

Obstacle Detection and Ship Recognition System for Unmanned Surface Vehicles

Sevda Sayan^{1,2} ^a and Hazım Kemal Ekenel^{2,3} ^b

¹ASELSAN, Defense System Technologies, Turkey

²Istanbul Technical University, Department of Computer Engineering, Turkey

³New York University Abu Dhabi, Division of Engineering, U.A.E.

Keywords: Obstacle Detection, Ship Classification, Vision Transformers, Maritime.

Abstract: This study investigates obstacle detection and ship classification via cameras to ensure safe navigation for Unmanned Surface Vehicles. A two-stage approach was employed to achieve these goals. In the first stage, the focus was on detecting ships, humans, and other obstacles in maritime environments. Models based on the You Only Look Once architecture, specifically YOLOv5 and its variant TPH-YOLOv5 —specialized for detecting small objects— were optimized using the MODS dataset. This dataset contains labeled images of dynamic obstacles, such as ships, humans, and static obstacles, e.g., buoys. TPH-YOLOv5 performed well in detecting small objects, crucial for collision avoidance in Unmanned Surface Vehicles. In the second stage, the study addressed the ship classification problem, using the MARVEL dataset, which contains over two million images across 26 ship subtypes. A comparative analysis was conducted between Convolutional Neural Networks and Vision Transformer based models. Among these, the Data-efficient Image Transformer achieved the highest classification accuracy of 92.87%, surpassing the previously reported state-of-the-art performance. In order to further analyze the classification results, this study introduced a generic method for generating attention heatmaps in vision transformer based models. Unlike related works, this method is applicable not only to Vision Transformer but also to its variants. Additionally, pruning techniques were explored to improve the computational efficiency of Data-efficient Image Transformer model, reducing inference times and moving closer to the speed required for real-time applications, though Convolutional Neural Networks remain faster for such tasks.


1 INTRODUCTION


Unmanned Surface Vehicles (USVs) are types of robotic vehicles that can operate autonomously or by remote control in marine environments, performing tasks without human intervention. Today, the expanding operational areas of USVs have marked a turning point in the maritime sector. They are effectively used in various fields such as marine research, environmental monitoring, military reconnaissance and surveillance, and search and rescue operations. Object detection and ship classification in marine environments are challenging tasks that are critical for maritime safety and navigation. This requires handling image distortions caused by factors such as changing weather conditions, wave motion, reflections, and lighting. These factors can significantly im-

part the accuracy and reliability of the systems. Ship classification becomes even more complex due to the diversity, sizes, and movements of ships on the sea.

Advanced image processing technologies and deep learning methods are employed to overcome these challenges. In recent years, object detection algorithms like YOLO (You Only Look Once) (Redmon et al., 2016) and innovative approaches such as Vision Transformers (ViTs) have made significant progress in this field. These methods make object detection and ship classification in marine environments faster, more accurate, and more effective, opening new horizons for maritime safety and navigation.

YOLOv5 (Jocher, 2020) is an advanced version of the original YOLO, renowned for its speed and accuracy in real-time object detection. It has gained significant traction in the field of marine object detection due to its several key advantages. Its lightweight architecture allows it to perform detections quickly,

^a  <https://orcid.org/0009-0005-1121-8974>

^b  <https://orcid.org/0000-0003-3697-8548>

which is crucial for the dynamic and challenging environment where objects such as ships, buoys, and marine life need to be identified promptly.

Recent advancements in maritime surveillance have necessitated more effective object detection methodologies, particularly for small, difficult-to-detect objects. The TPH-YOLOv5 (Zhu et al., 2021) is an enhanced version of the YOLOv5 model, incorporates Transformer Prediction Heads (TPH) and the Convolutional Block Attention Module to address these challenges. CBAM aids the model in focusing on relevant areas within dense scenes, thereby enhancing detection accuracy. This model significantly improves the detection of small-scale objects by leveraging the self-attention mechanism of transformers, which provides superior feature representation.

Vision Transformers (Dosovitskiy et al., 2021), Data-efficient image Transformers (DeiT) (Touvron et al., 2021), Swin Transformers (Liu et al., 2021), and ConvNeXt (Woo et al., 2023) are recent advancements in neural network architectures that have transformed how we approach image classification tasks. ViT applies transformers directly to image patches and treats them as tokens. It offers a different approach compared to CNNs, achieving excellent results when pre-trained on large datasets. DeiT further optimizes this approach by introducing techniques like distillation to train more data-efficient models without reliance on extensive computational resources. Swin Transformers introduce a hierarchical structure that uses shifted windows to limit self-attention computation to local windows while allowing cross-window connection. Lastly, ConvNeXt modernizes the traditional CNN architecture by integrating transformer-like elements, such as layer scale and inverted bottlenecks, improving performance on par with more advanced transformer models. These architectures offer powerful options for handling various image classification tasks.

