

Comparative Study of YOLOv10, YOLOv11 and YOLOv12 Lightweight Models for Multi-Class Maritime Search and Rescue Using UAV Imagery

Juliana Lyn Satore¹, Jazzie Jao^{1,2}, Red Castilla¹, Edgar Vallar¹, Maria Cecilia Galvez¹

¹ Department of Physics, De La Salle University Manila, Manila, Philippines - (juliana_satore, red_m.castilla, maria.cecilia.galvez, edgar.vallar)@dlsu.edu.ph

² Department of Software Technology, De La Salle University Manila, Manila, Philippines- jazzie.jao@dlsu.edu.ph

Keywords: Maritime object detection, UAV imagery, YOLOv10, YOLOv11, YOLOv12.

Abstract

The maritime Search and Rescue (SAR) operation requires effective and accurate object detection systems capable of identifying various targets in dynamic sea environments and low-light situations. The paper presents a comparative study of the YOLOv10, YOLOv11, and YOLOv12 networks in multi-class marine detection using UAV images. The SeaDronesSee Odv2 dataset has been preprocessed using physics-based augmentation that mimics environmental changes, such as fog, noon, sunset, dawn, and cloudy scenarios. A multi-resolution tiling procedure was implemented to preserve the image consistency of small objects. Results show that YOLOv11s is the model that has the least accuracy-efficiency trade-off, with an mAP@0.5 of 0.888 and an F1-Score of 0.872 at a reasonable inference time. Precision-recall analysis has shown that large maritime objects were detected with high precision, while small objects were detected with average recall. The results show that multi-resolution preprocessing, as well as physics-based augmentation, enhance the robustness and generalization of the network. Altogether, YOLOv11s is the most stable version to use in real-time maritime SAR missions with UAVs due to its ability to handle a variety of visual conditions.

1. Introduction

Unmanned Aerial Vehicles (UAVs) are becoming increasingly essential tools in maritime monitoring, and surveillance missions in search and rescue (SAR) due to their ability to efficiently cover broad oceanic areas and provide real-time situational awareness (Abdelnabi and Rabadi, 2024). Their deployment significantly reduces human risk and operational costs while enabling rapid scanning of remote and hazardous environments. However, reliable object localization in maritime scenes remains a significant challenge due to the vastness of the ocean and the constantly changing visual conditions (Er et al., 2023).

Detecting and localizing objects in the maritime areas continues to be a difficult task (Spraul et al., 2020). The ocean's surface is never static, constantly shifting under the influence of light, wind, and waves. Reflections, glare, and shadows distort visual patterns, while small objects such as humans appear blurry and easily lost against the moving water (Er et al., 2023). Additionally, the ocean's appearance varies significantly at different times of day and under varying weather conditions. Furthermore, the ocean appears dramatically different depending on the time of day or weather, ranging from bright at noon to golden at sunset, and gray under overcast skies, forcing detection systems to adapt to a broad spectrum of visual conditions (Appiah and Mensah, 2024).

The SeaDronesSee Odv2 dataset is one of the most diverse datasets of evidence for real-world challenges, serving as a benchmark for maritime human and object detection. It includes UAV recorded images at various altitudes, resolutions, and under different environmental conditions, providing both a challenge and an opportunity to develop models (Varga et al., 2022). High-resolution images contain very detailed information; however, they require a significant amount of computational resources. In contrast, lower-resolution images are much easier to process, yet they can miss tiny and important objects.

In pursuit of these dilemmas, this paper conducts a comparative analysis of YOLOv10, YOLOv11, and YOLOv12 architectures in multi-class maritime object detection using UAV images. A multi-resolution tiling pipeline is applied to retain the features of small targets. Simultaneously, a physics-based approach to data augmentation models the variety of realistic maritime scenarios, such as fog, dawn, noon, sunset, and cloudy scenarios, to improve the models' ability to generalize in response to natural visual changes.

2. Related Work

Maritime computer vision has been an incredibly active field of study due to its use in environmental surveillance, object recognition, and autonomous tasks. In the study of George et al. (2023) they trained a convolutional neural network using transfer learning on DenseNet, creating a multi-class model that detects ten marine objects, including fish, corals, submarines, and wrecks. They concluded that variability of underwater imagery is caused by light scattering, light absorption, and turbidity, demonstrating that model performance depends strongly on data preprocessing and augmentation strategies.

