-individual variability is high, and much of the signal comes from unstable background features, which is precisely the opposite of what appearance-based ReID methods assume. Our experiments with several backbones (including OSNet and ResNet50-based variants) all converged to the same conclusion: appearance embeddings add little discriminative power beyond motion, whereas a carefully tuned motion-only tracker already captures the most reliable cues available in these dorsal UAV views. Similar observations were reported by Zhang et al. (2023), who noted that animal features are often more complex and less distinguishable than pedestrian appearances, making visual tracking in wildlife contexts more challenging.

7.4. End-to-end group size estimation

On the AIMM dataset, the end-to-end estimator exhibits a positive counting bias of +10 individuals relative to the ground truth (74 vs. 64; +15.6%). Computed frame-wise, the overall Mean Average Error (MAE) and Root Mean Squared Error (RMSE) are 1.242 and 2.282, respectively (Table 5). Performance varies across sequences: errors are small for Seqs. 2, 3, and 5 (MAE $\leq$ 0.612; RMSE $\leq$ 1.015), with Seq. 3 matching the total count. In contrast, Seqs. 0, 1, and 4 are more challenging (MAE up to 4.634; RMSE up to 5.342), and the scatter of predicted versus true counts shows overestimation in most sequences, with Seq. 1 as the

Table 5
Group size estimation performance on the AIMM dataset.

Seq.	MAE	RMSE	IDs_{pred}	IDs_{GT}
0	2.389	3.119	16	14
1	1.280	2.163	8	9
2	0.477	0.804	17	12
3	0.612	1.015	16	16
4	4.634	5.342	14	11
5	0.010	0.117	3	2
	1.242	**2.282**	**74**	**64**

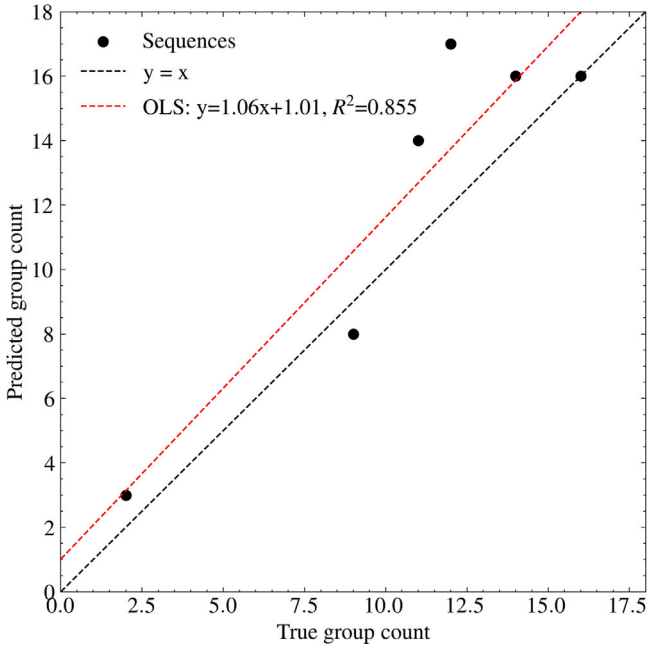


Fig. 12. Predicted versus true count for each sequence on the AIMM dataset. The dashed red-line represents an ordinary least squares regression of the datapoints. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

main underestimation outlier (Fig. 12). A least-squares fit (slope = 1.06, intercept = 1.01, R^2 = 0.855) lies above the identity line, indicating a consistent positive bias, especially at larger group sizes.

