RampWatch: An In-the-Wild Dataset and Text-Guided Detection Framework for Recreational Vessels

Malik Muhammad Asim[1,3,*] Claire B. Smallwood[3] Abdullah Tariq[1]
Johnny Lo[1,2,†] Syed Zulqarnain Gilani[1,4,†]

[1] Centre of AI & ML (CAIML), School of Science, Edith Cowan University, Australia
[2] Mathematical Applications & Data Analytics (MADA) Research Group, School of Science, Edith Cowan University, WA, Australia.
[3]Aquatic Science and Assessment, Department of Primary Industries and Regional Development, Hillarys, WA, Australia
[4]Dept of Computer Sciences, The University of Western Australia, Australia

1 Supplementary Material

This section provides additional results, implementation details, and analysis to complement the main paper. The following appendices are included in this supplement:

- **Appendix I:** Qualitative Analysis of Detection Results
- **Appendix II:** Per-Class Performance comparison YOLO-TG with baseline
- **Appendix III:** Annotations
- **Appendix IV:** Controlled Experiment

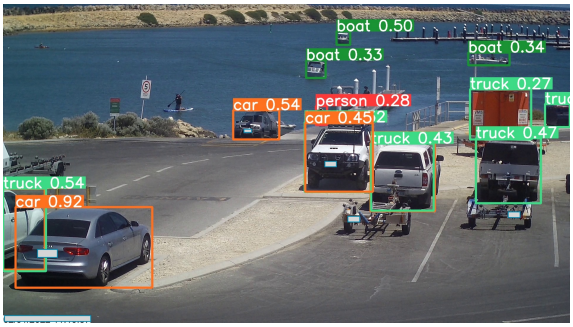
[*] *Corresponding author: m.asim@ecu.edu.au*
[†] *These authors contributed equally to this work.*

Appendix-I: Qualitative Analysis of Detection Results

In the initial configuration, the pre-trained YOLOv11 model was used without fine-tuning, limiting object detection to the MS COCO categories, which include generic classes such as person, car, and boat (Figure 1a). To overcome this constraint, the model was fine-tuned in the RampWatch data set, which includes diverse images of recreational vessels captured under varying conditions (Figure 1b). Fine-tuning led to improved detection performance, with boat instances more accurately recognized according to the RampWatch class taxonomy. However, some vessels remained undetected, particularly those absent from the label space of the dataset or heavily occluded.

To address these challenges, a Self-Attention (SA) module was incorporated, enhancing the model's ability to aggregate global contextual information (Figure 1c). This resulted in improved detection of vessel types in visually cluttered or dense scenes. However, the model still struggled with heavily occluded boats and categories not represented in the training data.

To further extend the model's capabilities, a text encoder was integrated to enable zero-shot detection via text prompts (Figure 1d). This module allows the model to identify previously unseen objects without requiring additional annotated data. For instance, providing the prompt "Surfer" enabled the model to detect a surfer and correctly assign it to the generalized "Others" category, despite "Surfer" not being part of the training labels. This demonstrates the capacity of the model for open-vocabulary recognition, improving its adaptability to real-world maritime environments.



(a) Pre-trained YOLOv11

(b) Fine-tuned (RampWatch) YOLOv11

(c) Feature Enhancement (YOLOv11+SA)

(d) YOLO-TG (Text guidance)

Figure 1: Qualitative comparisons across four configurations: (a) Pre-trained YOLOv11 baseline, (b) Fine-tuned (RampWatch) YOLOv11 (c) Feature Enhancement (YOLOv11+SA), (d) YOLO-TG using text guidance.

Appendix-II: Per-Class Performance comparison YOLO-TG with YOLOv11 (baseline)

The evaluation across several vessel classes among the RampWatch, SPSCD, and SMD datasets reveals consistent performance gains achieved by YOLO-TG over the baseline YOLOv11. These improvements are particularly pronounced in classes characterized by low inter-class variance and challenging visibility conditions.

RampWatch Dataset: In the RampWatch dataset, YOLO-TG consistently outperforms YOLOv11 across all metrics and classes (Table 1). In particular, YOLO-TG achieves the largest gains in recall

and mAP@50 for underrepresented and visually ambiguous categories such as JetSki, Kayak, and Rescue boats, where YOLOv11 struggles due to limited training examples and high intraclass variability. For example, the recall for the Rescue class increases from 0.51 to 0.65, the precision from 0.72 to 0.84, and mAP@50 improves by 23.6%. The dominant PowerBoat class also sees significant improvements, affirming YOLO-TG's robustness in both majority and minority classes. These gains suggest that YOLO-TG have better feature representations due to inclusion of the attention mechanism and text encoder. This improves the model's ability to recover low-resolution or partially visible objects, which are prevalent in the SMD dataset.

Table 1: Per-Class Performance comparison of YOLO-TG with YOLOv11 (baseline) on the RampWatch dataset.