This study first concentrates on object detection within marine environments utilizing YOLOv5 based models. Our study shows TPH-YOLOv5's better performance on the MODS (Bovcon et al., 2022) dataset, highlighting its potential to detecting small objects in USVs for maritime surveillance. Subsequently, we performed ship classification using the increasingly prominent vision transformer based models and compared their class-specific accuracies. To the best of our knowledge, this is the first study to use vision transformers for ship classification. With this approach, we outperformed the state-of-the-art. Additionally, we utilized the attention layers within these models to generate attention flow maps — a special-

ized type of heatmap designed to visualize the focused areas by the transformers. This study also proposes a general method for applying attention maps to different versions of the ViT model. These visualizations provide crucial insights into the regions of the images that the models primarily target during the classification process. Furthermore, we show the affect of post-training pruning techniques on inference times without significantly affecting accuracy.

All experiments in this study were conducted on the NVIDIA Orin AGX Developer Kit, a high-performance computing platform designed for edge AI and robotics applications. This platform's advanced GPU architecture and efficient parallel processing capabilities were essential for training and evaluating the deep learning models used in this research.

The remainder of this paper is structured as follows: Section 2 presents a review of the literature relevant to this field. Section 3 describes the datasets employed in our analysis. Section 4 details the implementation and results, which are divided into object detection and ship classification. Finally, Section 5 summarizes the findings.

2 RELATED WORKS

When examining the current literature on object detection and classification in marine environments, it is evident that Convolutional Neural Network (CNN)-based learning algorithms are predominantly preferred. In this context, most researchers initially address the problem from the perspective of object detection. Moreover, in their studies, authors frequently emphasize that factors such as reflections and adverse environmental conditions posed by the marine environment negatively impact the results. This section will examine and discuss the methodologies and results of similar studies.

MODS is a dataset presented at the MaCVi'23 (Kiefer et al., 2023) competition. This competition holds a central position in demonstrating the latest methods in the field of marine environment object detection and segmentation. The participating teams faced the necessity to establish a crucial balance between achieving high accuracy and reasonable inference speeds. Most approaches focused on improving the detection of small objects, which is of critical importance in maritime surveillance. In USV object detection challenge, teams improved upon the baseline method, Mask R-CNN (He et al., 2017), using advanced models. The Fraunhofer IOSB team took first place with their DetectorRS model, noted for its abil-

ity to detect smaller objects in aquatic environments, although it sometimes misidentified water reflections as objects.

Another prominent study involves the use of WaSR (Water Segmentation and Refinement) (Bovcon and Kristan, 2022), an advanced network focused on detecting obstacles in water environments using semantic segmentation. It effectively distinguishes between water, obstacles, and sky in images. Another study, eWaSR (Teršek et al., 2023), is a variant of WaSR designed for computationally limited embedded devices. It maintains similar detection performance with a very minor decrease in F1 score (0.52% less than WaSR) but significantly reduces computational requirements, operates 10 times faster on a standard GPU, and can work on embedded sensors where WaSR cannot due to memory constraints. Lastly, WaSR-T (Žust and Kristan, 2022) enhances WaSR by incorporating texture information over time, better handling challenges such as reflections. This version improves performance in challenging lighting and water conditions, making it more robust in dynamic marine environments.

The study (Aguilar et al., 2023) explores obstacle detection and avoidance in USVs using CNNs and semantic segmentation. It utilizes the Mastr1325 dataset (Bovcon et al., 2019), containing 1,325 pixel-wise annotated images for semantic segmentation, and the Marine Obstacle Detection Dataset (MODD) (Bovcon et al., 2018), consisting of 12 videos (4,454 frames) captured by real USVs under diverse conditions. These datasets allow the study to explore semantic segmentation for identifying obstacles and calculating safe routes based on regions of interest within segmented images. This study emphasizes reducing computational complexity through pre-segmentation and horizon line detection. Their method uses semantic segmentation to distinguish between sky and water regions, enabling more efficient processing and reducing false positives. The method achieves over 90% accuracy in identifying obstacles under various environmental conditions, including diverse lighting, weather, and high maritime traffic scenarios. These results show the applicability of their methodology for real-time navigation, emphasizing its practical utility in improving the safety and autonomy of USVs in complex maritime environments.

When we look at the ship classification task, the largest and most detailed dataset, MARVEL (Gundogdu et al., 2017), stands out. The creators of this dataset have performed ship classification using a deep learning model called AlexNet (Krizhevsky et al., 2012). Despite class imbalances in the dataset, equal numbers of examples from each class were se-



Figure 1: Sample images from the MARVEL (Gundogdu et al., 2017) dataset and their classes. From left to right classes container, fishing vessel, dredger and tug.

lected to build training and testing sets, 8192 and 1024, respectively. The model's classification accuracy for 26 classes was found to be 73.14%, a significant improvement over the 53.89% accuracy obtained using a Support Vector Machine (Hearst et al., 1998).

