ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING - DISI

SECOND CYCLE DEGREE IN ARTIFICIAL INTELLIGENCE

DETECTION AND METRIC LEARNING FROM PSEUDO-LABELS FOR UNSUPERVISED MARITIME VESSEL IDENTIFICATION

Dissertation in Machine Learning for Computer Vision

Supervisor:
Prof. Samuele Salti

Defended by:
Marco Sangiorgi

Graduation Session V of March 2026
Academic Year 2024/2025

A mia nonna, Rosa.
A mio zio, Mario.

Abstract

This thesis introduces an unsupervised framework that exploits model agreement, temporal coherence, and geometric priors from Automatic Identification System (AIS) geolocation messages to produce pseudo-labels of maritime vessels from raw surveillance streams suitable for vessel detection and vessel re-identification. Data sources include daylight and thermal camera frames, AIS data, and camera calibration information (field of view, azimuth, elevation, and intrinsic/extrinsic parameters). The first contribution is a pseudo-label generation pipeline for vessel detection, comprising a majority-voting ensemble of strong CNN and ViT [4] detectors, a SAM3-based [27] temporally consistent tracker, and ELoFTR-based keypoint matching for cross-modal detection transfer with homographies. A lightweight You Only Look Once (YOLO) [1] detector is trained on the generated annotations. The second contribution is a pseudo-label generation pipeline for vessel re-identification, cross-referencing AIS geodetic positions with YOLO detections via perspective projection. Finally, a metric-learning recipe is presented and evaluated with both classical ResNets and the latest DINOv3 [7] backbones. Experimental validation confirms the effectiveness of the proposed framework across both detection and re-identification tasks. Experimental validation confirms the effectiveness of the proposed framework across both detection and re-identification tasks, with the fine-tuned YOLO26l detector reaching 0.951 mAP@50 and 0.916 F1 score, while the best Re-ID setting based on DINOv3 ViT achieves 0.996 centroid mAP.

Contents

Chapter 1

Introduction

Maritime surveillance is a critical component of port security, coastal monitoring, traffic management, and search-and-rescue support. Training robust computer vision systems for these environments demand labeled datasets that capture the variability of real operational conditions. Most existing maritime vessel datasets are curated offline, frequently from port photographers' imagery or limited acquisition campaigns. These datasets are valuable for benchmarking, but they are not always representative of cameras deployed in the wild, where data quality is affected by weather, compression artifacts, night-time conditions, autofocus failures, and long periods with no vessels in view. Manual annotation at this scale is costly and hard to maintain over time.

In this thesis, an automatic dataset creation system is tackled from the ground up, starting with raw maritime surveillance streams. The result is a prototype pseudo-labeling system that can accommodate both vessel detection and re-identification. As a proof of concept, training and evaluation recipes of deep learning architectures are explored for both tasks.

Disclaimer. This work is conducted exclusively as a research activity. All potentially sensitive information, including geolocation data, vessel identities, and video streams, has been either adequately anonymized/corrupted or used only when properly authorized by the rightful data owners, in compliance with privacy and safety requirements.

1.1 Related Work

Research on maritime vessel detection and re-identification (Re-ID) has advanced significantly over the past decade, leveraging optical daylight (DLTV), infrared (IR), and Synthetic Aper-

ture Radar (SAR) imagery. Existing approaches can be broadly categorized based on model architecture, learning techniques, and dataset design.

Vessel Detection. Early detection approaches relied primarily on convolutional neural networks (CNNs) and segmentation models. In [17] the authors employed U-Net architectures with various CNN backbones (VGG, ResNet, DenseNet, EfficientNet) for vessel segmentation in Sentinel-2 optical imagery, demonstrating strong performance in class-imbalanced maritime scenes. Similarly, Spagnolo et al. [14] proposed CNN-based detection as part of a baseline dataset for coastal boat monitoring, integrating bounding-box regression with Re-ID embeddings.

Lightweight and multimodal detection strategies have been proposed to tackle real-time operational constraints and complex environments. Galdelli et al. [16] combined SAR and optical imagery using a YOLO-based multimodal pipeline with AIS-assisted cross-referencing to detect dark vessels. For visible-spectrum imagery, EL-YOLO [18] introduced an efficient neck and AWIoU loss to improve detection of small and overlapping ships, while PEW-YOLOv8 [19] specialized in infrared images, incorporating multi-path feature fusion and attention mechanisms to suppress background noise.

Vessel Re-Identification. Vessel Re-ID leverages embeddings that measure similarity between detected instances. Siamese networks have been applied to optical satellite imagery for linking detections across time and sensors [11], providing early benchmarks in identity-based vessel tracking. Thermal and IR modalities extend Re-ID to low-visibility scenarios; [20] introduced viewpoint-independent thermal embeddings to enable identification at night or under poor lighting conditions. In [15] the author explored RGB–IR multi-modal Re-ID, a promising path combining thermographic and visible-spectrum cues.

Transformer-based models have recently more effectively than CNN-only approaches for vessel Re-ID, particularly in addressing intra-class variability and occlusions. Liu et al. [12] introduced MCFormer, a multi-scale correlation-aware Transformer combining global and local correlation modules to align vessel features across views. Similarly, Lu et al. [13] presented SwinTransReID, which fuses Swin Transformer embeddings with side information such as vessel type, hull color, and orientation to improve identification robustness.

Related Datasets. High-resolution optical datasets, such as the WorldView VHR imagery and the xView dataset used by Matasci et al. [11], provide dense annotations ($\sim$12,770 ship instances overall) for detection and Siamese-based Re-ID. Large-scale vessel Re-ID datasets, e.g., VesselReID-1656 [13], contain over 135,000 images across 1,656 vessels, richly anno-

tated with vessel type, orientation, and color. Coastal datasets, such as the Porto Cesareo boat dataset [14], support Re-ID from low-altitude RGB cameras. Multi-modal datasets include paired RGB–IR imagery [15] and SAR-optical collections (HS3-S2) [16], enabling multimodal detection and tracking in complex maritime conditions. Thermal datasets [20] further extend capabilities to nighttime and low-visibility surveillance. Li *et al.* [23] proposed the VesselReID dataset with ∼30,587 images across 1,248 vessels, while Zhang *et al.* [24] targeted warship Re-ID under limited data by transferring knowledge from larger vessel datasets.

1.2 System Overview

This thesis proposes an end-to-end system that starts from raw, synchronized maritime surveillance streams (DLTV/IR video, AIS messages, and camera calibration) and automatically produces training-ready outputs for both vessel detection and vessel re-identification. The pipeline combines multi-model agreement, temporal consistency, and geometric constraints to reduce label noise and preserve physically plausible associations across modalities and time.

Compared with prior work, the main gap addressed here is integration: many studies focus on isolated components, while this work links data acquisition, pseudo-label generation, detector training, AIS-to-image association, and metric-learning evaluation in one coherent workflow. In addition, the approach targets realistic coastal operating conditions (domain shift, day/night variation, sparse labels, and long-tail vessel appearances).

Thesis Structure. The thesis is organized as follows. Chapter 2 presents the vessel-detection pipeline, from pseudo-label generation to detector training and deployment-oriented optimization. Chapter 3 introduces the vessel re-identification pipeline, including AIS-based association, pseudo-label construction, and metric-learning training/evaluation. Chapter 4 summarizes the main findings and discusses future work. The Appendix reports complementary derivations, additional qualitative/quantitative results, and implementation details.

Chapter 2

Vessel Detection

2.1 Automatic Pseudo-Labels Generation

This section covers the choices taken to build Detection Dataset 4K, a dataset that comprises 3997 pseudo-labels for vessel detection extracted from raw surveillance streams. The dataset is created by first generating candidate detections with the multi-model ensemble, then enforcing temporal consistency through tracking and finally retaining high-confidence vessel boxes with cross-modal detections transfer using key-point matching. This process reduces noisy frame-level labels and produces a training set aligned with real operational conditions (day/night, clutter, blur, and long-range targets), without requiring manual box annotation at scale.

2.1.1 Ensemble

Models. The pseudo-label generator uses a three-model detector ensemble comprising SAM3, GroundingDINO and Faster R-CNN, complemented by a pixel-level semantic prior heatmap given by Talk2DINO.

Faster-RCNN [28] is a well-known classic CNN that introduced the Region Proposal Network (RPN) to avoid the use of an external Selective Search algorithm and allowed running the backbone only once on the entire image.

GroundingDINO [25] is an open-vocabulary vision–language detector that integrates transformer-based object detection with language-guided queries, enabling zero-shot localization of objects specified through natural language prompts. In this work it is prompted with `"boat, ship"` to capture vessel-related semantic variants while generating bounding box proposals and con-

fidence scores.

SAM3 [27] is a promptable segmentation model that extends the Segment Anything paradigm with concept-aware reasoning and improved instance identification. When prompted with `"boat"`, it produces high-quality segmentation masks that are converted to bounding boxes, providing precise geometric localization cues to the ensemble.

Finally, L. Barsellotti, L. Bianchi *et al.* developed Talk2DINO [26] that leverages spatially detailed self-supervised DINO representations to produce open-vocabulary semantic heatmaps conditioned on text prompts. Using `"boat, ship"`, it generates a dense pixel-level relevance map indicating likely vessel regions.

Within the ensemble, Faster-RCNN, GroundingDINO and SAM3 provide complementary detection proposals with different vision–language priors and spatial granularity, while Talk2DINO supplies a semantic consistency prior used to validate fused detections at the pixel level.

Score Weighting. Let $\mathcal{B}_r, \mathcal{B}_s, \mathcal{B}_d$ be the sets of boxes returned by Faster R-CNN, SAM3, and Grounding DINO, with confidence scores $p_r, p_s, p_d \in [0, 1]$. The candidate boxes are geometrically fused when $\mathrm{IoU}(b_i, b_j) > \tau_{\mathrm{IoU}} \quad \forall b_i, b_j \in (\mathcal{B}_r \cup \mathcal{B}_s \cup \mathcal{B}_d)$ with $b_i \neq b_j$ and $\tau_{\mathrm{IoU}} = 0.5$. The candidate scores are re-weighted by three hyperparameters $\{w_r = 0.5,\ w_s = 4.0,\ w_d = 1.2\}$. The result is a cluster $\mathcal{C}$ that generates a coordinate-wise average box $\bar{b} = \frac{1}{|\mathcal{C}|} \sum_{b \in \mathcal{C}} b$, with a detector agreement confidence score defined as the sum of weighted confidence scores of all boxes assigned to $\mathcal{C}$:

$$S_{\mathrm{det}}(\mathcal{C}) = \sum_{k \in \{r,s,d\}} \sum_{b \in (\mathcal{C} \cap \mathcal{B}_k)} w_k \, p_k(b)$$

where $p_k(b)$ and w_k are respectively the confidence score of box b predicted by detector $k \in \{r, s, d\}$ and the weight associated with detector k.