Class	**Count**	**Precision (Pr)**		**Recall (Re)**		**mAP@50**		**mAP@50-95**	
		Y	**T**	**Y**	**T**	**Y**	**T**	**Y**	**T**
All	1075	0.84	0.89	0.57	0.72	0.63	0.71	0.38	0.44
Power Boat	856	0.84	0.91	0.68	0.75	0.71	0.77	0.39	0.48
JetSki	454	0.79	0.87	0.55	0.66	0.61	0.69	0.31	0.39
Yacht	376	0.76	0.86	0.61	0.69	0.67	0.73	0.37	0.47
Commercial	193	0.82	0.88	0.59	0.71	0.64	0.72	0.36	0.45
Kayak	270	0.77	0.85	0.42	0.56	0.51	0.63	0.25	0.35
Large Pleasure Boat	190	0.80	0.87	0.63	0.70	0.66	0.72	0.36	0.46
Rescue	89	0.72	0.84	0.51	0.65	0.55	0.68	0.29	0.41
Others	124	0.78	0.86	0.57	0.66	0.61	0.70	0.34	0.42

Note: Y = YOLOv11 (baseline), T = YOLO-TG.

SPSCD Dataset: On SPSCD, YOLO-TG again demonstrates superior detection capabilities over the YOLOv11 baseline (Table 2). The high precision and recall among both the YOLO-TG and baseline indicates that both models performed well on SPSCD due to the comparatively less challenging nature of the dataset. However, mAP@50 shows a consistent increase across all classes of SPSCD dataset.

Table 2: Per-Class Performance comparison of YOLO-TG with YOLOv11 (baseline) on the SPSCD dataset.

Class	**Count**	**Precision (Pr)**		**Recall (Re)**		**mAP@50**		**mAP@50-95**	
		Y	**T**	**Y**	**T**	**Y**	**T**	**Y**	**T**
All	2339	0.92	0.97	0.87	0.91	0.92	0.96	0.72	0.80
Small Craft	169	0.95	0.96	0.76	0.77	0.86	0.89	0.58	0.59
Small Fishing Boat	340	0.88	0.89	0.81	0.82	0.84	0.86	0.61	0.62
Small Passenger Boat	165	0.97	0.98	1.00	1.00	1.00	1.00	0.90	0.91
Fishing Trawler	71	0.97	0.98	0.86	0.87	0.95	0.97	0.75	0.76
Large Passenger Ship	308	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00
Sailing Boat	405	0.97	0.98	0.91	0.92	0.95	0.96	0.83	0.84
Motorboat	336	0.90	0.91	0.80	0.81	0.89	0.90	0.69	0.70
Pleasure Yacht	190	0.96	0.97	0.91	0.92	0.95	0.96	0.83	0.84
Medium Ferry	286	0.99	1.00	1.00	1.00	1.00	1.00	0.87	0.88
Large Ferry	108	1.00	1.00	1.00	1.00	1.00	1.00	0.87	0.88
High Speed Craft	157	0.97	0.98	0.99	1.00	0.99	1.00	0.89	0.90

Note: Y = YOLOv11 (baseline), T = YOLO-TG.

SMD Dataset: The results on SMD further highlight the advantage of YOLO-TG under challenging maritime conditions, including adverse weather and occlusion (Table 3). While YOLOv11 shows reasonable performance in prominent vessel classes, YOLO-TG notably improves recall and mAP@50, especially in smaller or less frequent classes, such as Kayak, Speed boat and FlyingBird/Plane.

Table 3: Per-Class Performance comparison of YOLO-TG with YOLOv11 (baseline) on the SMD dataset.

Class	Count	Precision (Pr)		Recall (Re)		mAP@50		mAP@50-95	
		Y	T	Y	T	Y	T	Y	T
All	1890	0.76	0.91	0.57	0.69	0.60	0.73	0.37	0.46
Ferry	478	0.79	0.92	0.44	0.73	0.53	0.78	0.32	0.53
Buoy	162	0.89	0.94	0.80	0.95	0.86	0.98	0.49	0.66
Vessel/Ship	1831	0.83	0.91	0.87	0.94	0.88	0.96	0.58	0.73
Speed Boat	465	0.78	0.88	0.41	0.68	0.48	0.74	0.22	0.45
Boat	124	0.74	0.89	0.36	0.40	0.39	0.56	0.25	0.30
Kayak	60	0.62	0.96	0.25	0.35	0.27	0.56	0.09	0.24
Sail Boat	183	0.96	1.00	0.90	1.00	0.97	1.00	0.75	0.84
Flying Bird/Plane	80	1.00	0.98	0.00	0.56	0.19	0.60	0.16	0.37
Other	818	0.68	0.82	0.41	0.80	0.53	0.82	0.22	0.40

Note: Y = YOLOv11 (baseline), T = YOLO-TG.

Appendix-III: Annotations

All images in the dataset were manually annotated using the Computer Vision Annotation Tool (CVAT). Each vessel was labeled with a tight bounding box. Annotations were exported in the YOLO format, which encodes each object as a single text line containing the class index and normalized values for the bounding box center coordinates x_center, y_center and dimensions $width, height$, relative to the image size. This format ensures compatibility with YOLO-based training pipelines and supports efficient real-time object detection (Figure 2).

