Automated Detection and Segmentation of Coastal Features from Remote Sensing Imagery Using Text-Prompted Zero-Shot Pipelines

Felipe Monsalvo[1,2], Harry Newton, Ph.D

[1] Center for Research in Computer Vision, University of Central Florida
[2] Department of Computer Science, Kean University

Abstract :

Submerged and semi-submerged maritime hazards such as rocks, shipwrecks, and reefs pose significant risks to navigation, particularly in dynamic coastal environments where visibility and tides rapidly change. This study develops and evaluates a modular computer vision pipeline that operates on high-resolution RGB satellite imagery, integrating Grounding DINO for text-prompted zero-shot open-vocabulary detection with Segment Anything Model 2 (SAM 2) for segmentation. Through prompt engineering, dynamic threshold tuning, and IoU-based filtering, Grounding DINO produced tightly aligned detections with high recall and strong spatial precision. These detections guided SAM 2 to generate polygon-ready masks, achieving high F1-scores and mIoU values across diverse coastal scenes. The pipeline demonstrated reliable geospatial alignment, with outputs exported to GIS-compatible formats for potential integration into charting and hazard-management systems. By focusing on modular detection and segmentation, this work highlights how combining language-guided detection with promptable segmentation yields interpretable, object-level results in challenging maritime imagery. These findings provide a foundation for future extensions, including multi-modal data integration for comprehensive coastal hazard assessment.

Index Terms: Open-vocabulary detection, zero-shot detection, text-prompted detection, instance segmentation, prompt engineering, remote sensing, Grounding DINO, Segment Anything Model 2 (SAM 2).

1 Introduction

Coastal environments are dynamic and often difficult to monitor at scale. Submerged and semi-submerged maritime hazards such as rocks, shipwrecks, and reefs pose significant risks to navigation and coastal infrastructure. These hazards are challenging to detect because they can appear visually distorted, blend with water patterns, or be partially obscured by turbidity, shadows, or surface reflections. Furthermore, seasonal changes and storm surges can cause significant shifts in the position of some features, particularly shipwrecks, and alter the depth of the surrounding seabed. Accurate, current charting of these features is critical for maritime safety, environmental assessment, and disaster response, where an undetected hazard can lead to severe operational and economic consequences.

Remote sensing has become a cornerstone of large-area maritime monitoring. In particular, ship detection in Synthetic Aperture Radar (SAR) imagery has received significant research attention, as SAR enables all-weather, day-night imaging. However, SAR returns are more responsive to metallic objects and engineered structures, leaving many natural or irregular hazards undetected. High-resolution RGB satellite imagery, by contrast, captures rich visual detail and texture, making it more suitable for identifying rocks, reefs, and other non-metallic features. Despite these advantages, there remains a lack of dedicated methods and benchmarks optimized for detecting and segmenting such hazards.

To address these challenges, this study develops and evaluates a modular detection–segmentation pipeline specifically designed for maritime hazard analysis. The pipeline integrates Grounding DINO, an open-vocabulary object detector, with Segment Anything Model 2 (SAM 2), a prompt-able segmentation model that generates precise pixel-level masks. This two-stage approach allows the detection and segmentation tasks to be tuned independently, enabling iterative refinement of prompts, thresholds, and filtering strategies for improved results.

To rigorously test the pipeline under real-world conditions, we created internal annotations on high-resolution RGB satellite imagery representing diverse coastal environments and hazard types. These internal datasets were sufficient to evaluate detection performance and segmentation fidelity across challenging scenarios. By leveraging prompt engineering, dynamic threshold selection, and input-to-output filtering, Grounding DINO produces bounding boxes that closely match ground-truth hazard extents. SAM 2 then refines these detections into polygon-ready masks, enabling integration with Geographic Information Systems (GIS) and operational mapping tools.

The results demonstrate that this modular pipeline can reliably identify and segment isolated hazards with strong spatial alignment, providing interpretable, object-level outputs. Beyond maritime safety applications, the flexibility of this approach offers a foundation for broader geospatial computer vision research. Future extensions may incorporate multi-modal data sources, including SAR, hyperspectral imagery, and historical nautical charts, to further enhance robustness and operational scope. By focusing on a detection-driven segmentation workflow and emphasizing interpretability and adaptability, this work contributes toward scalable, AI-enabled coastal monitoring in environments where reliable hazard identification is critical.