Let $H_t \in [0, 1]^{H \times W}$ denote the Talk2DINO heatmap that is used to score semantic consistency inside $\bar{b}$. Given the re-weighting hyperparameter $w_t = 2.1$ and $\Omega(\bar{b})$ as the set of pixels inside the fused box, the respective heatmap agreement confidence score is:

$$S_{\mathrm{heat}}(\bar{b}) = w_t \cdot \frac{1}{|\Omega(\bar{b})|} \sum_{(x,y) \in \Omega(\bar{b})} H(x, y)$$

The final normalized agreement score of the pseudo-label is:

$$S_{\mathrm{tot}} = \frac{S_{\mathrm{det}} + S_{\mathrm{heat}}}{w_r + w_s + w_d + w_t}.$$

where clusters containing boxes from only a subset of detectors receive lower normalized scores, implicitly encouraging multi-detector consensus.

A pseudo-label is retained if $S_{\text{tot}} > 0.42$ which correspond to roughly 3.3 out of 7.8, the maximum un-normalized overall agreement score.

Hyperparameters. The weights $\{w_r, w_s, w_d, w_t\}$, IoU threshold τ_{IoU} and final agreement score threshold are set manually to retain a robust ensable performance under small variations on the chosen values, guided by each component's contribution in the ablation studies reported below.

Evaluation. The ensemble is evaluated against the publicly available infrared-boat IR vessel dataset [21], achieving mAP (0.50–0.95) $= 0.6170$, mAP@50 $= 0.8412$, mAP@75 $= 0.7556$, Precision $= 0.8015$, Recall $= 0.7492$, and F1 Score $= 0.7744$.

Ablation Studies. Under the same evaluation protocol, the contribution of each component of the ensemble is tested. SAM3 is by far the strongest model in the ensemble on all metrics. Other models are retained because they make complementary errors (see Figure 2.1) and allow the ensemble to outperform each individual model in at least one metric (see Table 2.1).

Model	**mAP (0.50–0.95)**	**mAP@50**	**mAP@75**	**Precision**	**Recall**	**F1 Score**
Faster R-CNN only	0.3923	0.5618	0.4762	0.5161	0.4981	0.5069
SAM3 only	**0.6298**	0.8326	0.7467	**0.8015**	**0.7502**	**0.7750**
Grounding DINO only	0.4457	0.5861	0.5327	0.5474	0.5430	0.5452
Heatmap only	0.2557	0.3789	0.3057	0.6737	0.6615	0.6675
SAM3 + Heatmap	0.6096	**0.8428**	0.7439	**0.8015**	**0.7502**	**0.7750**
SAM3 + DINO	0.6129	0.8095	0.7319	<u>0.8000</u>	<u>0.7495</u>	0.7739
SAM3 + RCNN	<u>0.6199</u>	0.8284	0.7450	**0.8015**	**0.7502**	**0.7750**
SAM3 + DINO + Heatmap	0.6187	0.8395	<u>0.7552</u>	**0.8015**	0.7492	<u>0.7744</u>
Full Ensemble	0.6170	<u>0.8412</u>	**0.7556**	**0.8015**	0.7492	<u>0.7744</u>

Table 2.1: Ablation results for individual and combined ensemble components. Best results are in **bold** and second-best results are <u>underlined</u>.

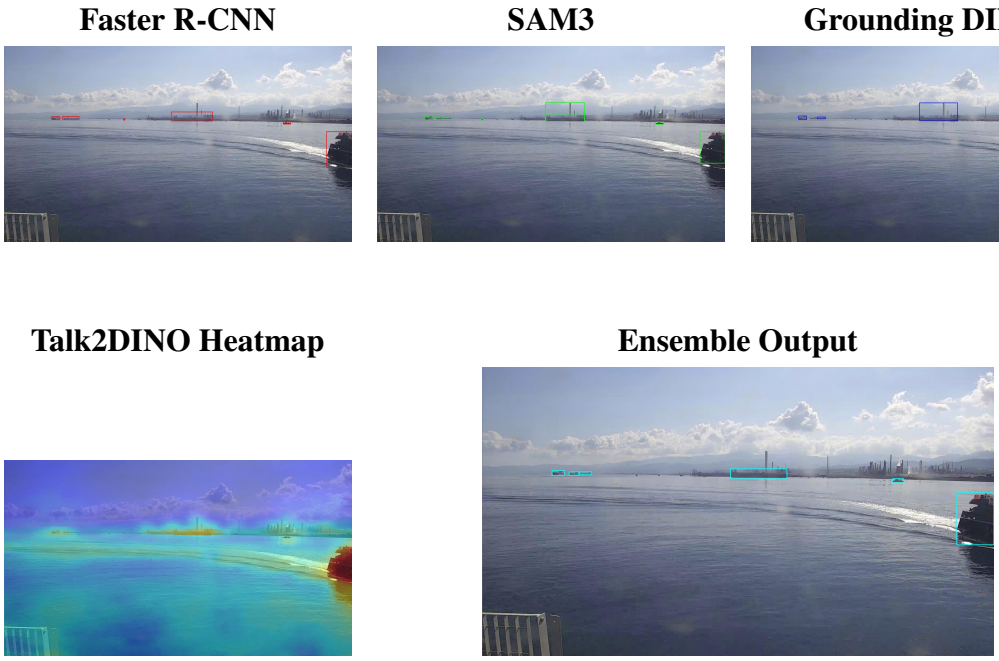


Figure 2.1: Qualitative comparison of detector components and final fusion output. Top row: individual detector predictions (Faster R-CNN, SAM3, Grounding DINO). Bottom row: Talk2DINO heatmap and final ensemble output.

2.1.2 Tracking

This stage stabilizes pseudo-label trajectories over time enforcing temporal consistency and detection linkage across consecutive frames through tracking.

Tracker. The SAM3 tracker [27] integrates an open-vocabulary segmentation model with a memory-augmented temporal tracking mechanism. This step interacts with the multimodal promptable memory-bank system that holds the object masks (masklets) found up to the current time step.

Salient points. Once the ensemble outputs a prediction, Talk2DINO [26] heatmaps are exploited again to extract the most prominent pixel in the DINOv3 feature space that better aligns with the prompt `"boat, ship"` inside each of the candidate bounding boxes. This pixel coordinate acts as a salient point to which SAM3 anchors for extracting a mask.

Results. Figure 2.2 reports a 2x3 visual comparison over three consecutive timesteps. The first column shows the ensemble detections, while the second column shows the corresponding SAM3 tracking outputs.

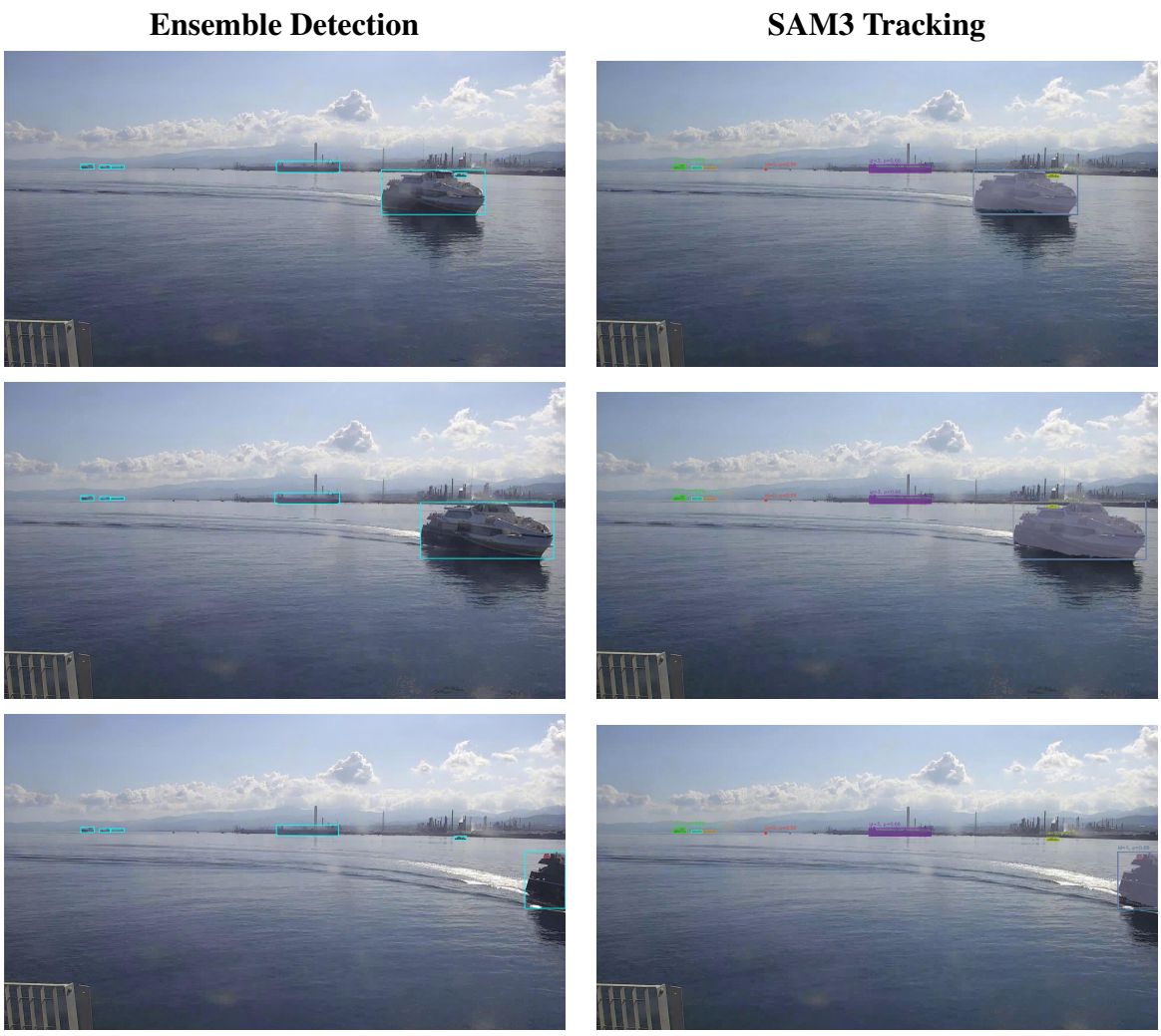


Figure 2.2: Tracking comparison at three timesteps of the same video stream. Left column: ensemble detections. Right column: SAM3 tracking outputs. Note the yellow target consistency across consecutive frames.