From an ecological perspective, our counting error ($MAE \approx 1.24$ dolphins per frame on the validation set) should be interpreted relative to the group sizes and survey objectives. In our data, groups range from single animals to aggregations of tens of individuals, so an average deviation of about one dolphin per frame corresponds to a small fraction of group size in most cases. Importantly, inspection of error patterns suggests that discrepancies arise mainly from missed individuals in crowded or highly cluttered conditions, rather than from systematic double-counting. This means that our pipeline behaves similarly to an additional "observer": it provides near-unbiased but slightly noisy counts at the surface, subject to the same kinds of perception limitations (brief surfacings, partial occlusions by waves) that constrain human observers. As with any visual survey, these counts do not directly correct for availability bias (animals below the surface) or imperfect detection at long range, so they should be viewed as a repeatable, standardized index of surface abundance rather than an absolute abundance estimator on their own. In practice, such indices can be integrated into existing distance-sampling or mark–recapture frameworks, where the key advantage of automation is the ability to process large volumes of UAV footage consistently and reproducibly across surveys, thereby strengthening temporal trend detection even when per-frame errors remain at the level of one or a few individuals.

7.5. Error analysis

The distribution of detection confidence scores provides further insight into the separation between true and false detections. True Positives (TP) (Fig. 13(a)) showed a clear skew towards high confidence, with most detections concentrated above 0.75, indicating that the model is generally reliable when assigning high scores. In contrast, false positives (FP) (Fig. 13(b)) exhibited a broader and flatter distribution, with a large proportion clustered at lower confidence values (<0.4), though a non-negligible number extended into higher ranges (>0.7). This overlap suggests that while score thresholds can filter out many false detections, they cannot eliminate them entirely, as some false positives receive confidence levels comparable to true detections. These results show the importance of combining confidence-based filtering with additional strategies (e.g., temporal consistency checks or background modelling) to further reduce false alarms.

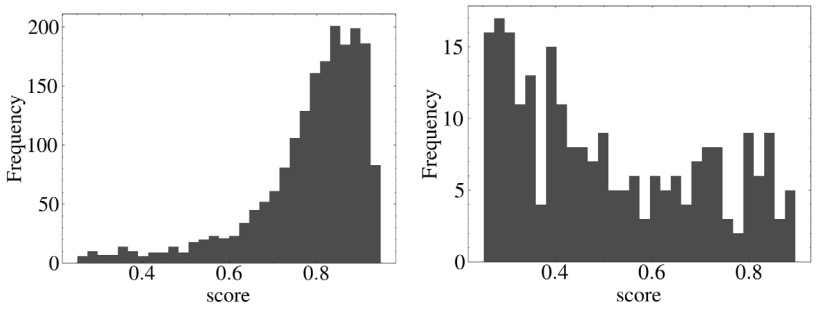
To better understand the limitations of our pipeline, we then extracted all frames in which the predicted count differed from the ground truth or where HOTA/IDF1 indicated association errors, and then manually inspected a stratified subset of these cases. Each failure was assigned to one or more of the following categories:

1. Low-contrast or distant individuals: dolphins close to the resolution limit or with very weak contrast against dark, textured water, leading to missed detections (Fig. 14(a)).
2. Sea-surface clutter and sun glint: confusion between dolphins and whitecaps, wave crests or strong specular highlights, typically causing false positives or uncertain localizations (Fig. 14(b)).
3. Dense aggregations and occlusions: tightly packed groups where individuals overlap in dorsal view or are partly hidden by waves, resulting in under-counting and fragmented tracks (Fig. 14(c)).
4. Border effects: animals entering or leaving the frame at the image borders, which often break tracks and generate short-lived identities.
5. Non-dolphin objects of similar size: surfers or small watercraft that occasionally trigger spurious detections despite being included as negatives during training (Fig. 14(d)).

Within this taxonomy, the most frequent failure modes are (1) and (3), which are also challenging for human observers. Tracking-specific errors are dominated by (3) and (4), where prolonged occlusions or frame exits interrupt otherwise stable motion trajectories. This categorization provides concrete targets for future improvements, such as adaptive thresholds in low-contrast regions, clutter-aware training augmentations, or explicit handling of entry/exit events at frame boundaries.

