Real-Time Maritime Object Detection in Search and Rescue Missions for On-Boat Camera Systems

Jonas Wortmann

Darmstadt University of Applied Sciences

h_da
HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES

Motivation

Thousands of people risk their lives annually crossing the Mediterranean Sea to escape war, persecution, and poverty, often using overcrowded and unsafe boats. Non-governmental Organizations (NGOs) make it their task to conduct search-and-rescue (SAR) operations to help people in distress. Traditionally, NGO crews manually scan the horizon with binoculars, a time-intensive and physically demanding task. However, video streams from cameras mounted on NGO vessels scanning the horizon present a valuable opportunity for automated analysis using computer vision. This work aims to develop a real-time computer vision-based solution to detect objects like people, boats, and other floating objects in video frames, enabling timely crew responses. The focus is on evaluating state-of-the-art (SOTA) object detection methods, comparing transformer-based and CNN-based models.

Dataset

The dataset, SAR-CAM, contains 5,632 high-resolution images including 4,151 annotated objects, captured during SAR missions by an NGO.

The following challenges are associated with this dataset:

- **Small Object Detection:** Boats on the horizon appear very small, covering only a few pixels, thus making them difficult to detect accurately.
- **Reflections:** The dataset contains reflections on the water surface and from the camera, which can create false positives and complicate the detection task.
- **Weather conditions:** The dataset includes images taken under various weather conditions, such as fog, overcast skies, and bright sunlight, which can affect visibility and detection performance.
- **Lens contamination:** The camera lens is contaminated by dirt, which creates additional noise and can lead to false detections.
- **Class imbalance:** The dataset exhibits a significant class imbalance, with a majority of the annotated objects being boats, while other relevant classes, such as people and floating objects, are underrepresented.

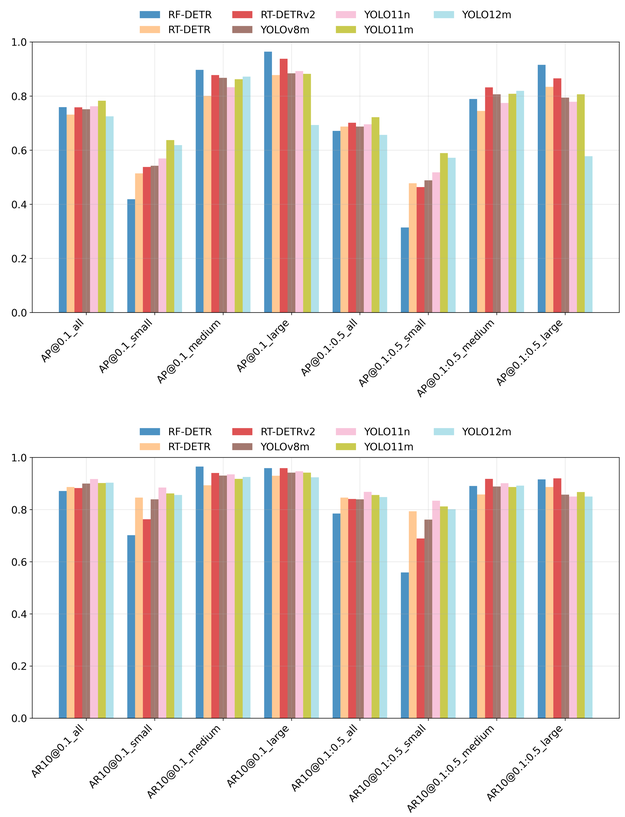
Finetuning SOTA object detection models

To address the challenges of the SAR-CAM dataset, various state-of-the-art (SOTA) object detection models are fine-tuned and evaluated. The model selection is based on the findings in [Ma+25] and [LHX25], whose results indicate the effectiveness of transformer-based object detection models in a related SAR scenario. The models include:

- **CNN-based:**
  - YOLOv8
  - YOLO11
  - YOLO12
- **Transformer-based:**
  - RT-DETR
  - RT-DETRv2
  - RF-DETR

The metrics used for evaluation include Average Precision (AP) and Average Recall (AR) at different Intersection over Union (IoU) thresholds, with a segmentation into small, medium, and large objects.