2.1.3 Keypoint Matching

Keypoint matching is used between synchronized IR and DLTV frames, to improve robustness at night and across modalities. Homographies estimated from matched keypoints allow transferring detections from clearer IR views to degraded DLTV frames, especially in cases affected by blur or autofocus failures. This step increases annotation coverage in difficult scenarios, under the assumption that the same homography can be calculated once and applied for

all frames iff the FOV of both IR and DLTV channels does not change over time.

SIFT. A first attempt exploit a classic algorithm in keypoint matching presented by David G. Lowe in 2004 [29], the Scale-Invariant Feature Transform (SIFT). SIFT extracts distinctive local features that are invariant to image scale and rotation and generate descriptors using local gradient histograms. The results are keypoints that can be compared on a pair of images to find a match, i.e. hopefully the same correspondance of an query object.

Given a synchronized pair of IR and DLTV frames, SIFT keypoints and descriptors are extracted from both grayscale images. Descriptor matching is performed using a FLANN-based nearest-neighbor search with a KD-tree index. To remove ambiguous correspondences, matches are filtered using the Lowe ratio test with threshold $\tau_L = 0.8$, reducing the amount of false positive matches. A geometric transformation between the two views is then estimated by computing a homography using RANSAC [30] sampling algorithm.

Learned Matching Models. Recent works have proposed learning-based alternatives to classical keypoint pipelines. In particular, MatchAnything [31] introduces a universal feature matching framework capable of establishing correspondences across heterogeneous image modalities. The method builds upon a detector-free transformer architecture derived from Efficient LoFTR[32] and predicts semi-dense correspondences directly between image regions, avoiding the explicit keypoint detection and descriptor computation stages required by traditional pipelines such as SIFT. The model is pre-trained on a mixture of large-scale datasets and synthetic cross-modal image pairs, DLTV-IR pairs among others.

Results. Qualitative comparisons in Figure 2.3 show that the ELoFTR-based model (MatchAnything) consistently overcomes SIFT on IR–DLTV pairs, as expected for cross-modal matching. In particular, learned correspondences are more stable and produce a more reliable geometric mapping, while SIFT often yields noisier or less repeatable matches.

There's a catch: ELoFTR does not perform equally well on all raw full-frame pairs, and cropping the DLTV image around the approximate search region is usually required. This effectively injects a weak prior about where correspondences are expected, and in this regime ELoFTR works significantly better. Also, a good homography is obtained only if it is possible to collect a well-distributed set of points across the image plane, also from different image pairs. For this reason, this is the step that requires the most supervision overall, but if handling of blurry images is not a requirement it can be skipped.

Figure 2.4 shows the final result of using keypoint-matching pseudo-label transfer from an IR image to a DLTV one.

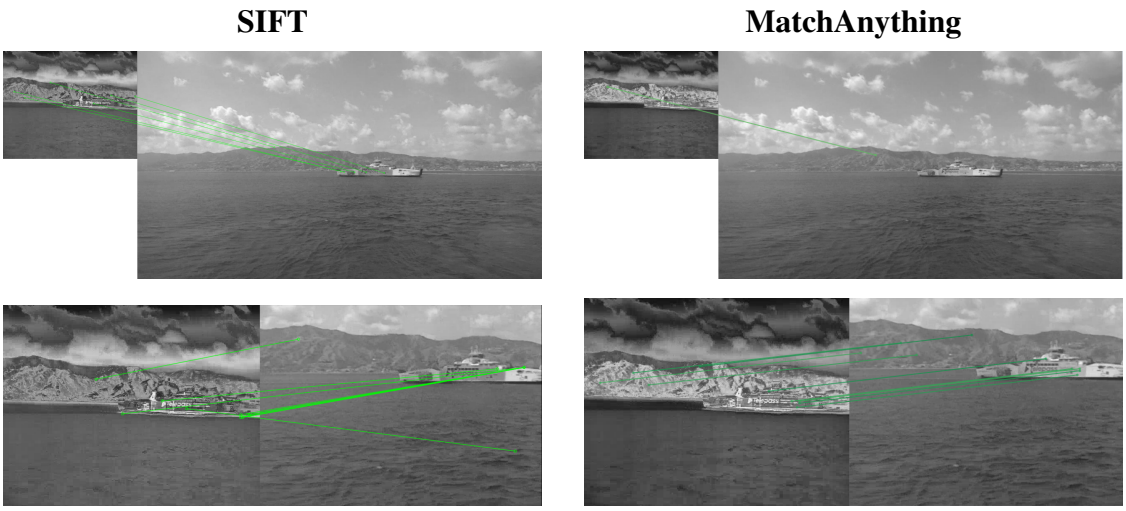
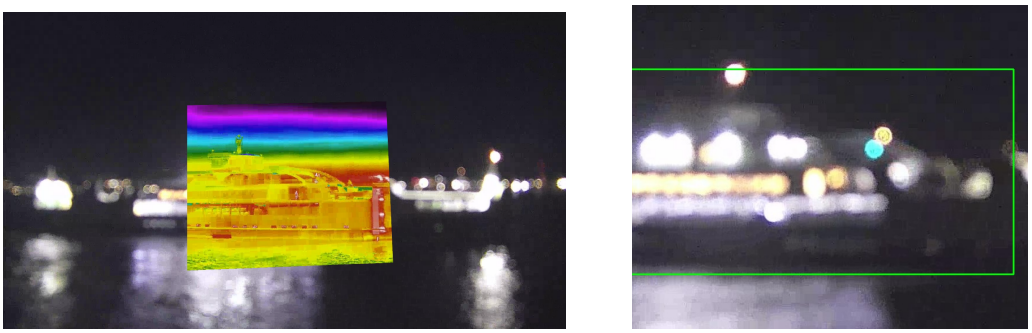


Figure 2.3: Qualitative keypoint-matching comparison of SIFT and ELOFTR-based approach (MatchAnything) on the same IR–DLTV image stereo pair under two pre-processing setups. First row has a full IR to full DLTV matching. In the second row the DLTV is cropped to match IR size and position approx.



Warped IR onto DLTV frame | Transferred pseudo-label (IR → DLTV)

Figure 2.4: Example of homography-based transfer using MatchAnything correspondences. Left: IR frame warped into the DLTV view. Right: corresponding detection pseudo-label transferred from IR to DLTV through the estimated homography.

2.2 YOLO Training

The Detection Dataset 4K introduced in Section 2.1 is used to train surrogate YOLO detectors for vessel detection.

Two large detection backbones are considered: YOLOv8l [2] and YOLO26l [3]. Both are evaluated first in their base pretrained configuration and then after fine-tuning on generated dataset. In practice, the fine-tuned models produce substantially cleaner vessel boxes, fewer spurious detections on sea clutter, and stronger recall on distant or low-contrast vessels.

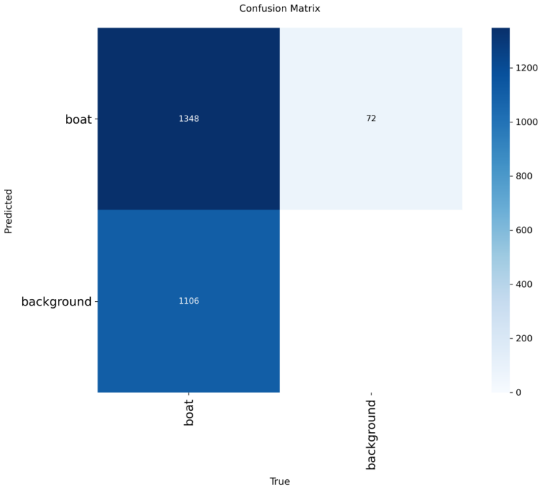
For reproducibility, both YOLOv8l and YOLO26l are trained with the same core setup: 100 maximum epochs with early stopping patience of 15 epochs, batch size 32, and input resolution 640. Detailed fine-tuning curves are reported in Appendix A (Figures A.1 and A.2) for further training diagnostics.

Table 2.2 summarizes the detection results before and after training on the pseudo-labeled dataset.

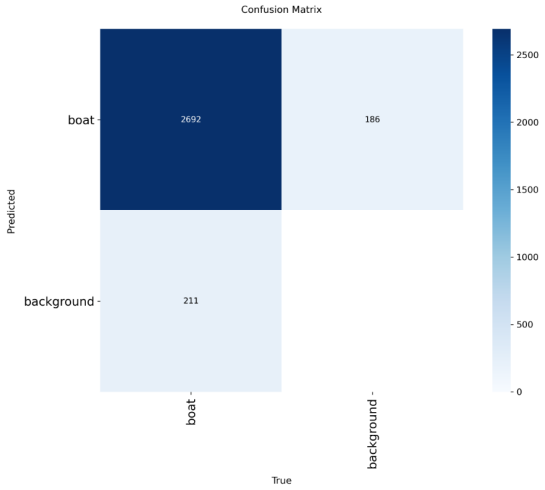
Model	**mAP@50**	**mAP@50–95**	**Precision**	**Recall**	**F1**
YOLOv8l	0.052	0.023	0.092	0.080	0.086
YOLOv8l Fine-tuned	<u>0.809</u>	<u>0.540</u>	<u>0.876</u>	<u>0.663</u>	<u>0.755</u>
YOLO26l	0.028	0.012	0.051	0.037	0.043
YOLO26l Fine-tuned	**0.951**	**0.675**	**0.950**	**0.885**	**0.916**

Table 2.2: YOLO detection performance before and after training on Detection Dataset 4K (Section 2.1). Best results are in **bold** and second-best results are <u>underlined</u>.