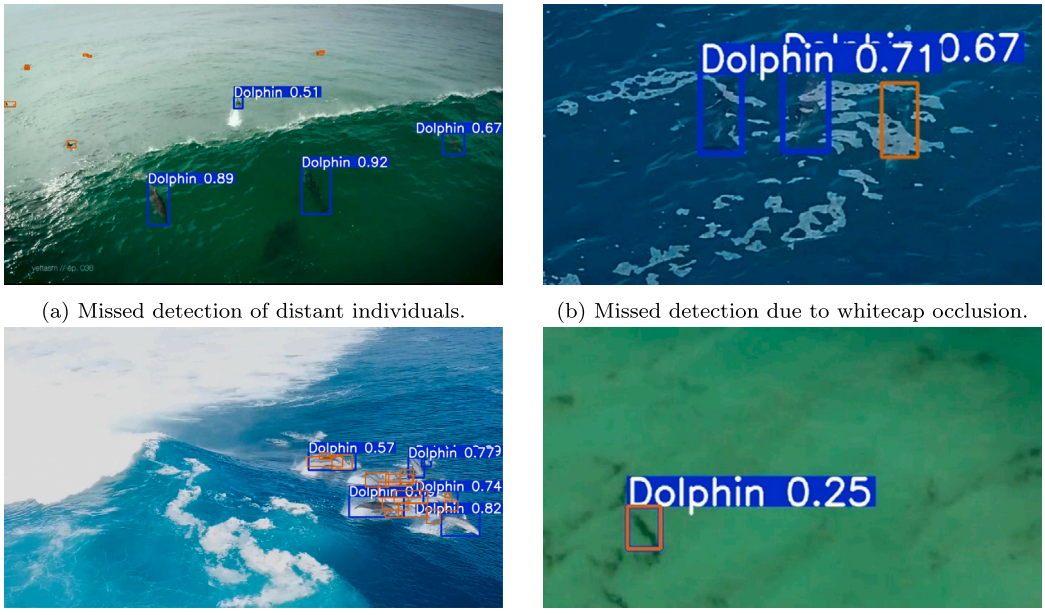
8. Limitations and future work

Our study has several limitations that suggest concrete avenues for future work. A first limitation is that the model is currently trained and tuned on a single focal taxon. As a result, there is a clear risk of domain shift: detection and tracking performance may degrade when applying the pipeline to other cetacean species with distinct body shapes and surfacing behaviour. A rigorous cross-domain evaluation would require independent annotated UAV datasets from other areas and species, which are not yet available to us. We therefore explicitly flag cross-species generalization as an open question and a priority for future work as broader multi-species UAV datasets become available. A second limitation is that our tracking pipeline is strictly single-camera and motion-based: appearance features are not exploited, and identity preservation can degrade in highly cluttered conditions, at long ranges, or when groups split and merge. Finally, although the GA-based tuning improves BoT-SORT performance, it is run offline; the resulting configuration is fixed and may not be optimal when the operational domain (altitude, sea state, water colour) shifts.



(a) Frequency of True Positives (TP). (b) Frequency of False Positives (FP).

Fig. 13. Histograms showing the frequency of true positives (a) and false positives (b) for each confidence score.



(a) Missed detection of distant individuals.

(b) Missed detection due to whitecap occlusion.

(c) Missed detection due to dense group aggregation.

(d) False positive due to algae with dolphin-like shape.

Fig. 14. Examples of detection failures. The blue boxes indicate correct model predictions and orange boxes indicate misdetections, i.e., false positives and false negatives. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Future work will therefore focus on both expanding the data and integrating richer learning paradigms. On the data side, we plan to incorporate additional regions, species and sensing conditions (including higher sea states and near-shore surf zones), and to add explicit labels for other marine animals and anthropogenic objects to strengthen negative supervision. On the modelling side, one promising direction is to explore integrated multi-kernel/multi-cue learning strategies inspired by the work of Zhang et al. (2021) on device-free localization, where multiple kernels are used to fuse complementary spatiotemporal features in cluttered environments. In our setting, analogous ideas could be used to jointly exploit appearance cues, motion patterns and sea-state descriptors to obtain more robust dolphin representations across viewing conditions. A second direction is online and domain-adaptive learning: instead of training a static model, the detector and/or tracker could be updated incrementally as new UAV surveys are collected, in the spirit of online domain adaptation frameworks developed for Wi-Fi-based localization in non-stationary environments (Zhang et al., 2025). Such approaches could help the pipeline remain accurate as survey areas, water properties or acquisition protocols evolve over time.