Balakrishnan et al. (2022) performed the comparative analysis of the various YOLO architectures (YOLOv3, YOLOv5, Tiny-YOLO, and Darknet-based versions) in real-time object detection. The results of their study showed that deeper backbones, such as Darknet-53, made higher precision, while lighter models provides faster inference. In terms of preprocessing, Çetin and Yıldız, (2022) separate data preprocessing into cleaning, transformation, and reduction, emphasizing that it enhances data quality and model reliability. They have noted that ensemble filtering, noise removal, and normalization directly impact the accuracy of learning algorithms, especially in high-variance data, such as in the case of maritime imagery. Equally, Smith et al. (2020) emphasized the application of tiling and subsampling as

preprocessing steps when processing huge and high-resolution images in deep learning pipelines. They have demonstrated that fine-grained features can be processed by cutting images into smaller and overlapping tiles, which the model processes without compromising contextual information or cutting objects at the boundary.

All these studies have highlighted the fact that both architectural creativity and customized preprocessing can have a positive influence on contemporary object detection in the maritime environment. Based on these discoveries, the current paper compares the behavior of YOLOv10, YOLOv11, and YOLOv12 on a multi-resolution UAV imagery pipeline and physics-based augmentations to evaluate their ability to withstand various environmental factors.

3. Materials and Methods

3.1 Dataset and Multi-Resolution Preprocessing

The dataset used in this study was the SeaDronesSee Odv2 dataset, which contains various maritime UAV images with resolutions ranging from 1225 x 926 to 5456 x 3632 pixels in size (Varga et al., 2022). The data has several classes of objects used in both maritime surveillance and search and rescue (SAR) missions. The images were divided into resolution-scaling grids, which had an overlap ratio of 30 percent between adjacent tiles to minimize the number of boundary cuts and preserve context. COCO annotations were translated into tile coordinates to align them, followed by conversion into the YOLO format. All tiles were scaled to 640x640, and adaptive padding was introduced to preserve the objects' original shape and prevent distortion during input normalization. A background filtering procedure was also used to regulate the representation of classes within the dataset. In every original image, there was a maximum of one non-class (background-only) tile remaining. This was not done to directly address inter-class imbalance, but rather to ensure that the dataset was not filled with empty samples. In this way, enough contextual background information was preserved, and at the same time, model learning became more efficient and less biased by non-informative areas.

3.2 Data Augmentation

To enhance the model's generalization and robustness, a physics based augmentation pipeline was employed to model the environmental uncertainty associated with maritime UAV missions. The augmentations were used to simulate a real-world environment by creating effects of lighting and visibility at different times of day, such as sunset, dawn, noon, fog, and cloudy. The photometric augmentation was supplemented with rotations and flips to prevent the model from overfitting. The implementation of these augmentation schemes on all splits was done to ensure that the object and background regions were exposed to different conditions.

3.3 Experimental Platform

All experiments were conducted on the Lenovo Legion R9000 laptop, equipped with a GeForce RTX 4060 graphics card (8 GB VRAM) and an AMD Ryzen 7000 series processor. The models have been implemented using the Ultralytics framework (CUDA-accelerated) in PyTorch and run within an Anaconda environment, ensuring reproducibility of the models. A batch size of 16 was used in 100 epochs of training.

3.4 Evaluation Metrics

In order to have a thorough assessment of both detection performance and computational efficiency, the following measures were used:

- **Precision (P)**, as illustrated in Equation 1, is the number of human targets that are correctly identified (True Positives, TP) divided by the number of predicted targets, that is, correctly and incorrectly identified (False Positives, FP).

$$P = \frac{TP}{TP + FP} \quad (1)$$

- **Recall (R)**, as shown in Equation 2, is a ratio of the number of correctly identified human targets (TP) to the number of ground-truth (including missed detections, False Negatives, FN).

$$R = \frac{TP}{TP + FN} \quad (2)$$

- **F1-Score (F1)**, is a harmonic mean of Precision and Recall that balances the accuracy and completeness of detection.