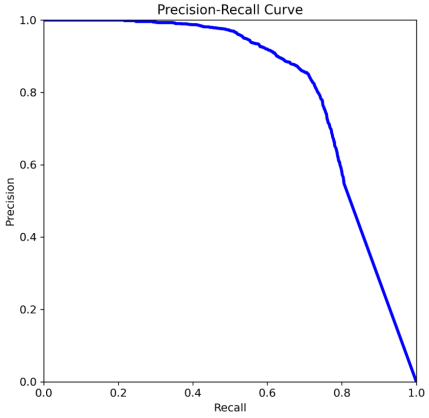
Overall, the results indicate that pseudo-label training is highly effective as shown in Figure 2.5 and Figure 2.6: both architectures improve by large margins across all metrics, with YOLO26l achieving the strongest final accuracy and best precision–recall trade-off.



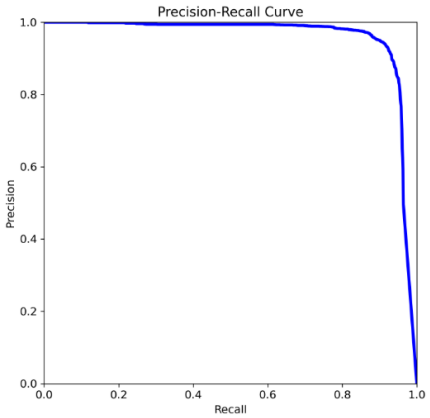
YOLOv8l fine-tuned confusion matrix



YOLO26l fine-tuned confusion matrix



YOLOv8l fine-tuned precision–recall curve



YOLO26l fine-tuned precision–recall curve

Figure 2.5: Evaluation visualizations for fine-tuned YOLO models trained on the generated pseudo-labels. Top row: confusion matrices for YOLOv8l (left) and YOLO26l (right). Bottom row: precision–recall curves for YOLOv8l (left) and YOLO26l (right).

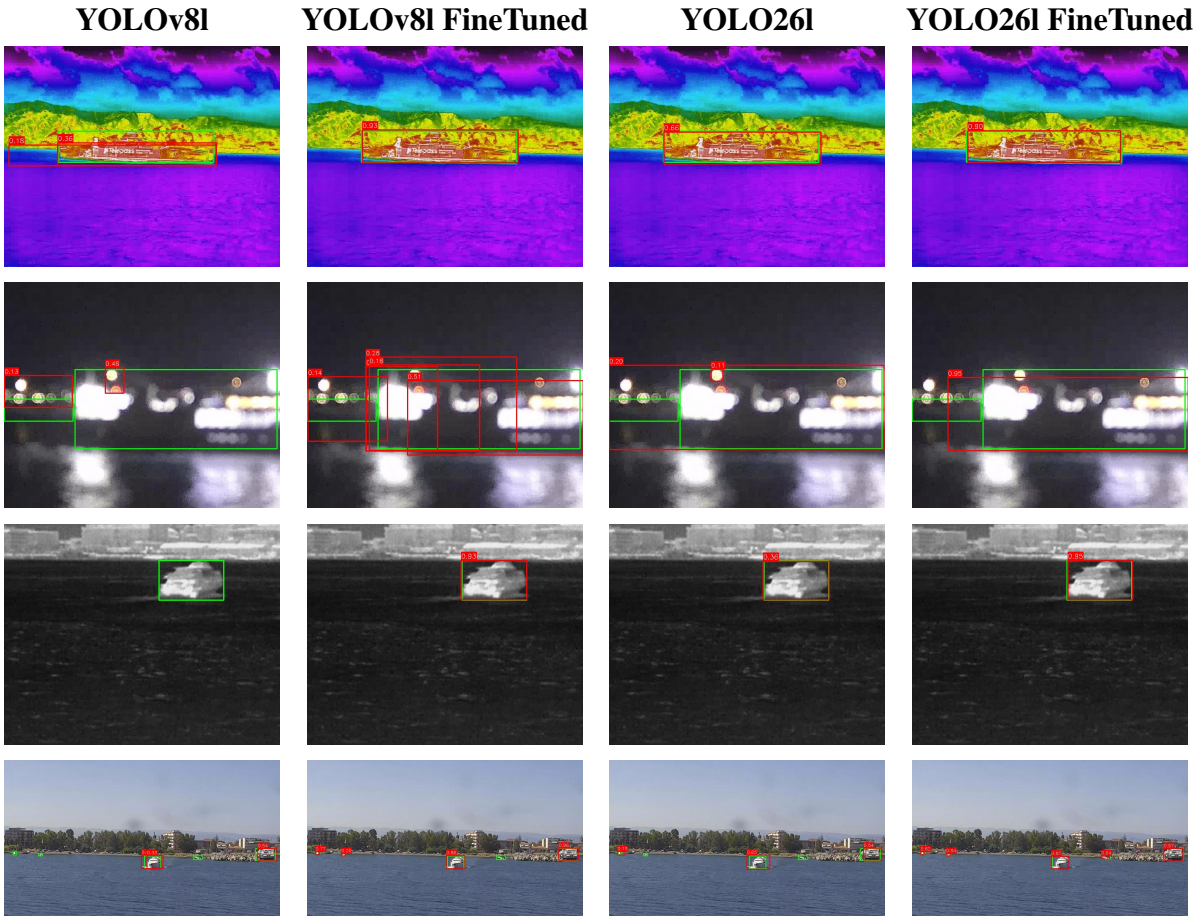


Figure 2.6: Qualitative detection comparison across models. Columns correspond to models (YOLOv8l base, YOLOv8l finetuned, YOLO26l base, YOLO26l finetuned), while rows correspond to four different input images (colored IR, night-time DLTV, grayscale IR, day-time DLTV). Pseudo-labels are in green, prediction in red.

Chapter 3

Vessel Re-Identification

This chapter presents the vessel re-identification pipeline built on top of the detection outputs introduced earlier. It starts from geometric association between AIS trajectories and image detections through perspective projection, then describes pseudo-label generation for vessel identities, and finally reports the metric-learning setup and evaluation protocol used to train and benchmark re-identification models.

3.1 Perspective Projection

This section presents the overall algorithm used to associate AIS positions with visual vessel detections. The Automatic Identification System (AIS) is a maritime tracking system in which vessels broadcast identity, position, course, speed, and navigation status through VHF radio transponders at regular intervals, while nearby shore stations and vessels receive these messages for monitoring and collision-avoidance purposes. The pipeline is organized in three stages: (i) conversion within the World Reference Frame (WRF), (ii) transformation from WRF to the Camera Reference Frame (CRF), and (iii) projection from CRF to the image plane. The results are shown in Figure 3.2, where synchronized daylight and infrared stereo images are the background for the overlayed AIS data, while Figure 3.1 provides the geospatial context of the same AIS observations and the optics fields of views.

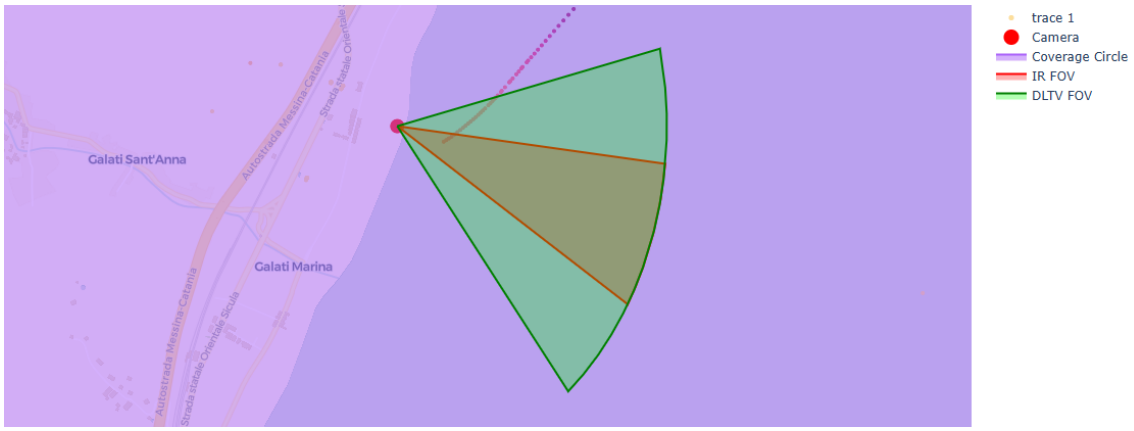
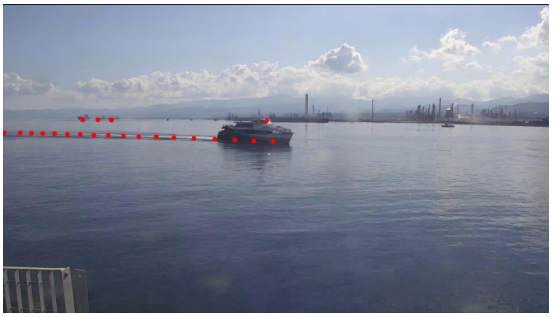


Figure 3.1: Map view of AIS data. AIS points are color-coded with darker tones for higher velocity values; the camera position is marked in red; the approximately 10 km AIS visibility limit is shown in violet; IR and DLTV fields of view are shown for the stereo configuration.

IR frame at the same timestamp of map in Figure 3.1.



DLTV frame at the same timestamp of map in Figure 3.1.

Figure 3.2: Synchronized stereo frames corresponding to the same timestamp used in AIS-to-image association.

In this context, WRF denotes the global/local world coordinate system used to represent vessel positions in physical space, while CRF denotes the camera-centered coordinate system whose origin is the camera and whose axes follow camera orientation.

3.1.1 WRF

ECEF. AIS streams provide vessel positions in non-Cartesian geodetic coordinates (latitude, logitude, height) that are converted to ECEF (”Earth-Centered, Earth-Fixed”), a global Carte-

sian reference frame where the origin is at the Earth's center of mass, and the axes rotate with the Earth. This simplifies distance computation and vector operations.