Another limitation of this study is that we do not benchmark against a wide range of recent detection and tracking architectures. Training, tuning, and fairly comparing multiple state-of-the-art pipelines on our large-scale dolphin detection (64,705 images, 225,305 boxes) and tracking (54,274 frames, 603 tracks) datasets would require computational resources and time beyond the scope of the present work. Instead, we focus on a recent and competitive baseline (YOLO11L + BoT-SORT) and provide detailed ablations and GA-based hyperparameter tuning. A broader multi-architecture benchmark on the datasets is therefore left as an important direction for future work.

Regarding strategies to increase ReID performance, one could pursue a fin-centric pipeline, cropping the dorsal fin and adding a contour stream. Also, training a view-robust descriptor exploring techniques like margin-based softmax (Gu et al., 2022) plus batch-hard triplet losses (Hermans et al., 2017) and tracklet-aware sampling with large language models (Wu et al., 2025) could deem effective. Applying domain-specific augmentations for glare/scale, and aggregating embeddings over short clips are simple techniques that could improve these results.

Beyond these directions, semi-supervised and active learning could further reduce annotation costs by exploiting the large volume of unlabelled UAV footage typically available, while multi-task formulations that combine detection, tracking and behavioural labelling may improve both data efficiency and ecological utility. We leave the systematic exploration of these learning paradigms to future work.

9. Conclusions

This work presented an end-to-end framework for dolphin monitoring from UAV imagery, combining semi-supervised dataset development, a custom annotation tool, YOLO11 detection, and BoT-SORT tracking with genetic algorithm–based optimization. The approach achieved reliable detection (precision of approximately 0.93 across a range of sea states) and improved tracking continuity (reduced ID fragmentation by about 29% relative to default settings), while also surpassing technical benchmarks to demonstrate group-size estimation as an ecological application.

Detection importance maps suggest that the detector prioritizes biologically relevant features, which may explain its ability to generalize across varying viewing angles and environmental conditions.

By integrating UAVs, detection-based tracking, and ecological objectives, this study provides a scalable, non-invasive basis for advancing automated marine mammal monitoring.

Beyond the specific architectural and optimization choices made here, our results highlight that reliable dolphin monitoring from UAVs is now within reach for routine ecological surveys, provided that models are carefully tuned to the visual statistics of the marine environment. At the same time, the limitations discussed above, notably the restricted geographic and taxonomic coverage, the absence of explicit multi-species annotations, and the reliance on a fixed, offline-tuned tracker, indicate that our framework should be seen as a strong baseline rather than a final solution.

In future work, we plan to build on this baseline by incorporating richer learning paradigms. Integrated multi-kernel or multi-cue formulations could allow the detector and tracker to combine complementary spatiotemporal descriptors of dolphins and sea state within a unified model, while online and domain-adaptive learning would enable the system to update as new surveys are collected in previously unseen locations. In parallel, semi-supervised and active learning strategies could exploit the large volumes of unlabelled UAV footage typically available, reducing annotation effort and improving robustness to rare conditions. Taken together, these extensions will move the framework from a static, single-species pipeline towards a more flexible and general tool for automated marine mammal monitoring in diverse ocean environments.