1.1 Contribution

Advancing Modular Vision Pipelines for Maritime Hazard Analysis

This work contributes a tested and adaptable pipeline architecture that others can adopt and extend for maritime and coastal computer vision. Instead of introducing another closed or task-specific model, we demonstrate how a detection-driven, promptable workflow can be engineered to handle irregular and low-contrast hazards with strong spatial accuracy. We formalize a process for linking open-vocabulary detection (GroundingDINO)

with instance-level segmentation (SAM 2) using prompt iteration, dynamic thresholds, and Intersection over Union (IoU)-based filtering. This process yields interpretable object-level outputs, rather than solely scene-level masks, enabling researchers to inspect intermediate stages and tune performance without retraining end-to-end. We provide evidence that this modular design delivers measurable gains in recall, precision, and mask fidelity on challenging coastal imagery. The pipeline's outputs are compatible with GIS environments, demonstrating a clear path toward operational deployment in chart updates or automated monitoring tools. By separating detection and segmentation components, the framework invites experimentation; different detectors, segmentation backbones, or prompt strategies can be swapped in and evaluated without rebuilding the entire system. This positions the pipeline as a foundation for future integrations and as a testbed for evolving open vocabulary or few-shot techniques in geospatial vision.

1.2 Paper Organization

The remainder of this paper is structured as follows. Section 2 reviews relevant literature on open-vocabulary detection, promptable segmentation, and maritime feature analysis. Section 3 outlines the proposed methodology, detailing the modular detection and segmentation pipeline. Section 4 presents experimental results and evaluations, including precision, recall, and segmentation metrics. Section 5 introduces an ablation study to assess the impact of specific prompt and filtering strategies. Section 6 discusses potential directions for extending the framework. Finally, Section 7 provides concluding remarks, followed by references.

2 Literature Review

This literature review examines current research in remote sensing imagery, segmentation strategies for coastal feature extraction, and the evolving role of vision-language models (VLMs) in this field. Special emphasis is placed on the emergence of modular, prompt-driven computer vision pipelines and their potential to generalize across diverse coastal environments. We highlight recent progress in open-vocabulary detection and promptable segmentation, framing them as key enablers for maritime hazard mapping. These developments inform our approach, which integrates text-conditioned detection with adaptable segmentation techniques to address the unique challenges of submerged and semi-submerged coastal hazards.

2.1 Remote Sensing Imagery and Challenges in Coastal Segmentation

Satellite imagery provides imagery provides high-resolution RGB data critical for monitoring coastal environments. However, its susceptibility to water reflectance, atmospheric distortion, and inconsistent lighting introduces substantial noise, especially when identifying features such as reefs, rocks, or shipwrecks that visually blend into marine backgrounds [5][6]. These challenges elevate the importance of semantic segmentation in coastal remote sensing, where traditional object detection often fails to capture irregular or amorphous targets. Although models like FDSESI apply sea-land masking to enhance ship classification [7], most current techniques are optimized for above-water features and lack robustness for partially submerged or camouflaged targets. The absence of reliable scene segmentation under such ambiguous conditions remains a pressing limitation in the field.

2.2 Training-Free and Open-Vocabulary Segmentation

Open-vocabulary segmentation models aim to generalize beyond fixed class labels using natural language prompts. For instance, SegEarth-OV achieves training-free segmentation by aligning image-text embeddings via CLIP features [3]. Its capacity to label both discrete objects and continuous coastal surfaces makes it promising for broad geospatial analysis, particularly in contexts where explicit class training is limited. Yet, despite its generalizability, the model often struggles with features lacking strong semantic priors, such as "semi-submerged reef clusters" or "subsurface rock ridges," leading to coarse boundaries or misclassifications. Upsampling and context-aware refinements (e.g., SimFeatUp) help mitigate these issues, but the underlying reliance on pretrained language-image representations can still hinder accuracy when applied to visually ambiguous maritime scenes.

2.3 Modular Pipelines and Prompt-Driven Detection

Modular computer vision pipelines, such as Grounding DINO paired with SAM 2, decompose the task into separate detection and segmentation phases, offering greater interpretability and control. Grounding DINO localizes visual concepts based on natural language phrases [9], while SAM 2 generates instance-level masks using flexible prompts like bounding boxes or points [10]. This setup is particularly well-suited to coastal imagery, where the variability of object appearance and vague semantic definitions complicate end-to-end segmentation. Text-prompted detection supports granular target identification using domain-relevant language (e.g., "faint reef under surface"), enabling the system to detect visually subtle features that might be missed by standard classifiers. This modularity also supports experimentation with prompt phrasing, confidence tuning, and mask refinement—elements that are harder to control in monolithic models.