An optimized CNN-based system for classifying marine vehicles using deep learning and transfer learning is presented in (Salem et al., 2023). Various CNN models, including MobileNetV2 (Sandler et al., 2018) and EfficientNet (Tan and Le, 2019), were trained on the Game of Deep Learning Ship dataset available on Kaggle¹, and the top-performing model was further tested on MARVEL. The results, with a total of 10,000 images across 5 classes (cargo, military, cruise, carrier, and tanker), demonstrated a high classification accuracy of 97.04%, outperforming other methods. (Salem et al., 2022) also worked on MARVEL by utilizing pre-trained models like EfficientNet (B0-B5) (Tan and Le, 2019), ResNet-152 (He et al., 2016), and InceptionV3 (Szegedy et al., 2015). The researchers aim to reduce training time and resource consumption without compromising the performance of image classification tasks for 26 ship classes. Their experiments demonstrate that the EfficientNet B5 architecture is superior to the other models, achieving a top accuracy of 91.60%. This improvement is notable as it exceeds previous best results in maritime vessel image classification.

The enhancement of ship classification accuracy through deep learning methods, specifically CNNs, is explored in (Leclerc et al., 2018). The study focuses on the maritime domain, employing transfer learning with pre-trained CNN architectures, such as Inception (Szegedy et al., 2015) and ResNets (He et al., 2016), to adapt to a limited dataset of maritime vessel images. These models were initialized with weights from ImageNet (Deng et al., 2009), enabling them to refine and build upon prior work more effectively. This approach demonstrates a substantial improvement over the existing state-of-the-art methods.

¹Kaggle dataset: <https://www.kaggle.com/datasets/arpitjain007/game-of-deep-learning-ship-datasets>

Table 1: Detailed comparison of performance metrics for various YOLOv5 models on the MODS (Bovcon et al., 2022) dataset.

	Precision	Recall	F1	mAP@.5	mAP@.5:.95
YOLOv5-s (Jocher, 2020)	0.845	0.745	0.791	0.793	0.403
YOLOv5-l (Jocher, 2020)	0.897	0.812	0.852	0.844	0.487
TPH-YOLOv5 (Zhu et al., 2021)	0.894	0.830	0.860	0.863	0.493

Table 2: Class-specific performance metrics for YOLOv5 models on the MODS dataset.

	Pr / ship Re / ship mAP@.5 / ship	Pr / person Re / person mAP@.5 / person	Pr / other Re / other mAP@.5 / other
YOLOv5-s (Jocher, 2020)	0.924	0.714	0.812
	0.923	0.625	0.484
	0.947	0.687	0.607
YOLOv5-l (Jocher, 2020)	0.919	0.916	0.799
	0.938	0.818	0.545
	0.95	0.824	0.648
TPH-YOLOv5-s (Zhu et al., 2021)	0.915	0.895	0.794
	0.920	0.818	0.609
	0.941	0.852	0.666

3 DATASETS

This study is conducted on two datasets, MODS (Bovcon et al., 2022) and MARVEL (Gundogdu et al., 2017), which offer comprehensive benchmarks for object detection and segmentation in maritime environments and ship classification, respectively. The MODS dataset is specifically designed for USV operations, focusing on obstacle detection and segmentation in real-world maritime scenarios. It includes stereo images captured from a USV navigating diverse coastal areas. Each image in the dataset is annotated to include dynamic obstacles, such as ships, swimmers, and other moving objects, using bounding boxes. Additionally, static obstacles, such as buoys are carefully annotated. This detailed annotation ensures precise evaluation of both dynamic and static obstacle detection, which is critical for USV navigation and collision avoidance. Images were captured under varying weather conditions and includes challenging features such as sun-glitter, sea foam, and dense object scenarios. The MODS dataset includes detailed labels for both detection and segmentation tasks. It is particularly used for detecting obstacles that USVs may encounter, especially in small sizes.

On the other hand, the MARVEL dataset stands out as the most extensive and comprehensive dataset ever created for ship classification. This dataset, hosting about 2 million images, divides five main ship types (cargo, military, carrier, cruise, and tanker) into 26 different subtypes. Figure 1 shows some sam-

ple images from the dataset. The dataset’s large size and diversity make it ideal for exploring advanced deep learning architectures, including CNNs and ViTs. MARVEL not only supports the development of high-accuracy models for ship classification but also contributes to maritime safety and monitoring applications.

Preprocessing and Dataset Adaptation

Both datasets required preprocessing to align with the goals of this study:

- **MODS:** The annotations were converted into a YOLO-compatible format to optimize the training of YOLOv5-based models. Around 9,000 images have been divided into approximately 80% training, 10% validation, and 10% test data.
- **MARVEL:** To address the inherent class imbalance in the MARVEL dataset and enhance model generalization, data augmentation techniques were employed. These techniques included horizontal flipping, translation, random brightness and contrast adjustments, and scaling, all aimed at diversifying the training data and ensuring equitable representation across vessel classes. This preprocessing step was crucial for maintaining the dataset’s robustness and preventing model overfitting.

