SCIENTIFIC JOURNAL
OF MARITIME RESEARCH
[POMORSTVO]

University of Rijeka
FACULTY OF MARITIME STUDIES

https://doi.org/10.31217/p.38.2.6

# Zero-Shot Learning in Maritime Domain: Classification of Marine Objects using CLIP

Ivan Lorencin[1,*], Domagoj Frank[2] , Damir Vusić[2]

[1] Juraj Dobrila University of Pula, Faculty of Informatics, Rovinjska 14, 52100 Pula, Croatia, e-mail: ilorencin@unipu.hr
[2] University North, Trg dr. Žarka Dolinara 1, 48000, Koprivnica, Croatia, e-mail: dfrank@unin.hr; dvusic@unin.hr
[*] Corresponding author

ARTICLE INFO

**Original scientific paper**
Received 11 September 2024
Accepted 2 October 2024

**Key words:**
CLIP
Marine objects
Vision-language Models
Transformers
Zero-shot Learning

ABSTRACT

Maritime security and monitoring are essential for global trade, environmental protection, and national defense. Traditional machine learning models have been effective in recognizing and classifying maritime objects, but their reliance on large, labeled datasets poses significant challenges, particularly in dynamic environments where new and unforeseen objects frequently emerge. This study explores the application of Zero-Shot Learning (ZSL) to the maritime domain, leveraging the CLIP (Contrastive Language-Image Pre-training) model to classify maritime objects with minimal labeled data. A custom dataset comprising 1,438 images across five maritime object categories—boat, cargo, cruise, dock, and lighthouse—was curated for evaluation. Four CLIP model variants were examined: "clip-vit-base-patch16," "clip-vit-base-patch32," "clip-vit-large-patch14," and "clip-vit-large-patch14-336." The study's findings indicate that the CLIP models, particularly the "clip-vit-large-patch14-336" variant, achieve high classification accuracy, with AUC values approaching 1.0 across most classes. Performance was strongest in easily distinguishable categories such as dock and lighthouse, but challenges remain with rare or ambiguous classes such as cargo ships, where F2 scores suggest variability in recall and precision. Additionally, the study highlights the potential limitations of these models, including their dependency on dataset diversity and potential biases introduced by web-scraped images, which may not fully represent the complex, real-world conditions of maritime environments.

## 1 Introduction

Maritime security and monitoring are critical for global trade, environmental protection, and national security. Traditional machine learning approaches have shown success in recognizing and classifying maritime objects such as ships, buoys, and marine animals [1-3]. However, these models typically rely on large, labeled datasets for training, which are not always available for every possible object of interest. This limitation is particularly pronounced in dynamic environments like oceans, where new types of vessels, debris, or other objects may frequently appear. The effective management and surveillance of maritime environments necessitate robust systems for object detection and classification. While traditional computer vision techniques have shown promise, their reliance on large, meticulously annotated datasets poses significant challenges, particularly in the context of maritime imagery. Zero-shot learning, an emerging paradigm in machine learning, offers a potential solution by enabling the recognition of unseen object categories without requiring corresponding training examples. By leveraging semantic and visual information, this research explores the feasibility of applying zero-shot learning to the maritime domain. Our aim is to develop a model capable of detecting and classifying a diverse range of maritime objects, including vessels, buoys, and marine debris, with minimal reliance on labeled data. Successful implementation of this approach could significantly advance maritime situational awareness and support a wide array of applications.

This paper is structured to explore the application of Zero-Shot Learning (ZSL) in the classification of maritime objects, focusing on the challenges posed by the scarcity of labeled data in dynamic maritime environments. It begins with the literature review provides an overview of recent advancements in machine learning and image processing within the maritime domain. Following this, the paper discusses the theoretical foundations of ZSL and its potential to generalize across unseen object categories. A description of the dataset used for evaluation is included, along with the methods for data collection and preparation. The results section presents a comparative analysis of different model variants, focusing on their performance across various maritime object categories. The paper concludes with a discussion of the strengths, limitations, and potential applications of the proposed approach in real-world maritime scenarios, offering insights into its future development.