2.4 Vision-Language Alignment and Domain Adaptation

Recent advancements in vision-language alignment suggest that flexible, prompt-compatible architectures could play a central role in open-world remote sensing. Singh et al. [12] demonstrate how Grounding DINO can be adapted with minimal supervision for agricultural domains, while Jose et al. [13] explore detection-segmentation fusion through shared embeddings in unified models like DINOv2. These efforts show that open-set generalization can be enhanced through multimodal learning, particularly when models reason jointly over visual and textual contexts. However, domain transfer remains a challenge—especially in environments like coastal zones, where target categories are visually ambiguous and context-dependent. Successfully adapting these models to the maritime domain requires more than prompt tuning; it involves careful architectural choices that preserve flexibility while addressing the nuanced demands of geospatial scene understanding.

3 Methodology

3.1 Modular Detection and Segmentation Approach

This study presents a two-stage vision pipeline for detecting submerged coastal hazards in satellite imagery. Text-prompted detection is paired with pixel-level segmentation to localize features like rocks, reefs, and shipwrecks. The modular design enables prompt tuning, interpretable outputs, and flexible component refinement. We evaluate the system's accuracy, generalization, and suitability for maritime monitoring.

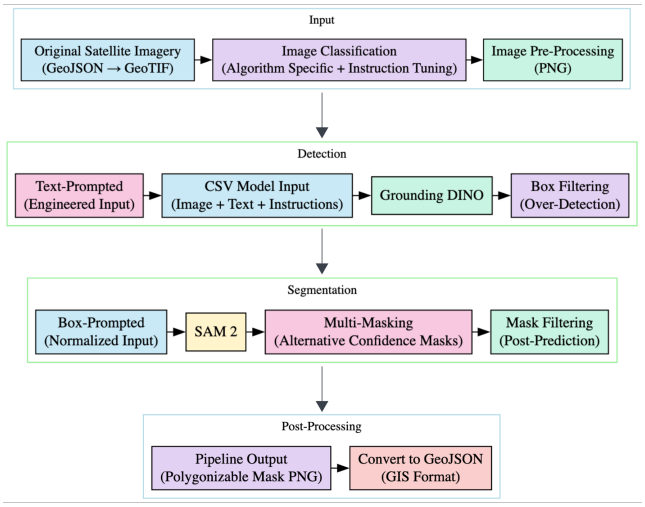


Figure 1. Modular pipeline for detecting and segmenting submerged coastal hazards in satellite imagery. It includes: (1) input preprocessing and classification (GeoTIFF to PNG); (2) detection via text-prompted Grounding DINO with filtering; (3) segmentation using SAM 2 with box prompts and multi-mask generation; and (4) post-processing with PNG-to-GeoTIFF export for GIS integration.

3.1.1 Text-Prompted Detection with Grounding DINO

We employ Grounding DINO, an open-set object detector pretrained on the Grounding-20M dataset containing over 20 million image–text pairs. This model integrates visual and language cues to perform open-vocabulary object detection. Additional detail is included in the *3.2 Experimental Setup* section below.

- **Prompt Engineering**: Text prompts such as "submerged shipwreck" or "isolated rock in water" are iteratively crafted and refined using maritime-specific vocabulary. This process enhances the model's ability to detect subtle or ambiguous features and allows for generalization to previously unseen object types within coastal imagery.
- **Dynamic Threshold Optimization**: Detection thresholds, such as box confidence scores and text relevance metrics, are dynamically tuned to balance precision and recall. Post-processing filters, including Intersection-over-Union (IoU) calculations, are applied to eliminate overlapping or low-confidence detections that could degrade segmentation performance.
- **Output Format**: The final output from the detection stage consists of bounding boxes linked to their originating text prompts. These spatially aligned boxes serve as structured inputs for the segmentation stage, enabling a distinct product to the next component of the pipeline.

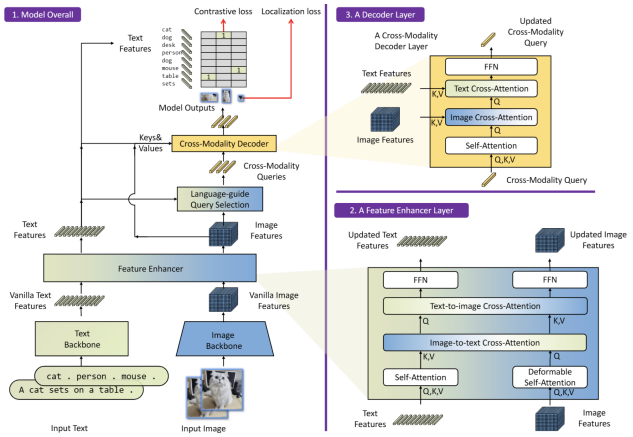


Figure 2. Grounding DINO architecture for text-prompted object detection. Text and image features are fused via a feature enhancer with bidirectional cross-attention, followed by language-guided query selection and cross-modality decoding. Outputs are optimized using contrastive and localization losses to align text prompts with visual regions.