For experimentation, the dataset was structured to include 3,600 training images, 800 validation images, and 800 test images per class, ensuring consistency and reliability in model evaluation.



Figure 2: Comparison of object detection performance across three YOLO model variants (YOLOv5-s (Jocher, 2020), YOLOv5-l (Jocher, 2020), TPH-YOLOv5 (Zhu et al., 2021)) on the MODS (Bovcon et al., 2022) dataset, illustrating detection outcomes in various maritime scenes (best viewed in digital format zoomed in).

This balanced approach allowed for comprehensive testing of classification algorithms, particularly for less frequent vessel types, while maintaining high generalizability and accuracy.

4 OBJECT DETECTION

One of the objectives of this study is to detect ships, humans, and objects in the marine environment. After reviewing the literature and conducting preliminary analyses, object detection was chosen over segmentation due to the need for fast and accurate responses in real-time, critical for USVs operating in dynamic and hazardous conditions.

Based on the analysis of the MODS dataset, the problem was defined as detection of small objects. TPH-YOLOv5, optimized for small objects, was compared with YOLOv5, which was chosen for its balance of accuracy, speed, and efficiency, especially for real-time applications. TPH-YOLOv5 was included due to its specialized design for this problem. Both models were initialized with pre-trained ImageNet weights and fine-tuned for 130 iterations on the MODS dataset.

Table 1 provides a detailed comparative evaluation of the overall performances using metrics such as precision, recall, F1 score, and mean average precision (mAP) at two intersection over union (IoU) thresholds. Here, TPH-YOLOv5 emerges as the superior model, demonstrating higher F1 scores and mAP values, which indicates its robustness in detecting objects with greater accuracy and consistency. Table 2 further delves into class-specific performance, highlighting how each model performs in detecting different objects like ships and persons. If we look at the increase in score for the person and other classes compared to other models, we can conclude that the

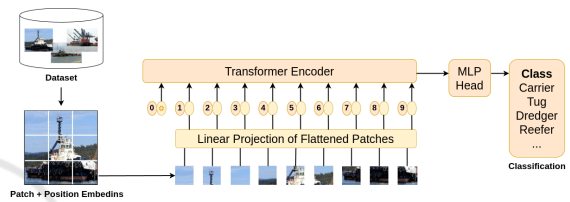


Figure 3: Graphical representation of training process of the vision transformers on MARVEL dataset.

ability to detect small objects has improved.

Figure 2 presents a comparative visualization of detection performance using three variants, as applied to the MODS dataset in diverse maritime settings. Each row corresponds to a different model, showing their efficiency in identifying and classifying objects under varying environmental conditions. Notably, TPH-YOLOv5 demonstrates a marked improvement in detecting smaller objects compared to the other models. This enhancement is attributed to the integration of Transformer Prediction Heads in the TPH-YOLOv5 architecture, which enhances the model's sensitivity to smaller-scale features and dynamic obstacles. This capability is critical for applications requiring high accuracy in cluttered and challenging environments, such as navigation and surveillance in maritime domains.

5 SHIP CLASSIFICATION

Traditional methods predominantly employ CNNs, which, while effective, are primarily designed to capture spatial hierarchies in images. Recent advancements in deep learning have introduced transformer-based models, which leverage self-attention mechanisms to process data in a manner that could potentially outperform conventional CNNs in terms of both accuracy and efficiency in certain tasks.

Table 3: Performance Comparison of Various Neural Network Architectures on the MARVEL(Gundogdu et al., 2017) Dataset, Evaluating Classification Accuracy Across Different Methods and Class Sizes.

Study	#class	method	acc.type	acc.(%)
Erhan et al.(Gundogdu et al., 2017)	26	AlexNet (Krizhevsky et al., 2012)	Val	73.14
Erhan et al.(Gundogdu et al., 2017)	26	SVM (Hearst et al., 1998)	Val	53.89
Leclerc et al.(Leclerc et al., 2018)	26	Inception-v3 (Szegedy et al., 2015)	Val	78.73
Leclerc et al.(Leclerc et al., 2018)	26	ResNet (He et al., 2016)	Val	75.84
Salem et al.(Salem et al., 2023)	5	EfficientNetB2 (Tan and Le, 2019)	Test	97.04
Salem et al.(Salem et al., 2022)	26	EfficientNet-B5 (Tan and Le, 2019)	Val	91.60
This study	26	ResNet50 (He et al., 2016)	Test	91.76
This study	26	ResNet101 (He et al., 2016)	Test	89.87
This study	26	ViT-B (Dosovitskiy et al., 2021)	Test	82.68
This study	26	DeiT-B (Touvron et al., 2021)	Test	92.87
This study	26	Swin-T (Liu et al., 2021)	Test	87.72
This study	26	Swin-B (Liu et al., 2021)	Test	90.44
This study	26	ConvNext-v2 (Woo et al., 2023)	Test	90.07

Table 4: Comparison of Image Classification Models Based on Standardized Image Size of 224x224, Floating Point Operations per Second (FLOPs) and Frame per Second (FPS).