## 2 Literature review

The maritime domain, characterised by vast and dynamic environments, has increasingly leveraged the capabilities of machine learning (ML) and image processing to address complex challenges. These technologies have been instrumental in enhancing maritime safety, efficiency, and environmental protection. Image processing techniques have been foundational in extracting meaningful information from maritime imagery. Image acquisition platforms, including satellites, aerial drones, and ship-borne sensors, generate vast amounts of visual data. Pre-processing techniques are applied to improve image quality and model performance [4]. Subsequently, feature extraction methods, including edge detection, texture analysis, and interest point detection, are employed to identify salient image regions [5]. Object detection algorithms, such as You Only Look Once (YOLO) [6-8], localise and classify maritime objects and recyclable materials within images. Application of deep learning in maritime environment has a vital role in protection an inspection of critical underwater infrastructure such as seafloor pipelines [9]. Furthermore, ML algorithms can also be used in modelling of the dynamics of tidal rivers and estuaries [10].

Machine learning algorithms have been integrated into maritime systems to analyse extracted image features and make informed decisions. While significant progress has been made, challenges persist in the application of ML and image processing to maritime domains. These include the variability of maritime conditions, the presence of occlusions and shadows in images, the scarcity of labeled data for specific maritime objects, and the computational demands of complex models. Addressing these challenges requires ongoing research and development to improve the robustness and reliability of maritime systems. Zero-shot learning (ZSL) has emerged as a promising approach to address the challenge of classifying unseen categories without corresponding training data. Early works primarily focused on attribute-based methods, where semantic attributes were used to bridge the gap between seen and unseen classes [11]. However, the reliance on handcrafted attributes limited their effectiveness.

Subsequently, generative models gained attention, aiming to synthesise training examples for unseen classes [12]. While these methods showed promise, they often suffered from mode collapse, and generating realistic samples remained challenging.

More recently, embedding-based approaches have emerged as a dominant paradigm in ZSL. By learning a joint embedding space for visual and semantic information, these methods enable similarity-based classification of unseen classes [13].

Despite the advancements in ZSL, its application to the maritime domain is relatively unexplored. The convergence of computer vision and natural language processing has given rise to language-vision models, a class of artificial intelligence systems capable of understanding and reasoning about both visual and textual information simultaneously. These models have emerged as a powerful tool for bridging the gap between the human-interpretable world of language and the machine-perceptible world of images. Early research in language-vision focused on establishing foundational connections between the two modalities. Tasks such as image captioning, which requires generating descriptive text for an image [14], and visual question answering, demanding an accurate textual response to a question about an image [15], were instrumental in developing techniques for multimodal representation learning. These initial studies laid the groundwork for understanding how to align visual and textual information, forming the basis for more complex language-vision models.

Building on these foundational works, the development of large-scale pre-trained language-vision models has marked a significant leap forward. Models like CLIP [16] and ViLBERT [17] have been trained on massive datasets of image-text pairs, enabling them to learn rich representations that capture complex relationships between visual and textual content. These pre-trained models have demonstrated exceptional performance on a wide range of tasks, including image classification, object detection, and visual grounding. By leveraging the knowledge acquired from extensive training data, these models have shown a remarkable ability to generalise to new tasks and domains.

This study hypothesises that the application of Zero-Shot Learning (ZSL) through language-vision models like CLIP can effectively address the challenges of maritime object classification, particularly in environments where labelled data is scarce or unavailable. We propose that these models, due to their ability to generalise from diverse image-text pairings, can accurately classify a wide range of maritime objects, including those that have not been seen during training. The conclusion of this research will evaluate the validity of this hypothesis by analysing the models' performance in classifying different maritime object categories and assessing their potential limitations in real-world maritime scenarios.

## 3 Zero-Shot Learning

Zero-shot learning (ZSL) is a challenging machine learning subfield focused on classifying unseen categories without corresponding training data. Unlike traditional supervised learning reliant on labelled data for each class, ZSL generalises knowledge from known classes to unknown ones. A core ZSL assumption is the existence of auxiliary information bridging the gap between seen and unseen classes. This information is often represented as semantic embeddings capturing object category semantics and attributes. These embeddings reside in a shared semantic space encompassing both seen and unseen classes [18]. The integration of language and vision modalities presents a promising avenue for enhancing maritime object detection and classification. By fusing the complementary strengths of these domains, we can develop more robust, accurate, and interpretable models. The incorporation of textual descriptions enriches object representations with semantic information, facilitating improved discrimination between object categories. Moreover, language-vision models have the potential to mitigate the challenges posed by limited labeled data in the maritime domain through techniques such as data augmentation and knowledge transfer from large-scale pre-trained models.