3.1.2 Zero-Shot Instance Segmentation with SAM 2

The Segment Anything Model 2 (SAM 2) is employed to convert the detected bounding boxes into high-resolution, pixel-level segmentation masks that precisely outline the spatial extent of each identified maritime feature.

- **Prompt-Based Segmentation**: Filtered bounding boxes produced by Grounding DINO are passed directly to SAM 2 as spatial box prompts. This enables precise, instance-level segmentation of each detected region, while preserving spatial alignment with the original satellite input. The approach ensures that segmentation is both guided by prior detection and responsive to object-scale variability in coastal imagery.
- **Multi-Mask Selection**: For each box prompt, SAM 2 generates three segmentation masks corresponding to different confidence levels. A heuristic scoring method is used to evaluate mask quality based on factors such as edge coherence, region compactness, and alignment with the prompt. The highest-scoring mask is selected for downstream use, enhancing consistency and accuracy across diverse object shapes and textures.
- **Output Format**: Segmentation results are exported as PNG files for visualization and analysis. When geospatial precision is required, these masks are converted to GeoTIFF, preserving spatial metadata for seamless integration with GIS systems and marine navigation datasets.

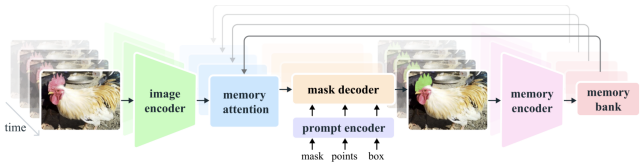


Figure 3. Architecture of Segment Anything Model 2 (SAM 2) for prompt-based segmentation. The model encodes input images and processes prompts (boxes, points, masks) through a dedicated prompt encoder. A memory module enables attention-guided refinement across frames or related inputs. The mask decoder generates high-resolution segmentation masks, leveraging both spatial context and temporal consistency.

3.2 Experimental Setup

3.2.1 Pipeline Overview Our workflow integrates Grounding DINO for text-prompted detection with SAM 2 for segmentation, as illustrated in the figure. The pipeline processes input imagery through preprocessing, detection, segmentation, and post-processing stages to produce polygon-ready masks.

3.2.2 Prompt Engineering for Detection Detection was guided by engineered text prompts using synonym stacking and descriptive modifiers. Phrases like *"submerged shipwreck," "partially submerged vessel," "isolated rock in water"* were iteratively refined to improve detection. We began with baseline terms, evaluated outputs, then added synonyms, spatial cues (e.g., "near shoreline"), and visual descriptors (e.g., "clustered," "shallow") to boost precision and recall across diverse scenes.

3.2.3 Detection Thresholds and Filtering GroundingDINO thresholds were tuned dynamically per image to balance recall and precision. We applied IoU-based filtering to remove overlapping or redundant detections, along with area constraints to exclude excessively large boxes.

3.2.4 Integration with SAM 2 Filtered bounding boxes were fed directly as box prompts into SAM 2. For each box, multi-mask output (3 masks per box) was enabled, and we selected the highest-confidence mask based on SAM 2's internal scoring. This approach produced accurate pixel-level segmentation aligned with the detected features.

4 Results and Discussion

4.1 Overview of Evaluation Setup

To evaluate RGB-based vision methods for maritime hazard mapping, we tested a two-stage pipeline using Grounding DINO for text-prompted detection and SAM 2 for box-guided segmentation. The image below shows the original satellite input, ground truth, and model predictions for both bounding boxes and masks, highlighting the system's ability to localize low-contrast hazards like rocks and wrecks.