Model	FLOPs	FPS
ResNet50 (He et al., 2016)	4.1G	52.66
ResNet101 (He et al., 2016)	7.8G	54.00
ViT-B (Dosovitskiy et al., 2021)	55.4G	0.95
DeiT-B (Touvron et al., 2021)	17.5G	1.31
Swin-B (Liu et al., 2021)	15.4G	0.51
Swin-T (Liu et al., 2021)	4.5G	1.14
ConvNext-v2 (Woo et al., 2023)	115G	0.13

In this study, we explore the applications of ViT (Dosovitskiy et al., 2021), DeiT (Touvron et al., 2021), Swin Transformers (Liu et al., 2021), and ConvNext-v2 (Woo et al., 2023) for the task of ship classification, utilizing the MARVEL dataset. Due to the large size of the MARVEL dataset, previous studies have typically used random subsets of the data, making direct comparisons across different models challenging. To maintain consistency in our experiments and ensure a fair evaluation, we followed a similar approach by utilizing a random subset of the dataset. Although our dataset size is similar to those in other studies, the processes we applied to reduce data size and address class imbalance issues introduced differences in our datasets. Due to these differences, we fine-tuned convolution-based networks to allow for a fair comparison with our transformer-based models. Additionally, we conduct a detailed analysis of each model’s performance, focusing on accuracy and computational speed.

Figure 3 illustrates the training process of Vision Transformers (ViTs), where an input image is divided into smaller patches, flattened into vectors, and then embedded with positional information to retain the spatial arrangement of the patches. These patch embeddings are passed through a transformer encoder,

which applies self-attention mechanisms to capture global relationships between different parts of the image. By doing so, the model can identify dependencies across patches, even if they are far apart in the original image, which is crucial for understanding the overall context. The output from the transformer encoder is then fed into a multilayer perceptron (MLP) head for classification, where the final decision is made based on the global information gathered from the attention layers. The attention mechanism, in particular, allows the model to focus on the most relevant patches, effectively classifying images without relying on traditional convolutional layers, which are commonly used in CNNs for local feature extraction. ViTs thus enable efficient image classification by modeling long-range dependencies across patches, making them a powerful alternative to convolution-based models.

The methodology includes resizing images to 224x224x3 pixels, and employs an AdamW (Loshchilov and Hutter, 2019) optimizer with a linear learning rate (LR) scheduler. The hyperparameters are optimized values tailored for each model to achieve optimal performance. ViT, DeiT, SwinT-base, and ConvNext-T all share an optimized learning rate of 1e-4, a weight decay of 0.01, and a linear LR schedule. SwinT-Tiny uses a slightly lower learning rate of 5e-5. All models, except ViT and DeiT, include a 0.1 warmup ratio. Each model is trained for 15 epochs with a batch size of 10 for training and 4 for evaluation. These configurations reflect careful tuning to align with the architectural differences and training requirements of each model, ensuring the best possible performance on the given tasks.

Table 3 presents the performance of various neural network architectures on the MARVEL dataset. While most previous studies, such as (Gundogdu

et al., 2017), (Leclerc et al., 2018), and (Salem et al., 2023), reported their models' performance on the validation set, we conducted evaluations on the test set. Notably, our method, using the DeiT model, achieved a test-set accuracy of 92.87%, surpassing the previous best score of 91.60%, which was recorded by (Salem et al., 2023) with the EfficientNet-B5. Both this study and the two previous works involved the classification of 26 distinct classes. This result highlights the potential benefits of transformer-based architectures, like DeiT, over conventional CNNs, particularly in handling complex image classification tasks within a challenging dataset like MARVEL.

When we examine the detailed comparison of several neural network models used for ship classification, Table 4 highlights computational characteristics and response times which are key metrics for real-time maritime operations. Traditional CNN models like ResNet50 and ResNet101, which showed good frame per second (FPS), also achieve high accuracy. This makes ResNet architectures a solid choice for balancing both speed and performance.

Transformer-based models, while having lower FPS—show promising accuracy, with DeiT-B achieving the highest accuracy of 92.87%. However, the high FLOPs and low FPS of these models indicate that they are more suited for scenarios where accuracy is prioritized over speed. ConvNext-v2, despite being the most computationally expensive, still manages to offer a strong accuracy of 90.07%. Thus, when considering both response times and accuracy, traditional CNN models remain more practical for real-time applications, while transformer models excel in accuracy but are better suited for offline tasks.

5.1 Pruning

Pruning is a model compression technique used in deep learning to reduce the size and computational requirements of large neural networks by removing redundant parameters. This process helps maintain high accuracy while cutting down on the storage and processing power needed, making the models more efficient. Pruning involves identifying and discarding low-impact weights and neurons, often those with low magnitude or similar activations. The result is a lighter model with improved generalization, and better suitability for deployment on low-resource devices like embedded systems.