## 4 Dataset description

The dataset used in this study was curated through a comprehensive web scraping process, aimed at gathering a diverse and representative collection of images corresponding to five distinct maritime classes: boat, cargo, cruise, dock, and lighthouse. The primary objective of this dataset creation was to build a robust image repository that could effectively support the training, validation, and testing of machine learning models, particularly those focused on classification tasks within the maritime domain.

The dataset was compiled using automated web scraping techniques, which involved crawling various publicly accessible online sources, including image search engines, maritime-themed websites, and relevant databases. The web scraping process was carefully designed to ensure that the images collected were of high quality and relevant to the specific classes of interest. Each image was manually reviewed to confirm its relevance to the corresponding category, thereby minimising the inclusion of irrelevant or mislabeled images. The scraping process was conducted over several weeks, during which various scripts were employed to capture images based on keyword searches associated with each class. The images were then processed to remove duplicates and standardise their formats. This preprocessing step was crucial to maintaining the dataset's integrity and ensuring that the models trained on this data would be exposed to a wide variety of images, reflecting real-world variability in maritime settings. The final dataset comprises a total of 1,438 images, distributed across the five maritime classes as shown in Table 1. The distribution of images was intentionally kept unbalanced to reflect the natural occurrence of these classes in typical maritime environments. For instance, "cruise" class images are more prevalent due to the high visibility and frequent documentation of cruise ships compared to other classes such as "cargo" or "dock."

**Table 1** Dataset Composition

| Class | Number of Images |
|---|---|
| Boat | 159 |
| Cargo | 112 |
| Cruise | 824 |
| Dock | 175 |
| Lighthouse | 168 |
| **Total** | **1438** |

### Class Descriptions

Boat: This class includes a variety of small to medium-sized vessels, typically used for leisure or short-distance travel. The images in this class capture boats in different settings, such as in marinas, on open water, and docked at piers. The diversity within this class provides a comprehensive representation of recreational and functional boating scenarios.

Cargo: The cargo class encompasses images of commercial freight ships, including container ships, bulk carriers, and other vessels used for transporting goods across international waters. This class is characterized by large ships often depicted in ports, at sea, or in the process of loading and unloading cargo.

Cruise: The largest class in the dataset, cruise, contains images of cruise liners, which are a prominent feature of the maritime tourism industry. These images vary in perspective and context, capturing cruise ships from aerial views, at sea, docked at ports, and in various lighting conditions.

Dock: Images in the dock class feature maritime docking facilities, including piers, harbors, and mooring points where boats and ships are stationed. This class is essential for understanding the infrastructure associated with maritime activities. The dock images are diverse, depicting both bustling commercial docks and quieter, smaller-scale docking areas.