Finetuning Results



While YOLO11m performs best AP at small object detection, its lightweight model YOLO11n has the best small object AR. Among DETR variants, RT-DETR achieves the best AR rate for small objects, while maintaining comparable AP performance to RT-DETRv2. Transformer-based models RF-DETR and RT-DETRv2 perform better than CNN-based models in large object detection. In medium object detection, transformer-based and CNN-based models perform similarly. Notably, RF-DETR achieves a very high AR10@0.1 on medium objects of 0.97. Interestingly, YOLO12m performs worse in large object detection compared to other YOLO variants.

Small Object Detection Enhancement Methods

To further enhance the detection of small objects, various methods are applied:

Hyperparameter Tuning: Optimizing the model's hyperparameters, by searching for the best combination of training parameters, such as learning rate, batch size, and number of epochs, and data augmentation parameters, such as rotation, scaling, and flipping. The optimization algorithm is based on a Bayesian optimization method called Tree-Parzen Estimator [Wat23].

Knowledge Distillation: To enhance the feature extraction capabilities of the model, knowledge distillation is applied, where a smaller student model learns from a larger teacher model. In context of the work, the student models are the best performing CNN-based and transformer-based models. The recently released open foundation model DINOv3 [Sim+25] is used as the teacher model, which shows strong feature extraction capabilities on the introduced SAR-CAM dataset. The student models are trained unsupervised on the SAR-CAM dataset by minimizing the distance between the feature representations of the teacher and student models, which is expected to enhance the performance of the student models. The knowledge distilled student models are then fine-tuned on the SAR-CAM dataset.

COCO class mapping: YOLO and RT-DETR variants are pre-trained on the COCO dataset, which contains 80 classes. To leverage the pre-trained weights, a class mapping is applied, where the *boat* class in SAR-CAM is mapped to the *boat* class in COCO.

Adding a P2 Layer to YOLO11's architecture: Adding a P2 layer to YOLO11's architecture improved small object detection performance on the VisDrone dataset [Wan+25]. The P2 layer, which operates at a lower level of the network can better capture detailed features, which is important to detect small objects. Compared to other layers in the Feature Pyramid Network, the P2 layer acts on high-resolution feature maps from an earlier backbone stage and can therefore preserve more local details of the targets. Local details can be edges, textures and shapes, which are crucial for the detection of small objects.

Results

- RQ1: *Which model architecture, CNN-based or transformer-based, best meet the real-time operational requirements of SAR-CAM, in terms of inference speed, detection accuracy (low false-positive rate and high recall), and compatibility with the NVIDIA TensorRT ecosystem?* While CNN-based architectures achieve higher detection accuracy, particularly for small objects, transformer-based architectures significantly outperform CNNs in inference speed. Among the evaluated models, YOLO11m with FP16 quantization meets all system requirements while achieving the highest AP, making it the best-suited model for the SAR-CAM system.
- RQ2: *How can the detection performance of current state-of-the-art object detectors be improved for small objects in maritime environments by keeping the detection accuracy of medium and large objects stable?*

