

Technical Assessment for Automated Ship Detection (California Coast)

Jorge Sánchez

Introduction

This report will cover first an exploratory data analysis, a brief discussion of a proposal automation ship detection pipeline, the limitations of the dataset and the proposed pipeline, and finally a recommendation to the Agency.

Exploratory Data Analysis (EDA) Summary

The initial dataset, consisting of 4000 satellite images crops (80x80 pixes), reveals significant limitations for robust model training. In Figure 1, we can see that the classes are inbalanced with 3000 images labelled as ‘no-ship’ and 1000 images labelled as ‘ship’.

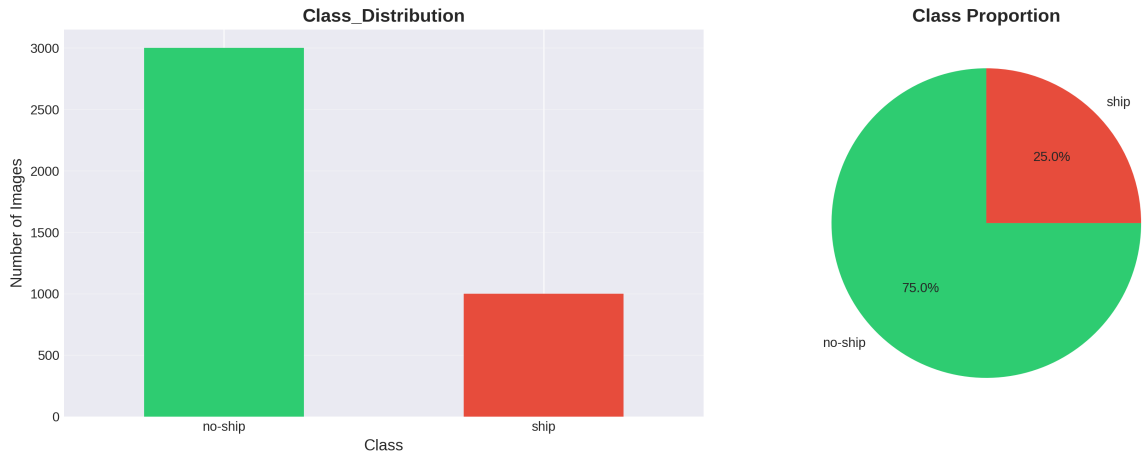


Figure 1: Class Distribution

Successfully loaded 4000 images from a total of 4000

Proposed Automation Pipeline: YOLO-Based Detection

To move towards automated counting, I proposed a state-of-the-art **You Only Look Once (YOLO)** object detection framework. This approach is superior to traditional methods like the Sliding Window technique for satellite imagery due to its speed and high recall.

Pipeline Outline - Data Preprocessing: Upscale the 1,000 'ship' cropped images from 80x80 pixels to standard 640x640 resolution. This is essential for modern Convolutional Neural Networks (CNNs). - Proxy Label Generation: Generate precise YOLO-format bounding boxes for the upscaled 'ship' images. This creates a high-quality positive training subset. - Model Selection & Training: Select a resource-efficient model like YOLOv5 for training. The model will be fine-tuned to classify and localize ships within the 640x640 positive subsets. - Inference: The trained YOLO model will then be applied directly to satellite scene images (not cropped 80x80 images) to provide real-time counts and geo-localization for all vessels within the scene.

Deployment Limitations & Critical Recommendations

Data Limitations The current dataset creates the major risks for deployment:

- Low Precision: The small sample size (1,000 positive images) increases the risk of over-fitting resulting in poor performance on new, unseen data.
- Generalization Failure: The lack of diverse environmental samples (varying weather, lightning sea state, and vessels sizes) will cause the model to perform poorly outside the specific conditions seen in the training data.
- Geo-Bias: The limited focus on two geographic zones prevents accurate monitoring of the broader coastline.

****Recommendation for Real-Time Monitoring*** To achieve scalable and reliable monitoring, I recommend a shift in data strategy: - Mandate Satellite Image Acquisition: Immediately procure a large, diverse dataset of full, high-resolution satellite scene images covering the entire California coastline. The model must be trained and tested on a target imagery (fully scenes), not just cropped tiles, as the Sliding Window method is inefficient for scaling. - Data Augmentation & Rebalancing: Implement advanced techniques (e.g MixUp, Mosaic) to artificially inflate and diversify the 'ship' class data - Continuous Learning Loop: Deploy the model into a monitoring environment where human-validated detections can be continuously fed back into the training dataset.