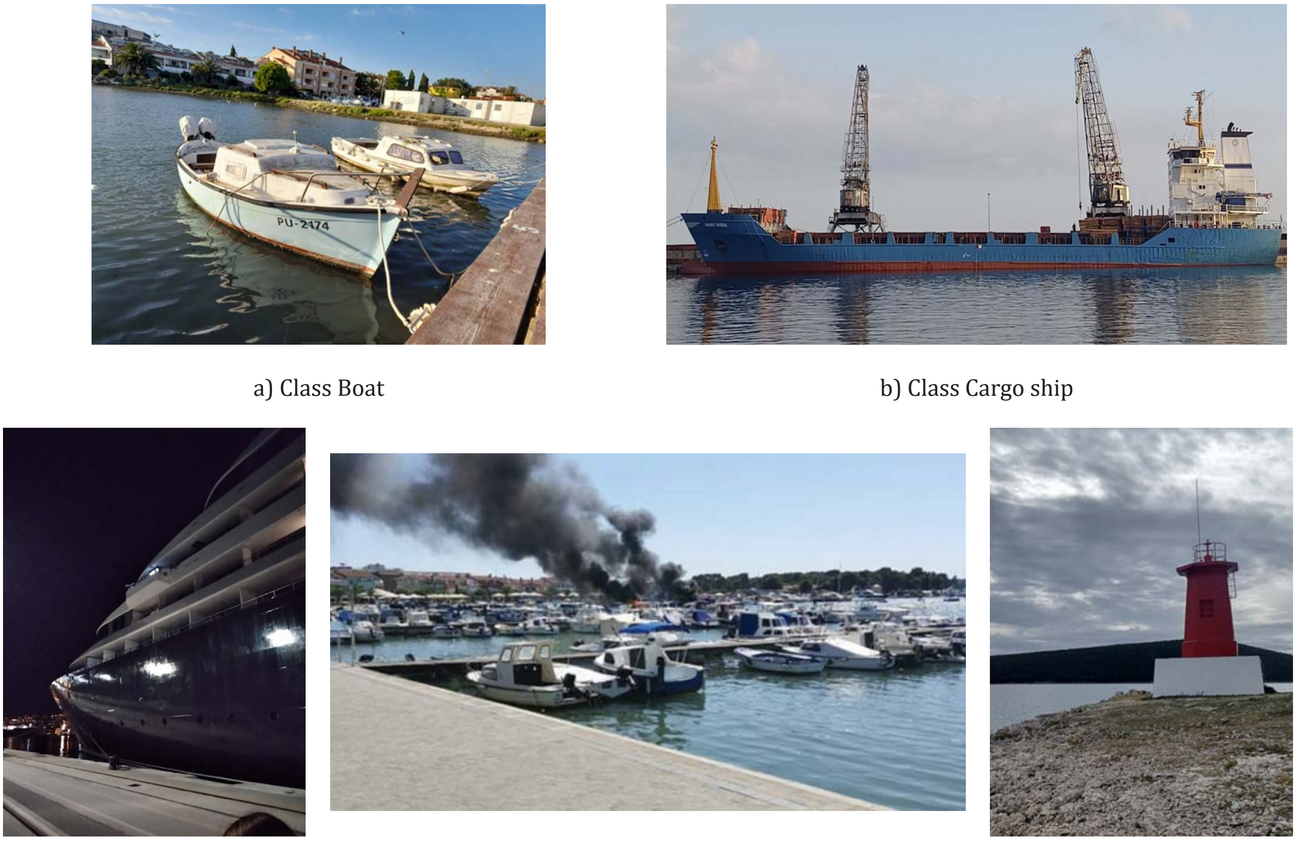
Lighthouse: The lighthouse class includes images of lighthouses situated in coastal areas, often serving as navigational aids for maritime vessels. These images capture lighthouses in various settings, from isolated rocky coastlines to more urbanized harbor areas. The inclusion of lighthouses provides a unique perspective on coastal features relevant to maritime navigation.

It is important to note that the dataset presented in this study was used exclusively for the testing and evaluation of different CLIP architectures. The dataset was not intended for training the models but rather to assess their zero-shot learning capabilities in the maritime domain. This approach allows for a more accurate evaluation of how well these models can generalize to new, unseen data, which is critical for their application in real-world maritime environments. The representation of each class is given in Figure 1.

The dataset used in this study reflects a notable class imbalance, with a large number of cruise ship images (824) compared to fewer images of other categories, such as cargo ships (112). This imbalance can be attributed to the varying availability and visibility of maritime objects, with cruise ships being more frequently documented in different settings, such as aerial views, ports, and at sea. Cruise ships can have diverse appearances, sizes, and conditions, necessitating a wide representation within this class to ensure comprehensive coverage. This abundance was intentionally included to capture as many variations of cruise ships as possible, as the dataset was primarily intended for testing the model's ability to generalize across different maritime objects. The dataset reflects the diversity of real-world maritime environments, while also acknowledging the challenges posed by this inherent class imbalance.

## 5 CLIP Vision Language model

CLIP (Contrastive Language-Image Pre-training) is a groundbreaking vision-language model developed by OpenAI that integrates visual and textual modalities in a shared representation space. This model is trained to understand and generate meaningful associations be-