$$F1 = \frac{2 \cdot P \cdot R}{P + R} \quad (3)$$

- **Mean Average Precision (mAP@0.50)**, presented in Equation 4, is the average of the values of Average Precision (AP) of the classes at an IoU threshold of 0.5.

$$mAP@0.50 = \frac{1}{C} \sum_{c=1}^C AP_c \quad (4)$$

Parameters (M) refer to the number of learnable weights in the network, indicating the model's complexity. Contrary to the **Model size (mb)**, the area of the weights is where the trained weights are stored for deployment. In the meantime, **FLOPs (G)**, which represent the count of floating-point operations required per forward pass, are used to measure computational cost and efficiency. Finally, **Inference (ms)** is the average time required to execute one image through the model, a convenient measure of real-time applicability which is critical to UAV-based SAR missions.

3.5 Model Architectures

This paper will compare three state-of-the-art YOLO architectures, including: YOLOv10, YOLOv11, and YOLOv12, using the nano (n) and small (s) versions. Each model was trained under the same conditions to enable a fair and objective assessment of its detection and computational performance.

• YOLOv10

The YOLOv10 architecture is built upon the continuous evolution of the YOLO family. One of the significant new features of the architecture is a decoupled detection head, in which classification and localization are computed using distinct computational pathways. This separation stabilizes the optimization process, leading to more precise feature learning during training (Hussain, 2024, Alif and Hussain, 2024). In addition, YOLOv10 employs an anchor-free detection mechanism, directly predicting object centers and bounding box dimensions, thereby eliminating the

Model	Variant	Params (M)	FLOPs (G)	Inference (ms)	Precision	Recall	F1
YOLOv10	n	2.71	4.18	3.14	0.872	0.795	0.832
	s	8.07	12.36	6.16	0.895	0.817	0.854
YOLOv11	n	2.59	3.21	2.55	0.889	0.840	0.864
	s	9.52	11.03	6.37	0.902	0.845	0.872
YOLOv12	n	2.56	3.22	4.04	0.900	0.832	0.865
	s	9.26	10.84	9.13	0.896	0.850	0.872

Table 1. Overall Performance Comparison of YOLO model variants.

els demonstrated good detection performance across all classes of maritime objects, although their behavior was found to be variable, depending on model complexity. YOLOv11s achieved the highest precision (0.902) and F1-score (0.872), indicating that the model has both high detection accuracy and good localization capabilities. The YOLOv12s model achieves a similar F1-score (0.872) at the cost of a higher inference time (9.13 ms), indicating that it is slightly less efficient when running in real-time. On the contrary, YOLOv12n, which had the lowest number of parameters (2.56 M), contributed to the best recall (0.832) and F1-score (0.865) with moderate inference speed (4.04 ms), which was highly effective despite a lightweight architecture. In general, the small variants revealed a consistent advantage over the nano models in terms of precision, recall, and F1-score, emphasizing the fact that all model sizes trade off detection accuracy and computational efficiency.

4.2 Class-wise Detection Performance

The heatmap of the results, in terms of class-wise mAP (at 0.5), shows that the six variants of YOLO differ in overall detection accuracy by a large margin, with the differences in their performance primarily due to the object types themselves, as shown in Figure 4. The easiest classes to differentiate were the Buoy and Jet Ski, which provided mAP values of more than 0.94 in all models. This high-performance ceiling, where the Buoy category reached its peak at 0.961 in YOLOv11n and YOLOv12s, implies that their size and high visual contrast enabled them to be easily distinguishable even in the varied physics augmentation to which the dataset was subjected. On the other hand, other more challenging classes included Life-Saving Appliance and Swimmer. The Swimmer class was the most challenging, with the lowest scores, ranging from 0.754 (YOLOv10n) to 0.788 (YOLOv11s), due to its size and low contrast, as well as the effects of physics augmentations on feature definitions.