CRediT authorship contribution statement

Leonardo Viegas Filipe: Writing – review & editing , Writing – original draft, Visualisation, Conceptualisation, Data curation. **João Canelas:** Writing – review, Conceptualisation, Data curation. **Mário Vieira:** Writing – review, Conceptualisation. **Francisco Correia da Fonseca:** Project administration, Supervision, Writing – review. **André Cid:** Writing – review, Data curation. **Joana Castro:** Writing – review, Data curation. **Inês Machado:** Supervision, Project administration, Conceptualisation, Writing – review.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the first author used ChatGPT to improve language and support literature search. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Declaration of competing interest

The authors declare that they have no conflict of interest.

Acknowledgements

The authors acknowledge Fundação para a Ciência e a Tecnologia (FCT), Portugal for its financial support via the projects LAETA Base Funding (DOI: 10.54499/UIDB/50022/2020) and LAETA Programatic Funding, Portugal (DOI: 10.54499/UIDP/50022/2020). The authors also acknowledge the EU-SCORES project, that received funding from the European Union's Horizon 2020 research innovation programme under the Grant Agreement number 101036457.

Data availability

Data will be made available on request.

References

Adžemović, E., Ayub, A., Malik, A., Chen, J., & Hussain, A. (2025). Deep learning-based multi-object tracking: a comprehensive survey. Preprint (arXiv:2506.13457). https://arxiv.org/abs/2506.13457. (Accessed 7 October 2025).

Aharon, N., Orfaig, R., & Bobrovsky, B.-Z. (2022). BoT-SORT: robust associations multi-pedestrian tracking. http://dx.doi.org/10.48550/arXiv.2206.14651, Preprint (arXiv:2206.14651).

Alsaidi, M., Al-Jassani, M., Bang, C., O'Corry-Crowe, G., Watt, C., Ghazal, M., & Zhuang, H. (2024). Localization and tracking of beluga whales in aerial video using deep learning. *Frontiers in Marine Science*, *11*, Article 1445698. http://dx.doi.org/10.3389/fmars.2024.1445698.

Álvarez-González, M., Suárez-Bregua, P., Pierce, G., & Saavedra, C. (2023). Unmanned aerial vehicles (UAVs) in marine mammal research: a review of current applications and challenges. *Drones*, *7*(667), http://dx.doi.org/10.3390/drones7110667.

Canelas, J., Clementino, L., Cid, A., Castro, J., Machado, I., & Vieira, S. (2025). Automated cetacean detection in UAV imagery using AI models: a case study on delphinid species. *International Journal of Data Science Analytics*, *20*, 3695–3979. http://dx.doi.org/10.1007/s41060-024-00704-9.

CVAT. ai Corporation (2024). Computer vision annotation tool (CVAT) (v2.16.1) [computer software]. http://dx.doi.org/10.5281/zenodo.12771595, Zenodo.

Dutta, A., & Zisserman, A. (2019). The VIA annotation software for images, audio and video. In *Proceedings 27th ACM international conference on multimedia* (pp. 2276–2279). http://dx.doi.org/10.1145/3343031.3350535.

Gu, H., Li, J., Fu, G., Yue, M., & Zhu, J. (2022). Loss function search for person re-identification. *Pattern Recognition*, *124*, Article 108432. http://dx.doi.org/10.1016/j.patcog.2021.108432.

Guan, Y., Fan, M., Bai, T., Li, X., & Xu, Z. (2025). Multi-object tracking review: retrospective and emerging trend. *Artificial Intelligence Review*, *58*, Article 235. http://dx.doi.org/10.1007/s10462-025-11212-y.

Hermans, A., Beyer, L., & Leibe, B. (2017). In defense of the triplet loss for person re-identification. http://dx.doi.org/10.48550/arXiv.1703.07737, Preprint (arXiv:1703.07737).

Jegham, N., Koh, C., Abdelatti, M., & Hendawi, A. (2025). YOLO evolution: a comprehensive benchmark and architectural review of YOLOv12, YOLO11, and previous versions. http://dx.doi.org/10.48550/arXiv.2411.00201, Preprint (arXiv:2411.00201).