a) Class Boat

b) Class Cargo ship

c) Class Cruise ship

d) Class Dock

e) Class Lighthouse

**Figure 1** Representation of classes contained in the data set

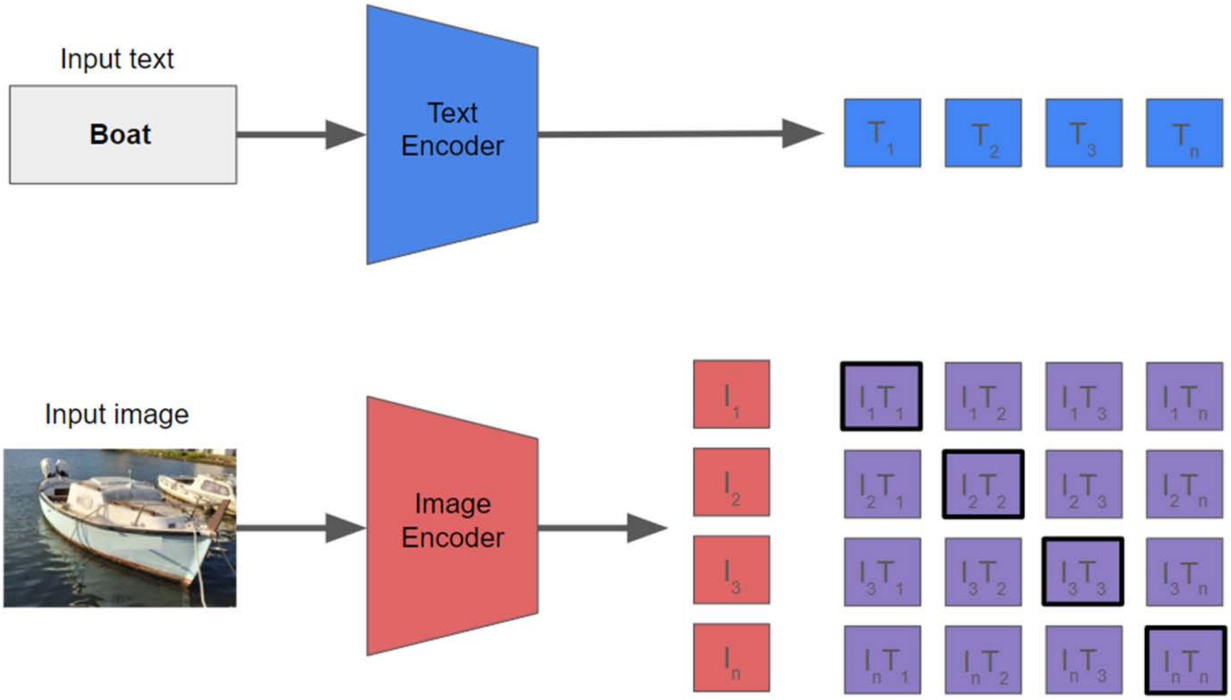


**Figure 2** Schematic representation of CLIP ViT model

tween images and their corresponding textual descriptions. CLIP leverages a large-scale dataset consisting of 400 million image-text pairs gathered from the internet, making it one of the most comprehensive pre-training datasets in the field of multi-modal machine learning [19]. The core of CLIP's architecture comprises two parallel neural networks: a Vision Transformer (ViT) and a Transformer-based text encoder [20]. The vision model processes images, while the text model processes corresponding descriptions. Both models project their respective inputs into a shared multi-dimensional space, enabling the system to align images with text based on their semantic similarity. During training, CLIP employs a contrastive loss function that maximizes the cosine similarity between matching image-text pairs while minimizing the similarity for non-matching pairs [21]. This approach enables CLIP to effectively learn robust, generalizable representations that capture the relationships between images and text. The vision encoder is typically a ViT model or a ResNet, pre-trained to extract high-level features from images [22]. The text encoder is a Transformer model, pre-trained on text-only corpora to understand the syntactic and semantic nuances of natural language. The alignment of these two modalities in a joint embedding space allows CLIP to perform tasks such as zero-shot image classification, where the model can classify images into categories without having seen labeled examples during training. This capability is achieved by comparing the embeddings of candidate labels expressed as text with the image embeddings, as presented in Figure 2.

One of the key innovations of CLIP is its ability to perform zero-shot learning across a wide range of vision tasks. By formulating classification problems as text-based prompts, CLIP can classify images without needing task-specific fine-tuning. For instance, given an image of a boat, the model can identify it as a "boat" by comparing the image's embedding with text embeddings of potential labels like "boat," "cargo ship," and "dock", as presented in Figure 3. This capability demonstrates the model's generalization power and its potential to be applied to diverse vision tasks with minimal adaptation [23]. Moreover, CLIP's transfer learning capabilities extend beyond classification. The model has been shown to perform well on tasks such as object detection, image segmentation, and visual question answering (VQA), even when transferred to these tasks without additional task-specific training. This versatility underscores the effectiveness of the shared representation space learned by CLIP and its utility in various applications where understanding the interplay between visual and textual information is crucial.