We observed that Vision Transformers, are not suitable for real-time applications due to their low FPS. Therefore, we experimented with various pruning techniques to enhance their efficiency. Among the models we tested, DeiT had the highest accuracy,

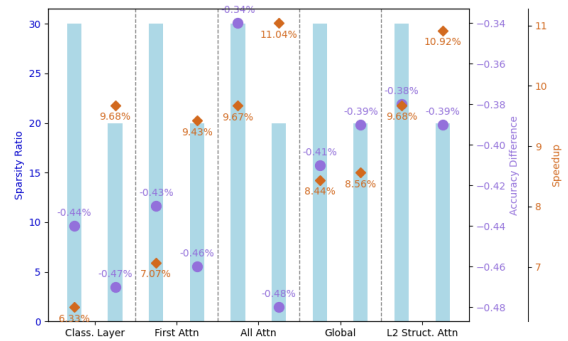


Figure 4: Sparsity vs Accuracy and Inference time change when apply pruning to our finetuned DeiT model (best viewed in digital format zoomed in).

which is why we focused our pruning experiments on this model. We applied various post-training pruning techniques on the DeiT model to assess their impact on the efficiency and accuracy of ship classification systems. These experiments used both local and global pruning strategies. Local pruning targeted specific model components, including the classification layer, the first attention layer, and all attention layers. Global pruning was applied across the model as a whole. Despite DeiT being the best model in terms of accuracy, its FPS was extremely low, making it inefficient for real-time applications. Pruning was necessary to reduce the model's computational load, aiming to provide a balance between maintaining high classification accuracy and improving inference speed.

Figure 4 illustrates the effects of various pruning strategies on the DeiT model, specifically targeting different components. For each pruning method, the sparsity ratio (in blue bars), speedup (in orange diamonds), and accuracy difference (in purple dots) are compared. The results reveal subtle changes in model performance, with accuracy reductions ranging from 0.34% to 0.48%. Meanwhile, inference time showed notable improvements, with speedups ranging from 6.33% to 11.04%, reflecting an important improvement in computational efficiency. The figure indicates that L2 structured pruning with 20% parameter reduction has optimal balance between maximum speedup and minimal accuracy degradation. It achieves a 10.92% speedup while maintaining a small accuracy drop of only 0.39%. However, despite the improvements, the speed-up achieved with DeiT still falls short of reaching the FPS seen in CNNs.

5.2 Attention Rollout

Heatmaps are visual tools used to highlight the regions within an image that a model focuses on during its decision-making process. They provide insight

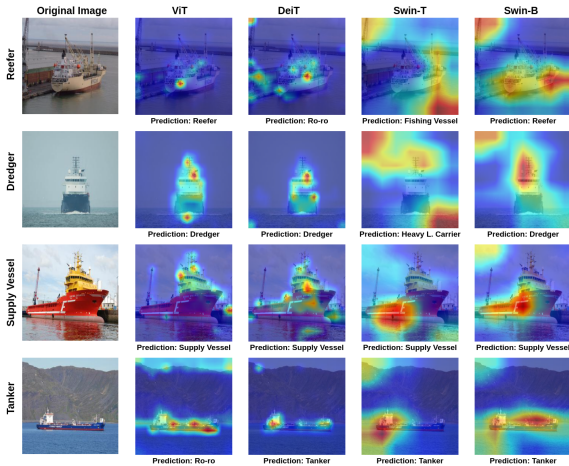


Figure 5: Attention rollout maps for true and false predicted samples on vision transformer models.

into the inner workings of machine learning models by revealing which features or areas are most influential in generating predictions. They are commonly used in image classification and text analysis to show which parts of the input the model focuses on. Colors typically range from cool (blue) for less attention to warm (red) for high influence. Heatmaps enhance transparency, making AI models easier to understand, helping identify biases or errors, and improving trust and performance by showing how the model processes inputs.

In this work, heatmaps are derived from the attention mechanisms of Vision Transformers (ViTs) and their variants, enabling the visualization of how these models process and prioritize different parts of an image. This approach helps identify whether the models are correctly focusing on relevant features, such as the shape or size of ships, while ignoring background noise like waves or reflections. Unlike previous work (Abnar and Zuidema, 2020), which often generated heatmaps specific to ViT, this study introduces a generic method that can be applied across various transformer architectures, including DeiT and Swin Transformers. This generalization allows for consistent interpretability and model validation for maritime applications.

To better understand which parts of an image the vision transformer models focus on, we used the attention layers to create rollout maps, a type of heatmap. These maps highlight the regions of the image that the model pays the most attention to during classification. By analyzing these visualizations, we gain valuable insights into how the model processes and prioritizes different parts of the image when making decisions.