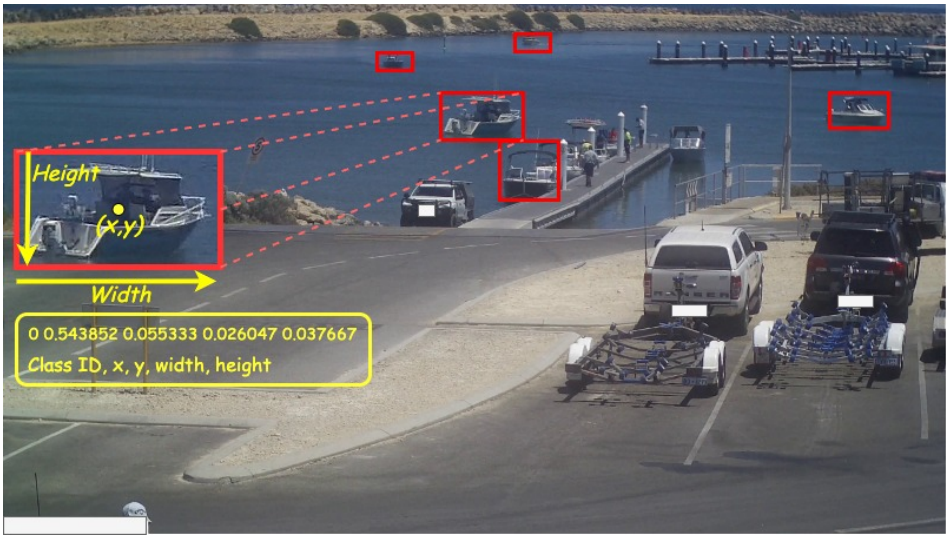


Figure 2: Example of annotated data using CVAT. Each vessel is enclosed by a bounding box and labeled with a corresponding class ID. The annotations are exported in YOLO format, which stores the class index followed by normalized values of the bounding box center coordinates (x, y) and dimensions (width, height), relative to the image dimensions. This format supports efficient loading in YOLO-based training pipelines.

Moreover, Figure 3 shows RampWatch dataset captures vessel activity under unconstrained real-world conditions, spanning both metropolitan and regional boat ramps. It encompasses diverse challenges such as varying lighting, weather changes, water-surface reflections, occlusion, and background clutter. This diversity ensures that the dataset reflects the true complexity of recreational vessel detection in coastal environments.

Figure 3: Representative samples from the dataset illustrate a range of challenging conditions, including variable lighting, diverse weather scenarios, and surface reflections, highlighting the diversity and complexity of the RampWatch dataset.

Table 4: Impact of prompt semantics on RampWatch. Using irrelevant prompts ("a bike", "a car") significantly degrades performance (nearly half the mAP compared to semantic vessel prompts), confirming that contextually relevant prompts are the primary driver of gains and irrelevant.

Prompt Type	Pr	Re	mAP@50	mAP@50–95	Δ@50 (%)	Δ@50–95 (%)
Irrelevant Prompt ("a bike", "a car")	0.67	0.40	0.411	0.223	–41.97	–49.32
Relevant Semantic Prompt (Ours)	**0.89**	**0.72**	**0.710**	**0.440**	**+14.52**	**+10.00**

Appendix-IV: Controlled Experiments

To examine which module contributes most to performance, we conducted controlled experiments by removing all additional blocks and equipping YOLOv11 only with the OVD module, followed by combining OVD with the SA module. Results in Table 5 indicate that the primary gains stem from semantic alignment. Moreover, to further validate this, we performed an experiment using irrelevant semantic prompts unrelated to recreational scenarios, which led to a drastic performance drop as shown in Table 4. The results demonstrate the importance of semantic alignment in guiding detection. When irrelevant prompts such as "a bike" or "a car" were used, the model's performance dropped sharply, with mAP@50 decreasing to 0.411 and mAP@50–95 to 0.223, corresponding to relative declines of –41.97% and –49.32%. In contrast, when relevant prompts aligned with recreational vessel semantics were provided, the model achieved substantially higher accuracy, with mAP@50 of 0.710 and mAP@50–95 of 0.440, yielding relative gains of +14.52% and +10.00%. These findings confirm that the observed improvements are primarily driven by semantic alignment rather than incidental architectural changes, and highlight the sensitivity of open-vocabulary detection to prompt relevance.

Table 5: Controlled Experiment results on the RampWatch dataset. Precision (Pr), Recall (Re), and mAP metrics are reported shows the most gain is due to OV module (semantic alignment). Δ values denote relative improvements over the YOLOv11 baseline.

Method	Pr	Re	mAP@50	mAP@50–95	Δ@50 (%)	Δ@50–95 (%)
YOLOv11 (baseline)	0.84	0.53	0.590	0.370	–	–
YOLOv11 (SA only)	0.85	0.57	0.599	0.383	+1.5	+11.62
YOLOv11 (OV only)	0.87	0.68	0.647	0.413	+9.66	+3.54
YOLOv11 (with SA & OV)	**0.89**	**0.72**	**0.710**	**0.440**	**+14.52**	**+10.00**