Comparing the models directly to each other, YOLOv11s is the most balanced and stable, performing as it scores the highest in both challenging classes. This high performance during challenging situations confirms its increased ability to generalize to an extensive scale of objects in terms of size and contrasts. The newer YOLOv12n was also the most competitive in the lightweight category, as demonstrated by the Jet Ski targets. The YOLOv10 variants typically made up the lower limit in performance, with the YOLOv10n variant recording the lowest score in the two most challenging classes.

4.3 Computational Analysis

Figure 5 illustrates the trade-offs in performance between computational and detection accuracy of the YOLOv10, YOLOv11, and YOLOv12 architectures, as measured by both Model Size (MB) and mAP at 0.5. There is an apparent pattern: precision

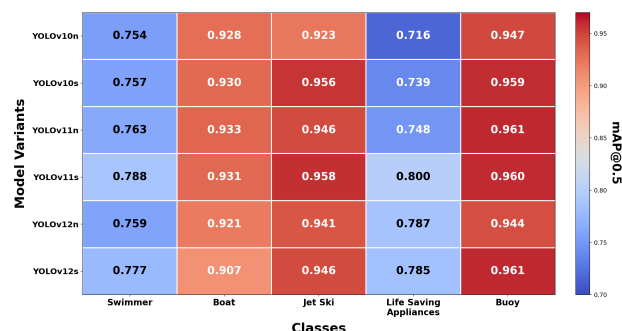


Figure 4. Performance Comparison Across Classes.

tends to increase with model complexity, and the small variants always have higher mAP at 0.5 than the corresponding nano ones.

The model with the highest overall mAP at 0.5 (0.888) is the YOLOv11s model, which has the smallest size (18.52 MB), and is therefore the most accurate model variant. On the other hand, the smallest YOLOv10n model achieved the lowest accuracy of 0.854 (5.49 MB). In the case of lightweight implementation, YOLOv12n was more efficient, achieving 0.870 mAP at 0.5, with a model size of 5.27 MB, and reaching the same accuracy as YOLOv11n (0.870 mAP at 0.5, 5.23 MB). This comparison validates that YOLOv11s is the most appropriate option when the application has high accuracy requirements, regardless of size. However, when resource constraints are limiting, YOLOv12n is the most suitable option due to its moderate accuracy in detection.

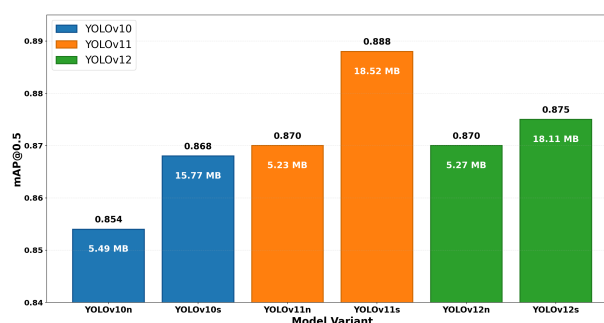


Figure 5. Overall Performance Detection.

4.4 Accuracy–Speed Trade-off Analysis

Figure 6 shows a trade-off between the Inference Time (ms) and Detection Accuracy (mAP at 0.5) of the YOLOv10, YOLOv11, and YOLOv12 model variants. In line with the concept of model complexity, detection accuracy tends to improve with longer inference times, which correlates with the complexity of

the computing network architecture. YOLOv11s achieves the highest accuracy with the highest mAP at 0.5, which is 0.888, along with an inference time of approximately 6.4 ms. This is where high accuracy is found at the extreme of the performance spectrum. On the other hand, the fastest model is YOLOv11n, which processes an image in about 2.5 ms and still achieves a competitive mAP of 0.5, at 0.870. This makes YOLOv11n the best option in applications that demand real-time implementation with a high level of latency consideration.

The nano versions (YOLOv11n, YOLOv12n) are instrumental on limited platforms, such as maritime UAVs. Both models achieve the same level of competitive accuracy (0.870). However, at the accuracy level, YOLOv11n is the most efficient option, as it is faster at 2.5 ms compared to YOLOv12n at 4.0 ms. Moreover, the YOLOv10 series is evidently inefficient in this measure; YOLOv10n is slower than YOLOv11n (3.1 ms vs. 2.5 ms) and has a lower accuracy (0.854). In general, the YOLOv11 models establish the performance boundary between the tested variants, with the fastest high-accuracy speed (YOLOv11n) and the highest average accuracy (YOLOv11s).