In our exploration of the attention rollout tech-

nique (Abnar and Zuidema, 2020), we adopted a methodology for quantifying attention flow in transformers that significantly enhances our understanding of how information propagates through the model’s layers. This approach presents novel post-hoc methods to approximate attention to input tokens using attention weights. These techniques, referred to as ”attention rollout” and ”attention flow”.

$$\text{Let } A_i = \text{attention matrix for layer } i, \quad (1)$$

$$\hat{A}_i = \max(A_i) \quad (\text{Fused Attention}) \quad (2)$$

$$A'_i = \frac{\hat{A}_i + I}{\sum(\hat{A}_i + I)} \quad (\text{Normalized}) \quad (3)$$

$$\text{Rollout}_i = A'_i \cdot \text{Rollout}_{i-1} \quad (4)$$

The implementation involved manipulating attention matrices, to remove less informative attention scores based on a specified discard ratio. Among the provided three options (max, min and mean) on the attention heads, we employed a maximum fusion strategy in Eq. 2 to better capture significant areas that any single head may highlight. This was complemented by a strategy to discard lower attention values, which significantly improved the clarity of attention maps by highlighting the strongest or most activated features. Additionally, residual connections are included by adding the identity matrix I to the attention matrix of each layer. To capture the cumulative attention across layers, the final attention map was generated by multiplying the refined attention matrices, as shown in Eq. 4, and then normalizing to maintain a consistent distribution of attention across the image.

This methodology allowed for a more focused and interpretable visualization of where the model directs its attention, providing insights into its decision-making process. By employing this technique, we effectively address the challenge of visualizing attention in deeper layers of the model. Our modified attention rollout method can be applied not only to the original ViT but also to its variants, such as DeiT and Swin Transformers.

For models like DeiT that include a classification token (CLS token), we extracted this token from all attention layers before combining them, as seen in Eq. 5. In this equation, the number of CLS tokens is represented by the variable k . This operation indicates that we select all elements from the last two dimensions starting from the index k onwards, effectively skipping or ignoring the first k entries in both dimensions.

$$\tilde{A}_i = A_i[\dots, k :, k :] \quad (5)$$

Similarly, for Swin Transformers, we averaged across the batch dimension to reduce window size, as shown in Eq. 6. This process takes the mean of the attention across the batch dimension (dimension 0).

$$\tilde{A}_i = \text{mean}(A_i, \text{dim} = 0, \text{keepdim} = \text{True}) \quad (6)$$

This enhanced method not only aligns with the findings from (Abnar and Zuidema, 2020) but also adapts their insights to improve the interpretability of deeper layers in other vision transformer models. This adaptation ensures that the attention mechanisms in ViTs and other transformer-based architectures can be effectively visualized, offering more intuitive explanations of model behavior across various vision tasks.

Figure 5 illustrates several input images that were correctly or incorrectly predicted. The rows correspond to specific vessel types, with the true class labels indicated on the far left of each row. Each column represents the output of a different model: the first shows the input image, and the remaining four columns show the outputs of ViT, DeiT, Swin-T, and Swin-B models, respectively. The maps in each cell represent the regions of the image where each model focused during classification, highlighting the areas that were most important in the model's decision-making process. Below each image, the predicted class for that specific model is displayed, showing how each model classified the vessel.

It is observed that inaccuracies in predictions are usually due to the models focusing on incorrect areas of the images. However, the rollout maps highlight the models' ability to focus on the relevant parts of the images without being distracted by background objects, particularly in correct predictions. This visualization helps in understanding how the models' attention mechanisms are engaged during prediction, confirming that they correctly identify and concentrate on pertinent features within the images.

6 CONCLUSION

This paper presents a comprehensive study on object detection and ship classification for Unmanned Surface Vehicles. By utilizing the MODS and MARVEL datasets, we applied advanced deep learning models, such as YOLOv5 and Vision Transformers, to address the challenges of detecting and classifying ships and other obstacles in dynamic maritime conditions. Our results demonstrate that the TPH-YOLOv5 model significantly outperformed other variants of YOLOv5, particularly in detecting small objects, achieving a high mAP score of 0.863.

For ship classification, Vision Transformer models, especially DeiT, achieved state-of-the-art performance with a classification accuracy of 92.87% on the MARVEL dataset. Although the DeiT model achieved a high classification accuracy of 92.87% on the MARVEL dataset, its low FPS performance makes it less suitable for real-time applications. The high computational demands of Vision Transformers can result in slower inference times, which is a critical limitation for time-sensitive tasks like ship classification in maritime environments. To address this, post-training pruning techniques were employed, reducing the model's complexity and improving its speed by up to 11%, with minimal impact on accuracy. Despite the improvements from post-training pruning, the speed and accuracy trade-off still led to the conclusion that ResNet-based CNN models are more practical for real-world applications.

Moreover, this research introduces an approach to visualize the attention areas on images, applicable not only to ViT but also its derivative models. The study also proposes strategies to improve response times, aiming to make these models more suitable for time-sensitive tasks in real-world maritime applications.