Model	Method	AP@0.1_small	AP@0.1:0.5_small	AR@0.1_small	AR10@0.1:0.5_small	AP@0.1:0.5_medium	AR@0.1:0.5_medium	AP@0.1:0.5_large	AR@0.1:0.5_large
YOLO11m	Baseline	0.64	**0.59**	0.86	0.81	0.81	0.89	0.81	0.87
	HPT	0.61	0.57	0.84	0.81	**0.85**	**0.91**	**0.87**	**0.92**
	KD	**0.65**	**0.59**	**0.88**	**0.82**	0.80	0.90	0.71	0.84
	class mapping	0.52	0.47	0.80	0.74	0.76	0.84	0.80	0.84
	P2 Layer	0.55	0.48	**0.88**	0.80	0.75	0.86	0.76	0.85
YOLO12m	Baseline	**0.62**	**0.57**	0.86	0.80	**0.82**	0.89	0.58	0.85
	HPT	0.59	0.54	**0.89**	**0.84**	0.81	**0.90**	**0.83**	**0.89**
RT-DETR	Baseline	0.51	0.48	0.85	0.79	0.74	0.86	**0.83**	**0.89**
	KD	0.43	0.40	0.78	0.70	0.73	0.84	0.77	0.85
	class mapping	**0.56**	**0.52**	**0.87**	**0.82**	**0.77**	**0.89**	0.82	**0.89**

Table: Small object detection performance and effect of small object detection enhancement methods. For medium and large objects, AP@0.1:0.5 and AR@0.1:0.5 are reported to verify stable performance.

No single method consistently improves small object detection. The impact of each method varied across different SOTA models. For a transformer-based architecture like RT-DETR, class mapping proved to be a simple yet effective way to enhance small object detection. With a longer training time, knowledge distillation could be a promising method for YOLO11m.

- RQ3: *How well do models fine-tuned on the SAR-CAM dataset generalize to benchmark maritime dataset SMD, and what does this imply about their domain adaptability?* While the general performance on medium and large objects is stable, the performance on small objects drops significantly. Similar to the observations on the SAR-CAM dataset, YOLO11m achieves the highest AP on small objects, while the lightweight YOLO11n achieves the highest AR on small objects. Among transformer-based models, RT-DETRv2 better generalizes to the small objects, while RF-DETR better generalizes to medium and large objects. Qualitative results suggest that detection errors due to reflections occur. Additionally, false positives are observed from detection on the sky, probably caused by clouds.

References

[LHX25] Jing Li, Yun Hua, and Mei Xue. "MSO-DETR: A Lightweight Detection Transformer Model for Small Object Detection in Maritime Search and Rescue". In: *Electronics* 14.12 (12 Jan. 2025), p. 2327. ISSN: 2079-9292. DOI: 10.3390/electronics14122327. URL: https://www.mdpi.com/2079-9292/14/12/2327 (visited on 06/21/2025).

[Ma+25] Shuai Ma et al. "OWRT-DETR: A Novel Real-Time Transformer Network for Small-Object Detection in Open-Water Search and Rescue From UAV Aerial Imagery". In: *IEEE Transactions on Geoscience and Remote Sensing* 63 (2025), pp. 1–13. ISSN: 1558-0644. DOI: 10.1109/TGRS.2025.3560928. URL: https://ieeexplore.ieee.org/abstract/document/10965796/figures (visited on 07/03/2025).

[Sim+25] Oriane Siméoni et al. *DINOv3*. Aug. 13, 2025. DOI: 10.48550/arXiv.2508.10104. arXiv: 2508.10104 [cs]. URL: http://arxiv.org/abs/2508.10104 (visited on 11/21/2025). Pre-published.

[Wan+25] Zhou Wang et al. "PC-YOLO11s: A Lightweight and Effective Feature Extraction Method for Small Target Image Detection". In: *Sensors* 25.2 (Jan. 9, 2025). ISSN: 1424-8220. DOI: 10.3390/s25020348. URL: https://www.mdpi.com/1424-8220/25/2/348 (visited on 01/02/2026).

[Wat23] Shuhei Watanabe. *Tree-Structured Parzen Estimator: Understanding Its Algorithm Components and Their Roles for Better Empirical Performance*. May 26, 2023. DOI: 10.48550/arXiv.2304.11127. arXiv: 2304.11127 [cs]. URL: http://arxiv.org/abs/2304.11127 (visited on 09/29/2025). Pre-published.