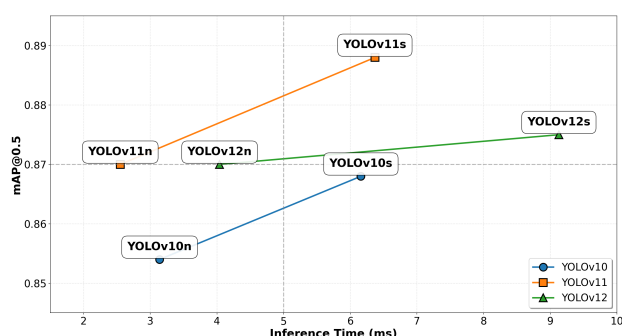


Figure 6. Speed-Accuracy Trade-off: Inference Time vs Overall Detection Performance.

4.5 Precision–Recall Curve Analysis

Figure 7 shows the Precision-Recall (PR) curves, which are used to measure the detection ability of the YOLOv11s model in all classes of maritime objects. This model has high overall reliability with a cumulative mAP@0.5 of 0.887. The curves indicate a substantial difference in performance regarding object size and visibility. The model achieves maximum precision and recall with Buoy (0.959), Jet Ski (0.958), and Boat (0.928), as they have prominent shapes that enable the model to have high confidence. On the other hand, the curves of Swimmer (0.791) and Life-Saving Appliances (0.800) indicate a sharp decline in accuracy with an increase in recall. This trend highlights the inherent difficulty in accurately identifying these smaller, visually subtle objects, which are often obscured by water or waves.

Even in the case of challenging targets, the effectiveness of physics-based augmentation is proven by the stability of these curves. This strategy was effective in this context because it could model changes in lighting (dawn, sunset, noon) and visibility (fog, cloudy weather), thereby reducing overfitting and improving model robustness, allowing YOLOv11s to perform well in a range of realistic maritime scenarios. This discussion supports the balanced trade-off of the model between precision and recall, and it is effective both for large and small objects, as well as for small-scale detection tasks with higher demands.

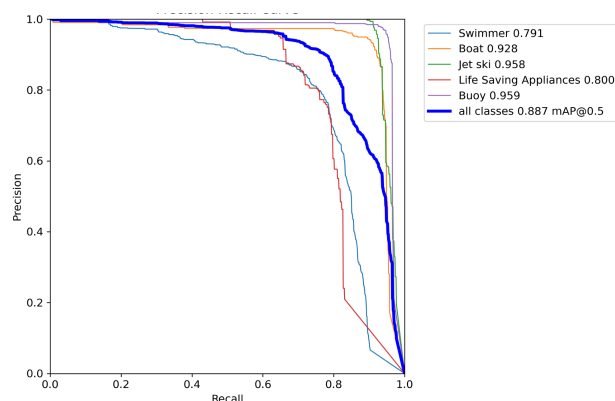


Figure 7. Precision-Recall curves for each object class and overall performance.

5. Conclusion

The paper presents a comparative analysis of the YOLOv10, YOLOv11, and YOLOv12 architectures for multi-class maritime object detection using UAV imagery. A multi-resolution tiling pipeline and physics-based augmentation were suggested to improve resilience to illumination and sea variation variability. The highest general balance of accuracy and efficiency was observed in YOLOv11s, with the highest mAP at 0.5 (0.888) and F1-score (0.872), and a moderate inference time. The Swimmer class was the most challenging to classify because it was small in scale, had low contrast, and was obscured by waves; however, the physically inspired augmentations significantly enhanced the consistency of detection in complex lighting conditions. The results demonstrate that YOLOv11, combined with a multi-resolution and physics-aware training pipeline, is a valuable framework for UAV-based maritime search and rescue operations.

It can be further extended to include attention or transformer-aided models to enhance the localization of small objects and provide more accurate detections, even in unfavorable weather conditions or when the object itself is in motion.