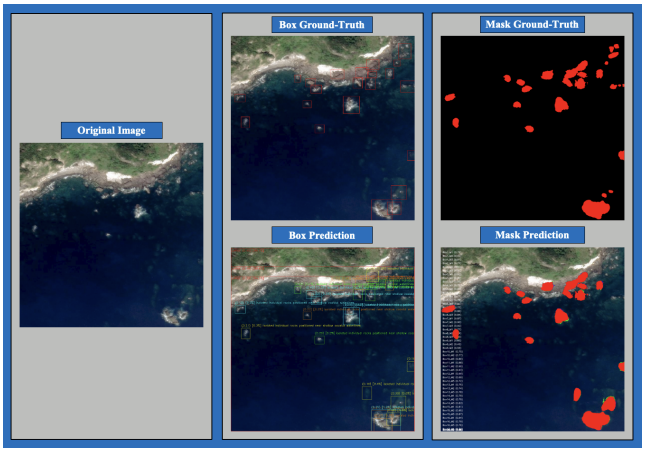


Figure 4. Detection and segmentation outputs compared to ground truth using Grounding DINO and SAM 2 on coastal imagery.

4.2 Detection Results (Grounding DINO)

Grounding DINO performed well in detecting submerged and irregular maritime hazards when guided by context-rich prompts. Phrases like "partially submerged shipwreck near shore" outperformed generic terms by adding spatial or visual cues (e.g., "gray," "near coastline"), boosting precision and mAP.

False positives were mainly triggered by visually similar non-target features, including underwater shadows, sandbars, and man-made structures like piers. To address this, thresholds were tuned: a box confidence of 0.18 and text relevance of 0.75 provided the best balance. IoU-based filtering and area constraints further reduced overlapping or oversized boxes.

These refinements led to high recall (92.23%) and strong spatial alignment (mIoU: 98.32%). Precision (76.61%) and F1-score (83.69%) remained moderate, reflecting the visual ambiguity of certain coastal features.

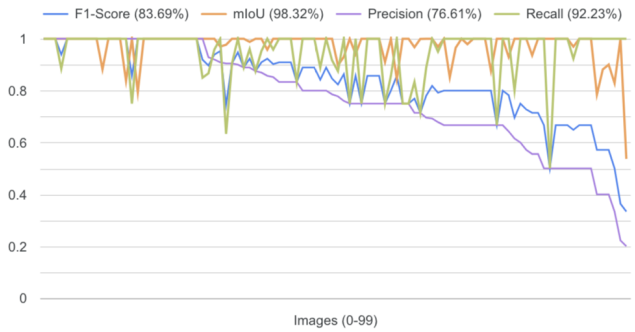


Figure 5. Per-image detection evaluation metrics for Grounding DINO across 100 satellite images.

4.3 Segmentation Results (SAM 2)

SAM 2 converted detection outputs into high-resolution, pixel-level segmentation masks. Each bounding box from Grounding DINO served as a prompt, with three candidate masks generated per region. Selecting the highest-confidence mask consistently produced strong segmentation performance.

On the test set, SAM 2 accurately captured the shapes of semi-submerged shipwrecks, boulders, and reef clusters, producing precise masks when boxes were well-aligned. However, segmentation quality declined when prompts were imprecise or incomplete. Overall, the model achieved high precision (92.04%), recall (87.52%), and strong boundary alignment, reflected in an mIoU of 84.48%, F1-score of 89.73%, and J&F-score of 87.11%.

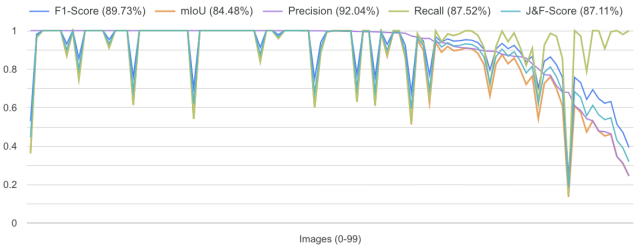


Figure 6. Per-image segmentation evaluation metrics for SAM 2 across 100 satellite images.

4.4 Confusion Matrix Analysis

To further assess model performance, confusion matrices were generated for both Grounding DINO (object-level detection) and SAM 2 (pixel-level segmentation). These matrices provide insight into class-specific behavior and the types of errors most commonly observed.

For Grounding DINO, the confusion matrix shows 629 true positives (72%) and 53 false negatives (6%), indicating that most annotated hazards were correctly detected. However, 192 false positives (22%) occurred, primarily due to visually similar but non-target features such as shadows, sandbars, or coastal infrastructure. True negatives are not represented, as evaluation was limited to annotated hazard regions.