ECEF uses a right-handed coordinate system:

- Z axis: toward the North Pole;
- X axis: intersection of Equator and Greenwich meridian (lat $= 0°$, lon $= 0°$);
- Y axis: intersection of Equator and $90°$ East longitude.

Coordinates are usually expressed in meters.

For geodetic-to-ECEF conversion, this work uses the standard WGS84 reference ellipsoid.

ENU. "East, North, Up" is a local observer-centric Cartesian frame centered at a chosen reference point. It makes global coordinates interpretable for tracking tasks.

- E: tangent to Earth, pointing east;
- N: tangent to Earth, pointing north;
- U: normal to the local surface, pointing upward.

Conceptually, the ENU conversion is performed by selecting a local reference point $(\text{lat}_0, \text{lon}_0, \text{alt}_0)$, converting both reference and target to ECEF, and then applying the local tangent-plane rotation to the relative ECEF displacement vector. The geometric relation between the global and local frames is illustrated in Figure 3.3.

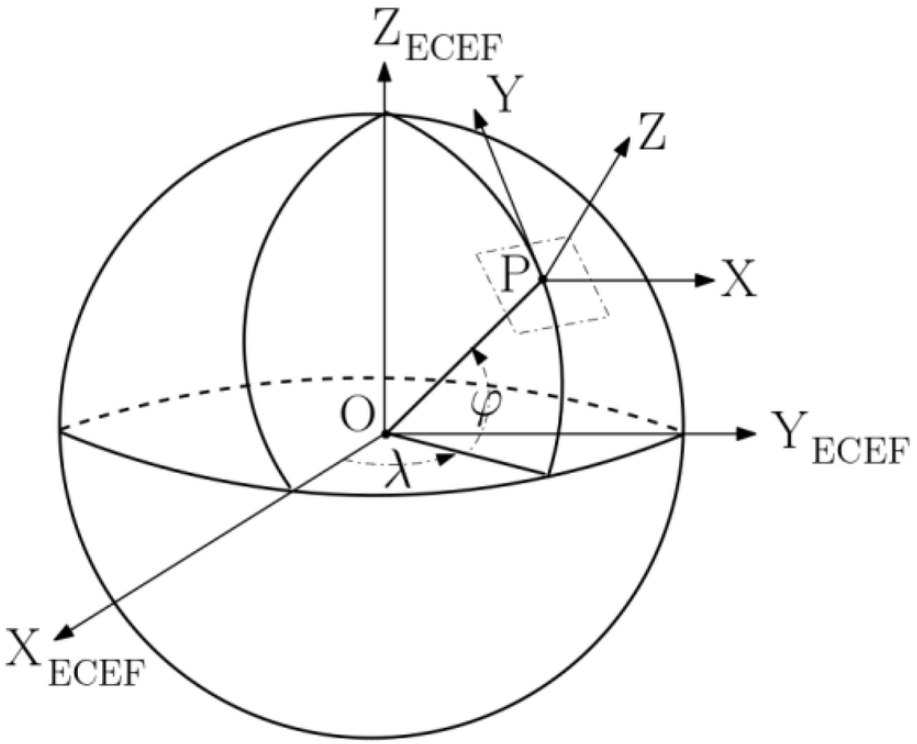


Figure 3.3: ECEF to ENU coordinate conversion. Point P is the reference geodetic coordinate, hence the camera position.

3.1.2 WRF to CRF

Given camera installation offset $\mathbf{c}_{ENU} = [E_c, N_c, U_c]^T$, azimuth α (clockwise from North, in degrees), and elevation β (degrees above the horizon), we first convert angles to radians:

$$\alpha_r = \alpha\,\pi/180, \qquad \beta_r = \beta\,\pi/180.$$

Following the implemented convention, the forward direction is built as

$$\tilde{\mathbf{f}} = \begin{bmatrix} \cos(-\beta_r)\sin(\alpha_r) \\ \cos(-\beta_r)\cos(\alpha_r) \\ \sin(-\beta_r) \end{bmatrix}, \qquad \mathbf{f} = \frac{\tilde{\mathbf{f}}}{\|\tilde{\mathbf{f}}\|}.$$

Let the world up vector be

$$\mathbf{u}_w = [0, 0, 1]^T.$$

The camera right and up vectors are then computed with cross products:

$$\tilde{\mathbf{r}} = \mathbf{f} \times \mathbf{u}_w, \qquad \mathbf{r} = \frac{\tilde{\mathbf{r}}}{\|\tilde{\mathbf{r}}\|}, \qquad \mathbf{u} = \mathbf{r} \times \mathbf{f}.$$

Using homogeneous coordinates, the rotation matrix is assembled with camera axes as rows (as in the code):

$$R = \begin{bmatrix} \mathbf{r}^T & 0 \\ \mathbf{u}^T & 0 \\ \mathbf{f}^T & 0 \\ 0\,0\,0 & 1 \end{bmatrix}.$$

The translation matrix is

$$T = \begin{bmatrix} 1 & 0 & 0 & -E_c \\ 0 & 1 & 0 & -N_c \\ 0 & 0 & 1 & -U_c \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Therefore, the world-to-camera view matrix is

$$V = RT.$$

For a world point in homogeneous ENU coordinates $\mathbf{p}_{w,h} = [E, N, U, 1]^T$, camera coordinates are obtained as

$$\mathbf{p}_{c,h} = V\mathbf{p}_{w,h}.$$

3.1.3 CRF to Image plane

The camera intrinsic matrix is derived from horizontal field of view, vertical field of view, and frame size. Let image width and height be W and H, and let ϕ_x and ϕ_y be the horizontal and vertical field-of-view angles (in radians). When ϕ_y is not provided, it can be obtained from aspect ratio as:

$$\phi_y = 2\arctan\left(\frac{H}{W}\tan\frac{\phi_x}{2}\right).$$

The focal lengths in pixels are:

$$f_x = \frac{W}{2\tan(\phi_x/2)}, \qquad f_y = \frac{H}{2\tan(\phi_y/2)}.$$

The principal point is assumed at image center:

$$c_x = \frac{W}{2}, \qquad c_y = \frac{H}{2}.$$

Thus,

$$K = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix},$$

where (f_x, f_y) are focal lengths in pixels and (c_x, c_y) is the principal point. This matrix maps 3D camera coordinates to the image plane and enables association between AIS-derived projections and vessel detections. For any ENU input point, the composed operator K $[R\ T\ ]$ provides the final image-plane coordinates used for AIS-to-detection association.

3.2 Re-Identification Pseudo-Labels Generation

After perspective projection, pseudo-label generation is performed by combining two complementary inputs: (i) vessel detections produced by the surrogate YOLO model and (ii) AIS-derived projected points on the image plane. The objective is to assign robust pseudo identity labels by associating each AIS target with the geodesic nearest detected bounding box. This is better than associating using pixel distances in the image plane, because perspective causes depth-dependent scale distortion errors which get amplified in AIS crowded images.

A single representative bounding-box point is selected at the midpoint of the box base, which is a robust proxy of vessel contact with the sea in a land-surveillance setting. This point is back-projected into the ENU world frame by calculating the ray intersection with the ground plane. More details are provided in Appendix A.

In practice, the distance computed from this inverse projection can be imprecise under slight camera miscalibration or malformed bounding boxes. The main reason is that distance scales nonlinearly (approximately hyperbolically) with respect to linear variations in the vertical image coordinate. A simple mitigation is to shift projected AIS points slightly downward in the image. This can be obtained by injecting a small pitch offset, i.e. modifying the camera elevation, but with the effect of an almost uniform image-space translation for all targets, as illustrated in Figure 3.4. A more effective correction is to adjust the assumed camera height instead: near-camera projections move downward more than points near to the horizon, miti-

gating the effect only for relevant boxes, as shown in Figure 3.5.

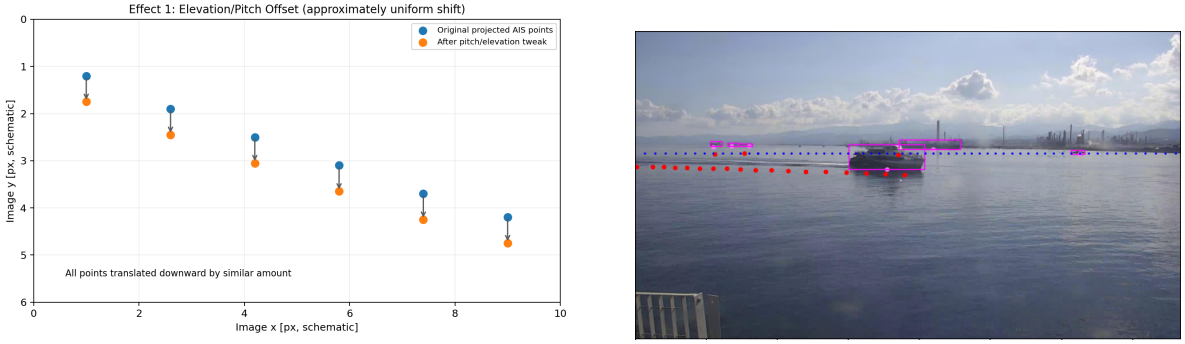


Figure 3.4: Effect of a small pitch/elevation correction on projected AIS red points: the displacement is approximately a uniform downward translation in image space also at the horizon (blue dots). YOLO detection are in violet.

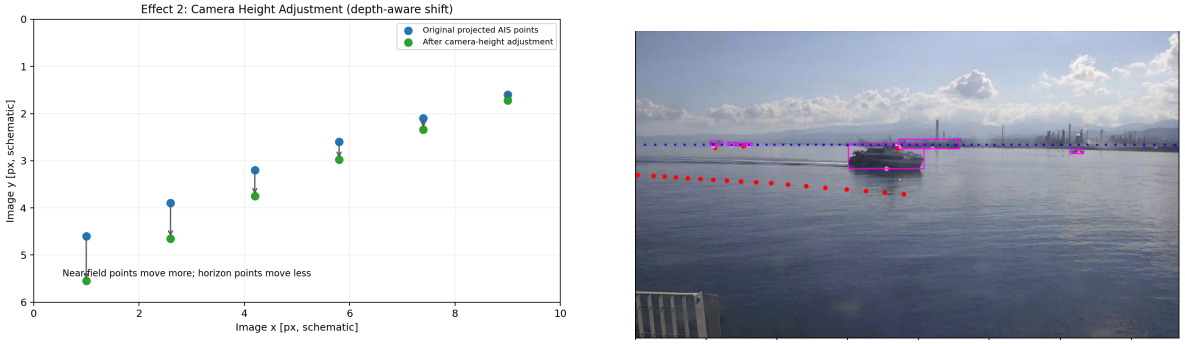


Figure 3.5: Effect of camera-height correction on projected AIS red points: displacement is depth-aware, with stronger shifts for near-camera points and smaller shifts near the horizon (blue dots).YOLO detection are in violet.