References

- Abdelnabi, A. A. B., Rabadi, G., 2024. Human detection from unmanned aerial vehicles' images for search and rescue missions: a state-of-the-art review. *IEEE Access*.
- Alashjaee, A. M., AlEisa, H. N., Darem, A. A., Marzouk, R., 2025. A hybrid object detection approach for visually impaired persons using pigeon-inspired optimization and deep learning models. *Scientific Reports*, 15(1), 9688.
- Alif, M. A. R., Hussain, M., 2024. YOLOv1 to YOLOv10: A comprehensive review of YOLO variants and their application in the agricultural domain. *arXiv preprint arXiv:2406.10139*.
- Apostolidis, K. D., Papakostas, G. A., 2025. Delving into YOLO Object Detection Models: Insights into Adversarial Robustness. *Electronics*, 14(8), 1624.
- Appiah, E. O., Mensah, S., 2024. Object detection in adverse weather condition for autonomous vehicles. *Multimedia Tools and Applications*, 83(9), 28235–28261.

Balakrishnan, B., Chelliah, R., Venkatesan, M., Sah, C., 2022. Comparative study on various architectures of yolo models used in object recognition. *2022 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, IEEE, 685–690.

Çetin, V., Yıldız, O., 2022. A comprehensive review on data preprocessing techniques in data analysis. *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi*, 28(2), 299–312.

Er, M. J., Chen, J., Zhang, Y., Gao, W., 2023. Research challenges, recent advances, and popular datasets in deep learning-based underwater marine object detection: A review. *Sensors*, 23(4), 1990.

George, G. et al., 2023. Performance evaluation of a multi-class classification model for marine objects. *2023 12th International Conference on Advanced Computing (ICoAC)*, IEEE, 1–8.

Hidayatullah, P., Syakrani, N., Sholahuddin, M. R., Gelar, T., Tubagus, R., 2025. YOLOv8 to YOLO11: A comprehensive architecture in-depth comparative review. *arXiv preprint arXiv:2501.13400*.

Hussain, M., 2024. Yolov5, yolov8 and yolov10: The go-to detectors for real-time vision. *arXiv preprint arXiv:2407.02988*.

Jegham, N., Koh, C. Y., Abdelatti, M., Hendawi, A., 2024. Yolo evolution: A comprehensive benchmark and architectural review of yolov12, yolo11, and their previous versions. *arXiv preprint arXiv:2411.00201*.

Ji, Y., Ma, T., Shen, H., Feng, H., Zhang, Z., Li, D., He, Y., 2025. Transmission Line Defect Detection Algorithm Based on Improved YOLOv12. *Electronics*, 14(12), 2432.

Khanam, R., Hussain, M., 2025. A Review of YOLOv12: Attention-Based Enhancements vs. Previous Versions. *arXiv preprint arXiv:2504.11995*.

Li, M., Yan, N., 2025. IPD-YOLO: Person detection in infrared images from UAV perspective based on improved YOLO11. *Digital Signal Processing*, 105469.

Rao, S. N., 2024. YOLOv11 architecture explained: next-level object detection with enhanced speed and accuracy. *Medium*. Available online: <https://medium.com/@nikhil-rao-20/yolov11-explained-next-level-object-detection-with-enhanced-speedand-accuracy-2dbe2d376f71> (accessed on 25 February 2025).

Sapkota, R., Meng, Z., Churuvija, M., Du, X., Ma, Z., Karkee, M., 2024. Comprehensive performance evaluation of yolov12, yolo11, yolov10, yolov9 and yolov8 on detecting and counting fruitlet in complex orchard environments. *arXiv preprint arXiv:2407.12040*.

Smith, B., Hermesen, M., Lesser, E., Ravichandar, D., Kremers, W., 2021. Developing image analysis pipelines of whole-slide images: Pre-and post-processing. *Journal of Clinical and Translational Science*, 5(1), e38.

Spraul, R., Sommer, L., Schumann, A., 2020. A comprehensive analysis of modern object detection methods for maritime vessel detection. *Artificial intelligence and machine learning in defense applications II*, 11543, SPIE, 13–24.

Varga, L. A., Kiefer, B., Messmer, M., Zell, A., 2022. Seadronesee: A maritime benchmark for detecting humans in open water. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2260–2270.