For future work, we plan to explore the integration of detection and classification tasks into a unified framework, leveraging the strengths of our models. Additionally, we aim to further optimize model performance through advanced techniques such as quantization and knowledge distillation, with a focus on improving real-time capabilities. Expanding the datasets and incorporating more challenging maritime conditions will also be key to enhancing model robustness.

REFERENCES

- Abnar, S. and Zuidema, W. (2020). "Quantifying Attention Flow in Transformers". In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197.
- Aguiar, R., Bertini, L., Copetti, A., Clua, E. W. G., Gonçalves, L. M. G., and Moreira, L. B. (2023). Semantic Segmentation and Regions of Interest for Obstacles Detection and Avoidance in Autonomous Surface Vessels. In *2023 Latin American Robotics Symposium (LARS), 2023 Brazilian Symposium on Robotics (SBR), and 2023 Workshop on Robotics in Education (WRE)*, pages 403–408.
- Bovcon, B. and Kristan, M. (2022). WaSR—A Water Segmentation and Refinement Maritime Obstacle Detection Network. *IEEE Transactions on Cybernetics*, 52(12):12661–12674.
- Bovcon, B., Mandeljc, R., Perš, J., and Kristan, M. (2018). Stereo obstacle detection for unmanned sur-

- face vehicles by IMU-assisted semantic segmentation. *Robotics and Autonomous Systems*, 104:1–13.
- Bovcon, B., Muhovič, J., Perš, J., and Kristan, M. (2019). The MaSTr1325 dataset for training deep USV obstacle detection models. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3431–3438.
- Bovcon, B., Muhovič, J., Vranac, D., Mozetič, D., Perš, J., and Kristan, M. (2022). MODS—A USV-Oriented Object Detection and Obstacle Segmentation Benchmark. *IEEE Transactions on Intelligent Transportation Systems*, 23(8):13403–13418.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houshy, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations, ICLR 2021*.
- Gundogdu, E., Solmaz, B., Yücesoy, V., and Koç, A. (2017). MARVEL: A Large-Scale Image Dataset for Maritime Vessels. In *Computer Vision – ACCV 2016*, pages 165–180.
- He, K., Gkioxari, G., Dollár, P., and Girshick, R. B. (2017). Mask R-CNN. In *ICCV*, pages 2980–2988.
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep Residual Learning for Image Recognition. In *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR '16*, pages 770–778.
- Hearst, M., Dumais, S., Osuna, E., Platt, J., and Scholkopf, B. (1998). Support Vector Machines. *IEEE Intelligent Systems and their Applications*, 13(4):18–28.
- Jocher, G. (2020). ultralytics/yolov5: v3.1 - Bug Fixes and Performance Improvements. <https://github.com/ultralytics/yolov5>.
- Kiefer, B. et al. (2023). 1st Workshop on Maritime Computer Vision (MaCVi) 2023: Challenge Results. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pages 265–302.
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, volume 25.
- Leclerc, M., Tharmarasa, R., Florea, M. C., Boury-Brisset, A.-C., Kirubarajan, T., and Duclos-Hindie, N. (2018). Ship Classification Using Deep Learning Techniques for Maritime Target Tracking. In *2018 21st International Conference on Information Fusion (FUSION)*, pages 737–744.
- Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *CoRR*, abs/2103.14030.
- Loshchilov, I. and Hutter, F. (2019). Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.
- Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788.
- Salem, M. H., Li, Y., and Liu, Z. (2022). Transfer Learning on EfficientNet for Maritime Visible Image Classification. In *2022 7th International Conference on Signal and Image Processing (ICSIP)*, pages 514–520.
- Salem, M. H., Li, Y., Liu, Z., and AbdelTawab, A. M. (2023). A Transfer Learning and Optimized CNN Based Maritime Vessel Classification System. *Applied Sciences*, 13(3).
- Sandler, M., Howard, A. G., Zhu, M., Zhmoginov, A., and Chen, L. (2018). Inverted Residuals and Linear Bottlenecks: Mobile Networks for Classification, Detection and Segmentation. *CoRR*, abs/1801.04381.
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2015). Rethinking the Inception Architecture for Computer Vision. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826.
- Tan, M. and Le, Q. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114.
- Teršek, M., Žust, L., and Kristan, M. (2023). eWaSR – An Embedded-Compute-Ready Maritime Obstacle Detection Network. *Sensors*, 23(12):5386.
- Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jegou, H. (2021). Training data-efficient image transformers & distillation through attention.
- Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I. S., and Xie, S. (2023). ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16133–16142.
- Zhu, X., Lyu, S., Wang, X., and Zhao, Q. (2021). TPH-YOLOv5: Improved YOLOv5 Based on Transformer Prediction Head for Object Detection on Drone-captured Scenarios. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 2778–2788.
- Žust, L. and Kristan, M. (2022). Temporal Context for Robust Maritime Obstacle Detection. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6340–6346.