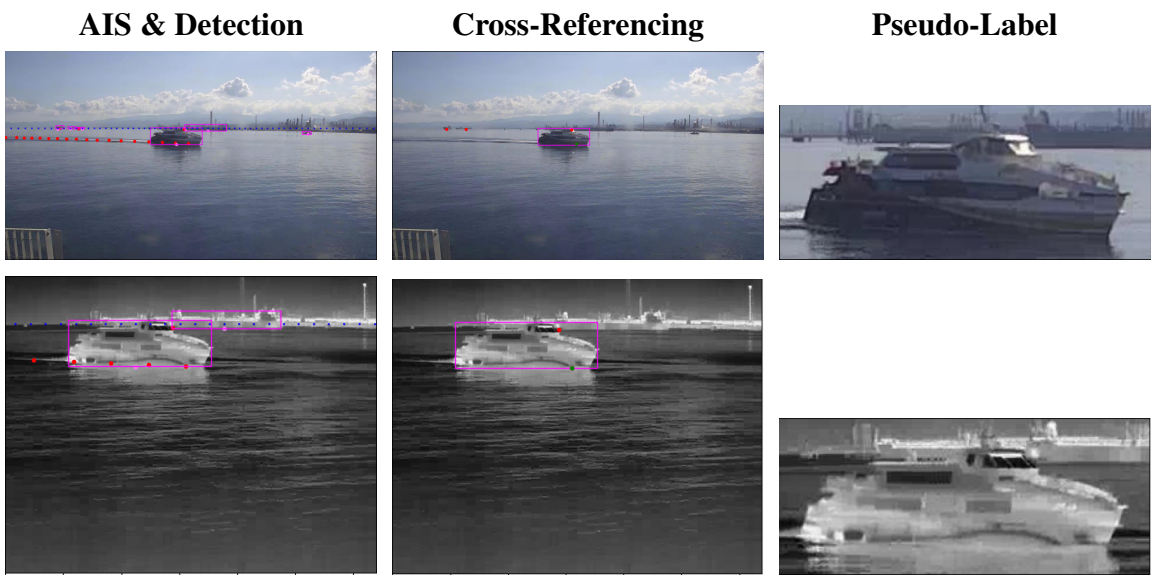


Figure 3.6: Three-stage cross-referencing pipeline for pseudo-label generation. Columns report (left to right): AIS with detection, cross-referencing (green dot), and final cross-referenced pseudo-labels. Rows show DLTV examples (top) and IR examples (bottom).

3.3 Metric Learning

This section summarizes the training methodology adopted for vessel re-identification using the pseudo-label dataset generated in the previous stages. The objective is to learn an embedding space where samples of the same vessel are close and samples of different vessels are well separated.

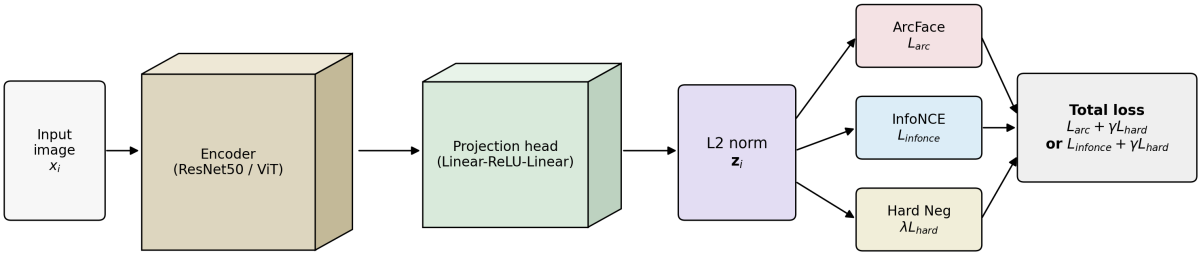


Figure 3.7: Metric-learning training network used for vessel re-identification. The embedding branch feeds ArcFace, InfoNCE, and hard-negative components, which are combined in the final training objective.

The pipeline shown in Figure 3.7 follows a contrastive metric-learning formulation with

three main components: (i) a visual encoder backbone, (ii) a projection head that maps encoder features into a compact embedding space, and (iii) a discriminative metric-learning loss. In the reference configuration, the encoder is initialized from pretrained weights and coupled with an ArcFace-based objective [33]. Hard-negative mining is integrated during training to focus optimization on difficult pairs and improve inter-class separation.

Given an input image x_i, the encoder produces a feature vector $\mathbf{f}_i$, which is mapped by a two-layer projection head and L2-normalized:

$$\mathbf{z}_i = \frac{W_2\,\sigma(W_1\mathbf{f}_i)}{\|W_2\,\sigma(W_1\mathbf{f}_i)\|_2},$$

where $\sigma(\cdot)$ is ReLU.

For supervised contrastive learning, the multi-positive InfoNCE [34] objective is:

$$\mathcal{L}_{\text{supcon}} = -\frac{1}{B}\sum_{i=1}^{B}\frac{1}{|P(i)|}\sum_{p\in P(i)}\log\frac{\exp\left(\mathbf{z}_i^\top\mathbf{z}_p/\tau\right)}{\sum_{a\neq i}\exp\left(\mathbf{z}_i^\top\mathbf{z}_a/\tau\right)},$$

where $P(i)$ is the set of samples in the batch with the same identity as i, and τ is the temperature. The batch size B is 128.

For ArcFace, cosine logits are margin-modified on the target class:

$$\cos\theta_j = \mathbf{z}_i^\top\hat{\mathbf{w}}_j, \qquad \ell_j = \begin{cases} s\cos(\theta_{y_i}+m), & j = y_i, \\ s\cos\theta_j, & j \neq y_i, \end{cases}$$

where $\mathbf{z}_i$ is the L2-normalized embedding of sample i, and $\hat{\mathbf{w}}_j$ is the L2-normalized class-prototype (classifier weight vector) for class j. With additive angular margin m (default to 0.5) and scale s (default to 30), the final loss is standard cross-entropy on ℓ_j.

When enabled, hard-negative mining adds a regularization term based on the mean similarity of the top-k hardest negatives per anchor:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{arcface}} + \lambda\,\mathcal{L}_{\text{hardneg}},$$

where λ controls the hard-negative penalty weight, defaults to 0.1.

Training and validation splits are generated with a stratified protocol to preserve class balance across identities. A complete implementation hyperparameters are reported in Ap-

pendix A

Evaluation is performed with complementary retrieval-oriented and separability-oriented metrics. Specifically, k-NN accuracy (higher is better) measures how often the nearest neighbors in embedding space share the correct identity; centroid-based accuracy (higher is better) measures whether each sample is closest to its class centroid; centroid mAP (higher is better) summarizes ranking quality across identities using centroid-based retrieval; positive similarity (higher is better) and negative similarity (lower is better) quantify intra-class compactness and inter-class separation, respectively; and the embedding-dispersion indicator (lower is better) measures the overall spread of embeddings, i.e. global compactness. This combination provides both performance and representation-quality perspectives.

Hyperparamters Ablation Studies. This first ablation group isolates the effect of training design by fixing the encoder to ResNet50 [8] and modifying only the optimization setup. We compare alternatives for scheduler and optimizer settings, loss composition, hard-negative mining, and augmentation policy to quantify which hyperparameter choices most influence vessel re-identification performance under a controlled backbone.

Ablation Term	**Value**	**Centr. mAP**	**KNN Acc.**	**Centr. Acc.**	**PosSim**	**NegSim**	**Emb. Std.**
/	/	**0.984**	**0.972**	<u>0.959</u>	0.661	0.010	0.062
ArcFace Margin	0.300	0.980	0.964	0.954	0.556	-0.004	0.062
ArcFace Margin	0.700	0.981	<u>0.971</u>	0.955	0.747	-0.000	0.062
InfoNCE Temp.	$\tau = 0.100$	0.940	0.948	0.947	0.795	0.163	<u>0.056</u>
InfoNCE Temp.	$\tau = 0.200$	0.883	0.919	0.911	<u>0.815</u>	0.152	<u>0.056</u>
Hard-negative Top-k	None	0.980	**0.972**	<u>0.959</u>	0.647	0.004	0.062
Hard-negative Top-k	10	**0.984**	0.968	0.951	0.648	-0.005	0.062
Hard-negative Top-k	20	<u>0.983</u>	**0.972**	0.956	0.638	-0.001	0.062
Projection Dim.	256	0.982	0.970	0.955	0.650	-0.003	0.062
Projection Dim.	512	<u>0.983</u>	0.969	**0.961**	0.640	0.007	**0.044**
Learning rate	1×10^{-3}	0.977	0.970	<u>0.959</u>	**0.871**	**-0.007**	0.062
Color jitter	Disabled	0.979	0.968	0.951	0.616	<u>-0.006</u>	0.062
Horizontal flip	Disabled	0.982	0.968	0.955	0.657	-0.001	0.062

Table 3.1: Results of ablation studies under hyperparameter variations. Best results are in **bold** and second best results are <u>underlined</u>.

Table 3.1 shows that the reference configuration is already near-optimal for retrieval qual-

ity, with the best centroid mAP and kNN accuracy obtained by the default setup. Variations of ArcFace margin and projection dimension produce only marginal gains in specific secondary metrics, while increasing the InfoNCE temperature consistently degrades retrieval performance. Hard-negative mining provides limited but stable benefits depending on k, suggesting that conservative settings are preferable. Overall, these results motivate keeping the baseline hyperparameter recipe as the main training configuration for the following experiments.