Jocher, G., Chaurasia, A., & Qiu, J. (2022). YOLOv5 (version 7.0) [computer software]. http://dx.doi.org/10.5281/zenodo.7347926, Zenodo.

Kellenberger, B., Marcos, D., & Tuia, D. (2018). Detecting mammals in UAV images: best practices to address a substantially imbalanced dataset with deep learning. *Remote Sensing of Environment*, *251*, Article 112105. http://dx.doi.org/10.1016/j.rse.2018.06.028.

Khanam, R., & Hussain, M. (2024). YOLOv11: an overview of the key architectural enhancements. http://dx.doi.org/10.48550/arXiv.2410.17725, Preprint (arXiv:2410.17725).

Li, S., Cao, J., & Xie, H. (2024). A review of detection-related multiple object tracking in recent times. In *Proceedings 2024 international conference on advanced communications technology* (pp. 1–6). IEEE.

Luiten, J., & Hoffhues, A. (2020). *TrackEval [computer software]*. GitHub, https://github.com/JonathonLuiten/TrackEval. (Accessed 7 October 2025).

Luiten, J., Osep, A., Dendorfer, P., Torr, P., Geiger, A., Leal-Taixé, L., & Leibe, B. (2020). HOTA: a higher order metric for evaluating multi-object tracking. *International Journal of Computer Vision*, *128*, 2025–2044. http://dx.doi.org/10.1007/s11263-020-01375-2.

Petsiuk, V., Das, A., & Saenko, K. (2018). RISE: randomized input sampling for explanation of black-box models. http://dx.doi.org/10.48550/arXiv.1806.07421, Preprint (arXiv:1806.07421).

Programmer, D. (2025). *Darkpgmr/DarkLabel [computer software]*. GitHub, https://github.com/darkpgmr/DarkLabel. (Accessed 14 October 2025). Original work published 2020.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: unified, real-time object detection. http://dx.doi.org/10.48550/arXiv.1506.02640, Preprint (arXiv:1506.02640).

Solawetz, J. (2025). What is mean average precision (mAP) in object detection? Roboflow Blog. https://blog.roboflow.com/mean-average-precision/. (Accessed 24 October 2025).

Ultralytics (2023). *Ultralytics YOLO11 [computer software]*. GitHub, https://github.com/ultralytics/ultralytics. (Accessed 7 October 2025).

Wu, Z., Wang, L., Wang, C., Teixeira, C., Ke, W., & Xiong, Z. (2025). Multi-tracklet tracking for generic targets with adaptive detection clustering. http://dx.doi.org/10.48550/arXiv.2508.05172, Preprint (arXiv:2508.05172).

Zhang, L., Gao, J., Xiao, Z., & Fan, H. (2023). AnimalTrack: a benchmark for multi-animal tracking in the wild. *International Journal of Computer Vision*, *131*, 496–513. http://dx.doi.org/10.1007/s11263-022-01711-8.

Zhang, J., Li, Y., & Xiao, W. (2021). Integrated multiple kernel learning for device-free localization in cluttered environments using spatiotemporal information. *IEEE Internet of Things Journal*, *8*(6), 4749–4761. http://dx.doi.org/10.1109/JIOT.2020.3028574.

Zhang, J., Xue, J., Li, Y., & Cotton, S. (2025). Leveraging online learning for domain-adaptation in wi-fi-based device-free localization. *IEEE Transactions on Mobile Computing*, *24*(8), 7773–7787. http://dx.doi.org/10.1109/TMC.2025.3552538.

Zhao, T., Zeng, Y., Fang, Q., Xu, X., & Xie, H. (2025). Semi-supervised object detection for remote sensing images using consistent dense pseudo-labels. *Remote Sensing*, *17*(1474), http://dx.doi.org/10.3390/rs17081474.

Zhou, K., & Xiang, T. (2019). Torchreid: a library for deep learning person re-identification in PyTorch. Preprint (arXiv:1910.10093). https://arxiv.org/abs/1910.10093.