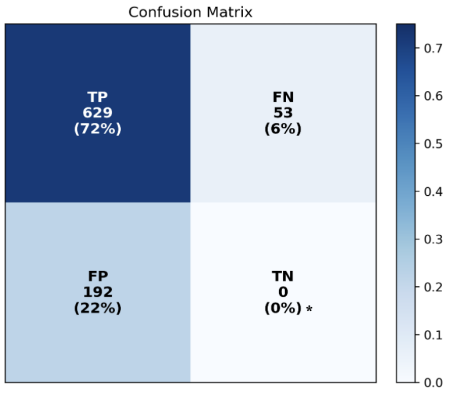


Figure 7. Grounding DINO Box Prediction Confusion Matrix

In SAM 2's pixel-level confusion matrix, over 8.8 million foreground pixels were correctly classified (81%), while false positives and false negatives accounted for 7% and 12% respectively. As with detection, true negatives were excluded, since evaluation focused solely on target-class pixels. The high true positive rate reflects SAM 2's ability to generate precise masks when provided with well-localized prompts, while the remaining error is largely attributed to imprecise box input or ambiguous object boundaries.

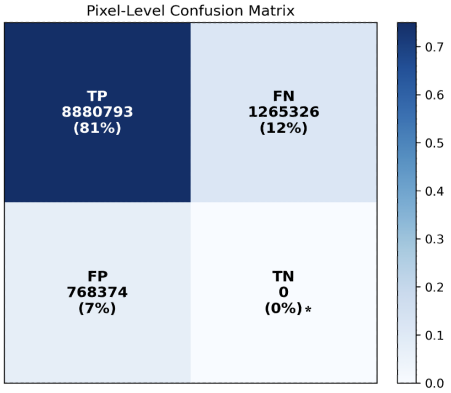


Figure 8. SAM 2 Mask Prediction Confusion Matrix

These matrices reinforce the importance of strong detection-stage outputs for accurate segmentation and highlight the need to mitigate common sources of false positives through prompt tuning and post-processing.

4.5 Geospatial Accuracy and GIS Integration

All segmentation results were exported to GeoTiff and overlaid on nautical chart layers using GIS software. The Grounding DINO + SAM 2 pipeline produced discrete, object-level outputs that aligned well with reference shoreline data, maintaining an average spatial error within a few meters. This level of accuracy is sufficient for vector-based chart updates, hazard databases, and further spatial analysis.

4.6 Runtime and Operational Feasibility

All experiments were conducted within a Docker container on WSL 2 (Ubuntu 20.04), using CUDA 11.8 and Python 3.10 to enable GPU-accelerated inference. Both Grounding DINO and SAM 2 were used with publicly available pretrained weights and required no additional training.

While the modular structure adds extra steps compared to a single end-to-end model, it enables flexible control over each stage. Users can refine prompts, adjust thresholds, and rerun detection or segmentation independently. This structure allows for better tuning of model behavior and easier debugging, especially when working with varied imagery or evolving hazard types.

4.7 Interpretation of Results and Trade-offs

The combined pipeline of Grounding DINO and SAM 2 is effective for tasks that require detecting and segmenting individual hazards with spatial clarity. It performs well when objects like isolated rocks, shipwrecks, or reef fragments must be separated and labeled independently. The use of prompts makes detection more interpretable, as each result is tied to a specific textual query.

The modular approach also allows tuning for different environments or use cases. However, performance depends on accurate bounding boxes. When target boundaries are unclear or when hazards appear as large continuous shapes, segmentation quality may drop. In such cases, more advanced segmentation tools or complementary methods might be needed to capture full spatial extent.

4.8 Limitations

While results are promising, several limitations remain:

- **Generalization to new terrains**: Performance tended to decline on satellite imagery from coastlines not represented in the test set, particularly those with different shoreline geometries, darker water reflectance, or unique environmental characteristics. This indicates a limitation in the pipeline's ability to generalize across geographies without region-specific prompt tuning or additional training data.
- **Class imbalance**: Small or rare hazards, such as underwater shadows, shallow algae blooms, isolated submerged rocks, and partially exposed shipwreck fragments, remain significantly underrepresented in the available data, reducing model robustness and generalization for these critical yet infrequent classes.
- **Modal exclusivity**: The current implementation relies entirely on RGB satellite imagery, which limits performance in turbid waters, cloudy conditions, or low-light scenarios. Future work will explore incorporating additional sensing modalities such as Synthetic Aperture Radar (SAR) or hyperspectral imagery to improve detection and segmentation under challenging environmental conditions.

5 Ablation Study

To evaluate the contribution of individual components in our RGB-based detection and segmentation pipeline, we conduct a series of ablation experiments. Each test isolates one factor model fine-tuning, prompt design, or post-processing using a fixed validation subset from the annotated dataset.