Encoder Ablation Studies. This ablation isolates the effect of the image encoder by keeping the metric-learning pipeline fixed and swapping only the backbone architecture. The compared models span CNN and transformer families, including ResNet [8], EfficientNet [10], ConvNeXt [9], ViT [4] Base, Swin Transformer [5], ViT DINO [6], and ConvNeXt/ViT DINOv3 [7]. The objective is to evaluate how backbone design impacts retrieval quality (centroid mAP, kNN accuracy, centroid accuracy) and embedding geometry (positive/negative similarity and feature spread) for vessel re-identification.

backbone	**Centr. mAP**	**KNN Acc.**	**Centr. Acc.**	**PosSim**	**NegSim**	**Emb. Std.**
ResNet50	0.984	0.972	0.959	0.661	0.010	0.0619
ResNet101	0.980	0.966	0.949	0.626	0.000	0.0618
EfficientNet-B3	0.982	0.970	0.958	0.813	-0.008	0.0619
ConvNeXt Small (IN12K)	0.991	<u>0.984</u>	0.970	**0.945**	-0.003	**0.0616**
ConvNeXt DINOv3	0.988	0.978	0.968	<u>0.942</u>	-0.004	**0.0616**
Swin Small	0.991	0.981	<u>0.971</u>	0.836	<u>-0.010</u>	0.0619
SwinV2 Tiny	0.990	0.978	0.962	0.891	-0.004	<u>0.0617</u>
ViT Base (AugReg)	<u>0.992</u>	0.983	<u>0.971</u>	0.917	**-0.012**	0.0619
ViT DINO	0.988	0.979	0.969	0.906	<u>-0.010</u>	0.0619
ViT DINOv3	**0.996**	**0.985**	**0.978**	0.902	-0.008	0.0618

Table 3.2: Encoder-backbone ablation results under the reference training protocol. Best results are in **bold** and second best results are <u>underlined</u>.

Table 3.2 indicates a clear advantage of transformer backbones over CNN alternatives under the same training recipe. ViT DINOv3 achieves the best overall retrieval performance, with the highest centroid mAP, kNN accuracy, and centroid accuracy, while ConvNeXt Small and ViT Base (AugReg) remain competitive secondary choices. CNN backbones such as ResNet

and EfficientNet are generally strong but consistently below the top transformer models on the main ranking metrics. Overall, these results support selecting ViT DINOv3 as the default encoder for subsequent experiments.

Data Scaling. This ablation analyzes how metric-learning behavior changes as training data scale increases, both in number of identities and number of images per split. We compare a classical CNN baseline (ResNet50) against a stronger transformer backbone (ViT DINOv3) across three dataset sizes (2K, 7K, 15K), while keeping the optimization recipe fixed. The goal is to measure whether additional data primarily improves retrieval/classification metrics or instead affects embedding geometry (positive/negative similarity separation and feature dispersion).

Backbone	**#Ids**	**#Imgs**	**Centr. mAP**	**KNN Acc.**	**Centr. Acc.**	**PosSim**	**NegSim**	**Emb. Std.**
ResNet50(Baseline)	64	2K	0.984	0.972	0.959	0.661	0.010	**0.062**
ResNet50	266	7K	0.968	0.967	0.947	0.613	0.000	**0.062**
ResNet50	484	15K	0.975	0.966	0.947	0.635	0.001	**0.062**
ViT DINOv3	64	2K	**0.996**	0.985	0.978	**0.902**	**-0.008**	**0.062**
ViT DINOv3	266	7K	0.992	0.985	0.978	0.855	-0.001	**0.062**
ViT DINOv3	484	15K	0.993	**0.988**	**0.981**	0.822	0.000	**0.062**

Table 3.3: Data-scaling ablation results under the reference training protocol. Best results are in **bold** and second best results are underlined.

Table 3.3 shows that the ViT DINOv3 backbone remains consistently stronger than ResNet50 at every data scale. Increasing data volume from 2K to 15K mainly benefits ranking/classification metrics for ViT DINOv3 (best kNN and centroid accuracy at 15K), while ResNet50 improves only marginally and remains below transformer performance. At the same time, similarity statistics become less separated as scale increases, suggesting harder identity discrimination despite better retrieval scores. Overall, these results confirm that larger datasets are beneficial when paired with stronger encoders, and they support the ViT DINOv3 + larger-scale setting as the preferred operating point.

Chapter 4

Conclusion and Future Works

This thesis set out to verify whether raw maritime surveillance streams can be transformed into usable training data for vessel detection and vessel re-identification with limited manual supervision. The results are consistent with this objective. The proposed pipeline shows that combining model agreement, temporal consistency, and AIS-driven geometric constraints can produce pseudo-labels that are sufficiently reliable to train competitive downstream models. In particular, detector fine-tuning on pseudo-labels yields large gains over base checkpoints, while the metric-learning pipeline demonstrates strong identity separability under multiple backbone and data-scale settings.

Beyond model accuracy, an important outcome is system-level feasibility: the work integrates data ingestion, pseudo-label generation, cross-modal association, training, and evaluation into a single reproducible workflow. This end-to-end perspective is one of the main practical contributions, because it moves from isolated algorithmic components toward a deployable methodology for real coastal monitoring scenarios.

At the same time, the current system has limitations that should be considered. First, Re-ID pseudo-label quality remains sensitive to calibration errors and difficult environmental conditions (night glare, haze, clutter, distant vessels), which can propagate errors to both detection and Re-ID training. Second, identity supervision is still indirect: AIS-based association can be ambiguous in crowded scenes or during crossings, introducing potential label drift. Third, evaluation is performed on a limited number of operational contexts; broader geographic and seasonal diversity is still needed to measure true generalization. Finally, this pipeline includes multiple stages and hyperparameters, which increases engineering complexity and may reduce robustness if components are not carefully maintained.

Future work should therefore focus on four directions. (i) Improve Re-ID pseudo-label reliability through uncertainty-aware filtering, confidence calibration, and stronger multi-frame/multi-camera consistency checks. (ii) Strengthen AIS–vision fusion with explicit temporal alignment models and probabilistic association to better handle missing or noisy transponder data. (iii) Expand validation with longer-term, multi-site benchmarks and standardized protocols for cross-domain testing. (iv) Increase deployment readiness by simplifying the pipeline, monitoring drift online, and introducing active-learning loops where human feedback is used only on high-uncertainty samples.

Overall, the thesis meets its core expectation: it demonstrates that weakly supervised, physically grounded pseudo-labeling is a viable strategy to bootstrap maritime detection and re-identification from real-world streams. The identified weak points are concrete and addressable, and they define a clear roadmap toward a more robust, scalable, and reliable maritime vision system.

Bibliography

[1] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[2] G. Jocher *et al.*, *Ultralytics YOLOv8*, software repository, 2023.

[3] R. Sapkota and M. Karkee, "Ultralytics YOLO Evolution: An Overview of YOLO26, YOLO11, YOLOv8 and YOLOv5," *arXiv preprint arXiv:2510.09653*, 2025.

[4] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *International Conference on Learning Representations (ICLR)*, 2021.

[5] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," arXiv preprint arXiv:2103.14030, 2021.

[6] M. Caron *et al.*, "Emerging Properties in Self-Supervised Vision Transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[7] O. Siméoni, H. V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, *et al.*, "DINOv3," *arXiv preprint arXiv:2508.10104*, 2025.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[9] Z. Liu *et al.*, "A ConvNet for the 2020s," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[10] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in *International Conference on Machine Learning (ICML)*, 2019.

[11] G. Matasci *et al.*, "Deep learning for vessel detection and identification from spaceborne optical imagery," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. V-3-2021, pp. 303–310, 2021.

[12] Y. Liu *et al.*, "Multi-Scale Correlation-Aware Transformer for Maritime Vessel Re-Identification," arXiv preprint arXiv:2511.14203, 2025.

[13] Lu *et al.*, "A New Large-Scale Dataset for Marine Vessel Re-Identification Based on Swin Transformer Network in Ocean Surveillance Scenario," *IET Computer Vision*, vol. 19, no. 2, pp. 101–115, 2025, doi:10.1049/cvi2.70007.

[14] G. Spagnolo *et al.*, "A new annotated dataset for boat detection and re-identification," CNR-ISTI, Italy, Tech. Rep., 2019.

[15] F. van Abbema, "Enhancing Vessel Re-identification with RGB-Infrared Multi-Modal Techniques," M.S. thesis, University of Twente, 2024.

[16] A. Galdelli, G. Narang, R. Pietrini, M. Zazzarini, A. Fiorani, A. N. Tassetti, "Multimodal AI-enhanced ship detection for mapping fishing vessels and informing on suspicious activities," *Pattern Recognition Letters*, vol. 180, pp. 45–59, 2025, doi:10.1016/j.patrec.2025.04.005.

[17] H.-C. Kim, H.-T. Lee, I.-S. Cho, "Vessel detection for maritime traffic management using U-Net with backbone networks," M.S. thesis, Ocean Engineering, 2025.

[18] D. Yang, M. I. Solihin *et al.*, "A streamlined approach for intelligent ship object detection using EL-YOLO algorithm," *Scientific Reports*, vol. 14, p. 12856, 2025, doi:10.1038/s41598-024-64225-y.

[19] T. Dong, M. Zhu, G. Tang, "Infrared ship target detection algorithm PEW_YOLOv8 in complex environments," *Scientific Reports*, vol. 15, p. 40574, 2026, doi:10.1038/s41598-026-40574-8.

[20] Y. Ginige *et al.*, "Vessel Re-identification and Activity Detection in Thermal Domain for Maritime Surveillance," arXiv preprint arXiv:2406.08294, 2024.

[21] NJUST, "infrared-boat Dataset," *Roboflow Universe*, Open Source Dataset, Aug. 2023. [Online]. Available: https://universe.roboflow.com/njust-oxpbo/infrared-boat. [Accessed: Mar. 8, 2026].

[22] H. Gupta, O. P. Verma, T. K. Sharma, H. Varshney, S. Agarwal, W. Pak, "Ship detection using ensemble deep learning techniques from synthetic aperture radar imagery," *Scientific Reports*, vol. 14, p. 80239, 2024, doi:10.1038/s41598-024-80239-y.

[23] J. Li *et al.*, "Unsupervised Maritime Vessel Re-Identification With Multi-Level Contrastive Learning," IEEE Transactions on Intelligent Transportation Systems, 2023, doi: 10.1109/TITS.2023.3243591.

[24] X. Zhang *et al.*, "A Transfer Learning-Based Approach to Maritime Warships Re-Identification," Engineering Applications of Artificial Intelligence, vol. 125, 2023, doi: 10.1016/j.engappai.2023.106617.

[25] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu, L. Zhang, "Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection," *arXiv preprint*, arXiv:2303.05499, 2023.

[26] L. Barsellotti, L. Bianchi, N. Messina, F. Carrara, M. Cornia, L. Baraldi, F. Falchi, R. Cucchiara "Talking to DINO: Bridging Self-Supervised Vision Backbones with Language for Open-Vocabulary Segmentation," ICCV, 2025.

[27] N. Carion, L. Gustafson, Y. Hu, S. Debnath, R. Hu, D. Suris, C. Ryali, K. V. Alwala, H. Khedr, "SAM 3: Segment Anything with Concepts," *arXiv preprint*, arXiv:2511.16719, 2025.

[28] S. Ren, K. He, R. Girshick, J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *Advances in Neural Information Processing Systems (NeurIPS)*, 2015, arXiv:1506.01497, doi:10.1109/TPAMI.2016.2577031.

[29] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004, doi:10.1023/B:VISI.0000029664.99615.94.

[30] M. A. Fischler, R. C. Bolles, "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981, doi:10.1145/358669.358692.

[31] X. He, H. Yu, S. Peng, D. Tan, Z. Shen, X. Zhou, H. Bao, "MatchAnything: Universal Cross-Modality Image Matching with Large-Scale Pre-Training, *arXiv preprint*, arXiv:2501.07556, 2025.

[32] Y. Wang, X. He, S. Peng, D. Tan, and X. Zhou, "Efficient LoFTR: Semi-Dense Local Feature Matching with Sparse-Like Speed, *arXiv preprint arXiv:2403.04765*, 2024.

[33] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," *arXiv preprint arXiv:1801.07698*, 2018.

[34] A. van den Oord, Y. Li, and O. Vinyals, "Representation Learning with Contrastive Predictive Coding," *arXiv preprint arXiv:1807.03748*, 2018.

Appendix A

Appendix

This appendix collects complementary technical material supporting the main chapters. It includes geometric derivations used in AIS-to-image association, additional qualitative and quantitative detection results, and the full reference configuration adopted for metric-learning experiments.

Inverse Distance Derivation. This section reports the geometric details used by the heuristic association algorithm introduced in Chapter 3.

The algorithm starts from each detected bounding box $b = (x_{\min}, y_{\min}, x_{\max}, y_{\max})$ and defines a single reference pixel at the midpoint of its base:

$$p_b = \left(\frac{x_{\min} + x_{\max}}{2},\ y_{\max} \right).$$

For land-surveillance cameras, this point is a good proxy for the vessel contact with sea level and is therefore more stable than the box center for inverse projection.

Given $p_b = (u, v)$, the inverse-projection pipeline follows the same camera model introduced above. First, the homogeneous pixel vector $\tilde{p} = [u, v, 1]^T$ is mapped to a camera ray using intrinsics:

$$\mathbf{r}_c = \frac{K^{-1}\tilde{p}}{\|K^{-1}\tilde{p}\|}.$$

Then the ray is rotated from CRF to ENU (world) coordinates through the inverse view transform:

$$\mathbf{r}_w = R^{-1} \begin{bmatrix} \mathbf{r}_c \\ 1 \end{bmatrix},$$

where only the first three components are used as the ray direction in ENU. Let the camera origin be $\mathbf{o}_w = [0, 0, h_c]^T$ and let the sea surface be approximated by the plane $z = h_g$ (typically $h_g = 0$). The back-projected 3D point is obtained by ray–plane intersection:

$$\mathbf{x}(t) = \mathbf{o}_w + t\,\mathbf{r}_w, \qquad t^* = \frac{h_g - o_{w,z}}{r_{w,z}}.$$

Valid solutions require $r_{w,z} \neq 0$ and $t^* > 0$. The resulting ENU point is

$$\mathbf{x}^* = \mathbf{o}_w + t^*\mathbf{r}_w.$$

Finally, each detection is represented directly in ENU as $\mathbf{x}^*$. Given an AIS target in ENU coordinates $\mathbf{a}$ at the same timestamp, matching is computed with Euclidean distance:

$$d(\mathbf{a}, \mathbf{x}^*) = \|\mathbf{a} - \mathbf{x}^*\|_2.$$

The detection with minimum d is selected, optionally under a distance gate.

Detection Training Results. Figures A.1 and A.2 report detailed loss and metrics graphs of the detection training procedure discussed in Chapter 2.

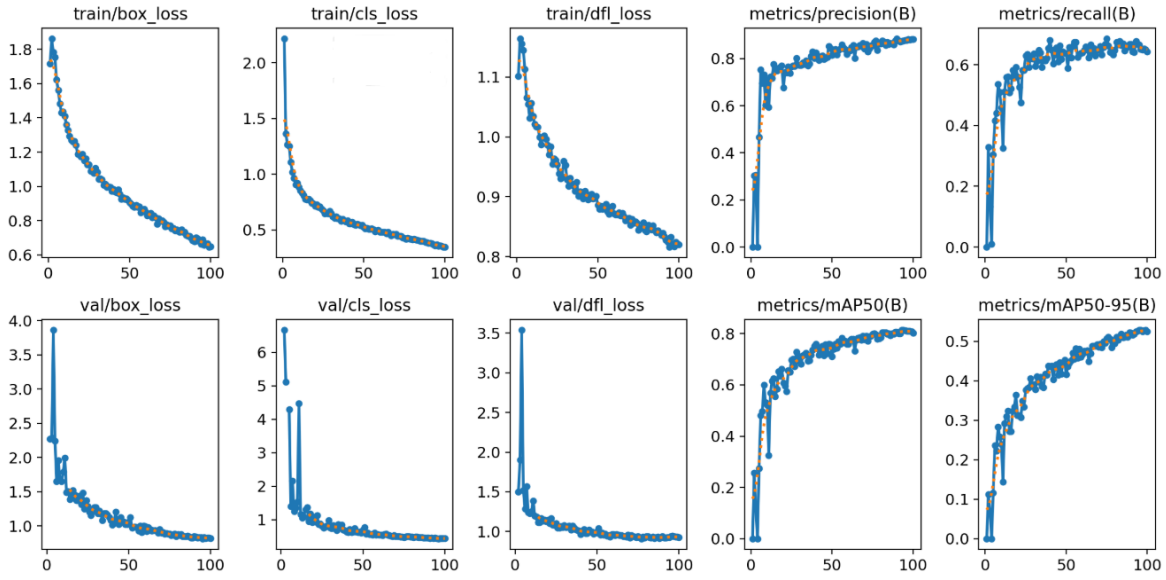


Figure A.1: Training and validation metric curves for the fine-tuned YOLOv8l model.

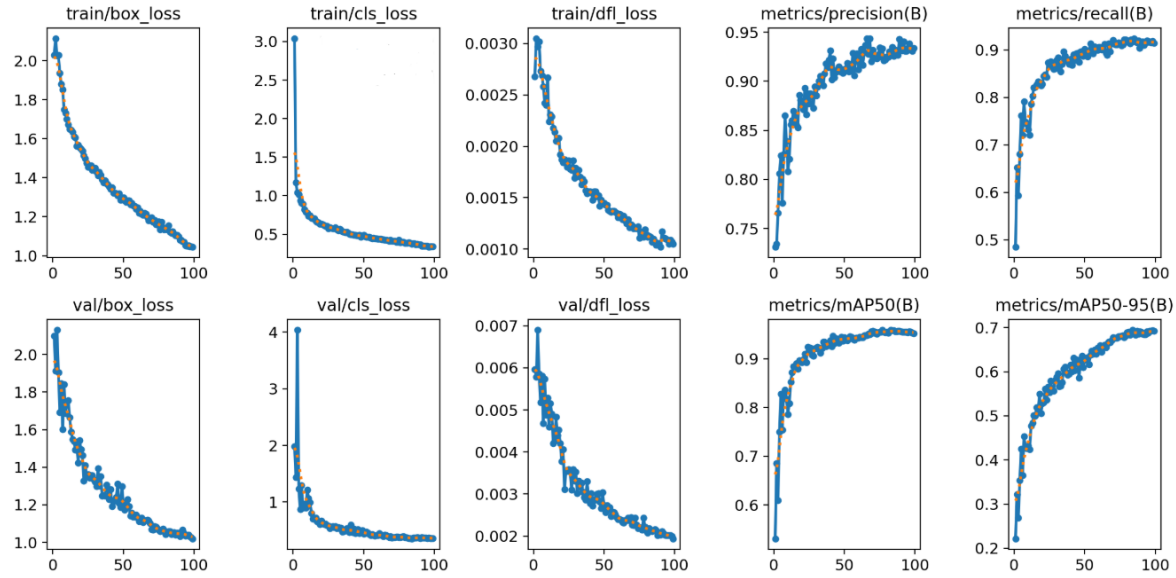


Figure A.2: Training and validation metric curves for the fine-tuned YOLO26l model.

Metric Learning Training Hyperparameters. Table A.1 reports the full training configuration used for the metric-learning pipeline summarized in Chapter 3.

Component	**Configuration**
Encoder	ResNet50 pretrained on ImageNet-1k
Projection head	Two-layer MLP with L2 normalization
Primary loss	ArcFace ($m = 0.5$, $s = 30.0$)
Optional contrastive setting	Supervised InfoNCE with temperature $\tau = 0.07$
Hard-negative mining	Top-k negatives, $k = 5$
Hard-negative weight	$\lambda = 0.1$
Optimizer	AdamW, learning rate 3×10^{-4}
Scheduler	Cosine annealing (`T_max`=100)
Epochs	150
Batch size	128 (train), 32 (validation)
Data split	Stratified 80/20 train/validation
Stabilization	Gradient clipping, max norm 1.0
Precision	Automatic mixed precision with gradient scaling
Augmentations	Color jitter, random crop, horizontal flip

Table A.1: Reference hyperparameter configuration for metric-learning training.

Acknowledgements

I would like to express my deepest gratitude to lawyer V. Laruffa of the Port Authority of the Strait of Messina for believing in this project and for making the data available.