5.1 Detection Prompt Sensitivity (Grounding DINO)

Location-aware prompts, which included spatial or contextual details such as "near brownish coastline" or "on the right side of image," produced the highest mAP scores. These outperformed both generic prompts like "rock" and domain-specific phrases such as "isolated submerged rocks." The improvement highlights how adding visual or positional cues helps Grounding DINO better interpret complex coastal imagery and distinguish true hazards from background features.

Prompt Type	mAP%
Generic	74%
Domain-Specific	77%
Location-Aware	81%

5.2 IoU-Based Filtering and Threshold Tuning

Adjusting the box confidence threshold revealed that 0.18 provided the best balance between detecting true positives and minimizing false positives. Lower thresholds increased sensitivity but introduced more noise, while higher thresholds improved precision at the cost of missing smaller or less distinct hazards. The addition of IoU-based filtering further reduced overlapping detections, improving the reliability of outputs passed to the segmentation stage.

Box-Threshold	mAP%	False Positives
0.13 (Low)	79%	High
0.18 (Optimal)	77%	Moderate
0.23 (Strict)	71%	Low

5.3 Segmentation Prompting and Mask Selection (SAM 2)

For SAM 2, selecting the highest-confidence mask from the three candidates generated per prompt consistently yielded the best segmentation accuracy. This "Best-of-N" method led to higher mIoU scores than using just the first mask or averaging all three. It proved especially effective in cases with complex or overlapping objects, where more flexible mask evaluation helped capture better boundary definitions.

Mask Strategy	mIoU (%)
Top 1	78%
Avg Masks	82%
Best-of-N (N=3)	85%

6 Future Directions

6.1 Integration of Multi-Modal Remote Sensing for Enhanced Detection

Future work will explore the integration of additional satellite sensing modalities, particularly Synthetic Aperture Radar (SAR) and Hyperspectral Imaging (HSI), to enhance the robustness and reliability of the detection pipeline [5, 6, 8]. SAR offers all-weather, day-and-night imaging by penetrating cloud cover, smoke, fog, and other atmospheric obstructions that hinder electro-optical (EO) systems [5, 6]. With modern spatial resolutions, SAR systems can detect structures, vehicles, and terrain features, even through foliage or dry soil [6]. Such capabilities are particularly valuable in critical applications like coastal monitoring, flood assessment, and IUU (Illegal, Unreported, and Unregulated) fishing detection, where EO data alone may be insufficient [1]. Despite SAR's inherent challenges, such as speckle noise, signal distortion, and lack of color information, recent advances in AI, including deep learning-based classification and semantic segmentation, show promise in automating interpretation and improving usability [8]. Models like DeepSAR [1] and SegEarth-OV [3] demonstrate the feasibility of high-precision object recognition and land-sea segmentation in SAR data, enabling improved situational awareness in complex coastal environments. Incorporating HSI data, known for its rich spectral detail, can further aid in material classification and habitat identification, providing complementary insights alongside SAR and EO imagery [9]. These multi-modal synergies offer a path forward for building robust, generalizable detection pipelines that are resilient to weather, terrain, and sensor limitations.

6.2 Context-Aware Image Tipping and Modality Recommendation

Another future direction involves the implementation of an adaptive, geospatially-aware tipping system that prioritizes new satellite imagery acquisitions based on environmental conditions linked to previous successful detections [9]. This strategy would trigger tipping based on geo-coordinates and tide levels equal to or below those associated with prior positive identifications, increasing the likelihood of successful repeat detections. The Florida Seafloor Mapping Initiative (FSMI), which emphasizes precise bathymetric data collection for coastal planning and resilience, underscores the value of high-resolution environmental context in supporting such systems [2]. An intelligent tipping framework could also incorporate real-time parameters such as cloud cover, sea state, and terrain occlusion, to dynamically recommend the most suitable sensing modality [3, 5, 6]. For instance, SAR would be prioritized when cloud cover degrades EO imagery quality, leveraging its all-weather capability [5]. Techniques like global feature enhancement and sea-land segmentation, as demonstrated by enhanced feature and land awareness, highlight the importance of preprocessing and modality selection in mitigating false positives and optimizing detection accuracy [7]. This context-aware, sensor-adaptive approach aligns with our ongoing plan, which seeks to develop an agentic vision system capable of integrating detection, segmentation, and real-time data selection, ultimately advancing autonomous coastal monitoring and hazard detection from space-based assets [9, 10].

7 Conclusion

This paper presents a focused study on applying modern computer vision techniques to detect and segment submerged and semi submerged maritime hazards such as rocks, shipwrecks, and reefs using high resolution RGB satellite imagery. We evaluate a two stage pipeline that combines GroundingDINO for text prompted open vocabulary detection with SAM 2 for precise instance segmentation. Our findings show that the GroundingDINO and SAM 2 pipeline offers strong control and interpretability when discrete hazard instances must be localized and reviewed. This makes it well suited for vector based chart updates, targeted hazard warnings, and integration into GIS workflows. Prompt engineering and dynamic threshold tuning proved critical to achieving high precision and recall, while IoU based filtering improved box quality before segmentation. SAM 2's multi mask selection further refined outputs to produce accurate polygon ready masks. All outputs were exported as GeoJSON and PNG for seamless use in navigation and coastal monitoring tools, demonstrating the pipeline's potential to enhance maritime situational awareness and support safer vessel routing. Looking forward, this approach lays the groundwork for expanding to multi modal pipelines that fuse additional sensing modalities and provide even richer interpretations of coastal risk in real time.

Acknowledgements

This research was supported by the National Science Foundation (NSF) through the Research Experiences for Undergraduates (REU) program at the Center for Research in Computer Vision (CRCV), University of Central Florida.

References

[1] M. S. Pillai, A. Bhattacharya, T. Baweja, R. Gupta, and M. Shah, "DeepSAR: Vessel detection in SAR imagery with noisy labels," in Proc. IEEE Int. Conf. Image Process. (ICIP), 2024. https://github.com/manupillai308/DeepSAR

[2] R. Goosen and C. Hapke, "Florida Seafloor Mapping Initiative," Sea Technology, Jun. 2, 2025. https://sea-technology.com/florida-seafloor-mapping-initiative

[3] K. Li, R. Liu, X. Cao, X. Bai, F. Zhou, D. Meng, and W. Zhi, "SegEarth-OV: Towards training-free open-vocabulary segmentation for remote sensing images," Poster presented at IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR 2025), Jun. 2025. https://cvpr.thecvf.com/virtual/2025/poster/35106

[4] C. Cuttano, G. Trivigno, G. Rosi, C. Masone, and G. Averta, "SAMWISE: Infusing Wisdom in SAM2 for Text-Driven Video Segmentation," Poster presented at IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR 2025), ExHall D, Poster #308, Jun. 13, 2025. https://cvpr.thecvf.com/virtual/2025/poster/34510

[5] European Space Imaging. (2024, May 23). What Is SAR Imagery? Introduction To Synthetic Aperture Radar. https://www.euspaceimaging.com/blog/2024/04/05/what-is-sar-imagery/

[6] Maxar Technologies. SAR Imagery for Civil, Commercial & Defense. https://www.maxar.com/maxar-intelligence/products/sar-imagery

[7] Sensors Journal. Accurate Ship Detection Using Electro-Optical Image-Based Satellite on Enhanced Feature and Land Awareness. MDPI. https://www.mdpi.com/1424-8220/22/23/9491

[8] DeepBlock. (2023, December 30). How AI can help overcome SAR imagery analysis challenges. https://www.deepblock.net/blog/sar-imagery-analysis-challenges

[9] Xiao et al., "Florence-2: Advancing a unified representation for a variety of vision tasks," arXiv preprint, arXiv:2405.10300, 2024. https://arxiv.org/abs/2311.06242

[10] N. Ravi et al., "SAM 2: Segment anything in images and videos," arXiv preprint, arXiv:2408.00714, 2024. https://arxiv.org/abs/2408.00714

[11] S. Liu et al., "Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection," arXiv preprint, arXiv:2303.05499, 2023. https://doi.org/10.48550/arXiv.2303.05499

[12] R. Singh, R. Bidese Puhl, K. Dhakal, and S. Sornapudi, "Few-Shot Adaptation of Grounding DINO for Agricultural Domain," arXiv preprint, arXiv:2504.07252, Apr. 2025. https://arxiv.org/abs/2504.07252

[13] C. Jose et al., "DINOv2 Meets Text: A unified framework for image- and pixel-level vision-language alignment," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2025. https://cvpr.thecvf.com/virtual/2025/poster/33482

[14] T. Ren et al., "Grounding DINO 1.5: Advance the 'Edge' of open-set object detection," arXiv preprint, arXiv:2405.10300, 2024. https://doi.org/10.48550/arXiv.2405.10300