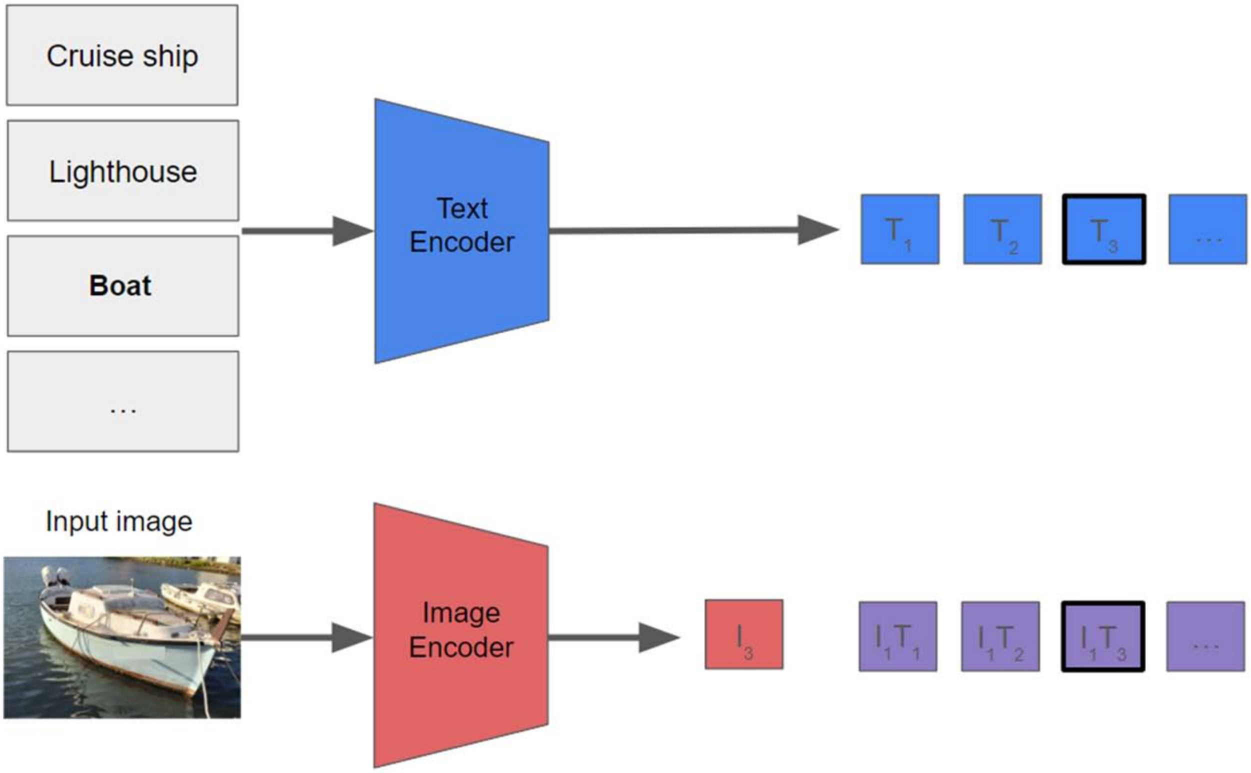


**Figure 3** Example of image classification using CLIP

CLIP's architecture and training methodology represent a significant advancement in the field of vision-language models. Its ability to generalize across tasks without fine-tuning opens new possibilities for AI systems to interact with the world in a more natural and flexible manner. Applications of CLIP span various domains, including content moderation, where the model can automatically identify inappropriate or harmful content based on textual descriptions, and in creative fields such as generating text-based image searches or guiding artistic creation through text prompts [24]. Additionally, CLIP's zero-shot capabilities make it particularly valuable in scenarios where labeled data is scarce or unavailable, offering a practical solution for deploying machine learning models in real-world environments with minimal human intervention [25].

In this research, four different CLIP architectures are used and compared, and these are:

- clip-vit-base-patch16,
- clip-vit-base-patch32,
- clip-vit-large-patch14, and
- clip-vit-large-patch14-336.

## 6 Results and discussion

The results for the "clip-vit-base-patch16" model across five classes—boat, cargo, cruise, dock, and lighthouse—are presented using AUC (Area Under the Curve) and F2 score metrics. The AUC values are very high across all classes, ranging from 0.988 for the "cruise" class to 0.999 for the "dock" and "lighthouse" classes, as presented in Figure 4. These values suggest that the model has a strong ability to correctly distinguish between true positive and false positive classifications across different categories. The F2 scores, which give more weight to recall than precision, show more variability. The scores range from 0.714 for the "cargo" class to 0.971 for the "lighthouse" class. This indicates that while the model performs well overall, there is some room for improvement, especially in the "cargo" class where the F2 score is significantly lower. This lower score suggests that the model might struggle more with recall or precision in this specific class, leading to more false negatives or false positives compared to other classes.

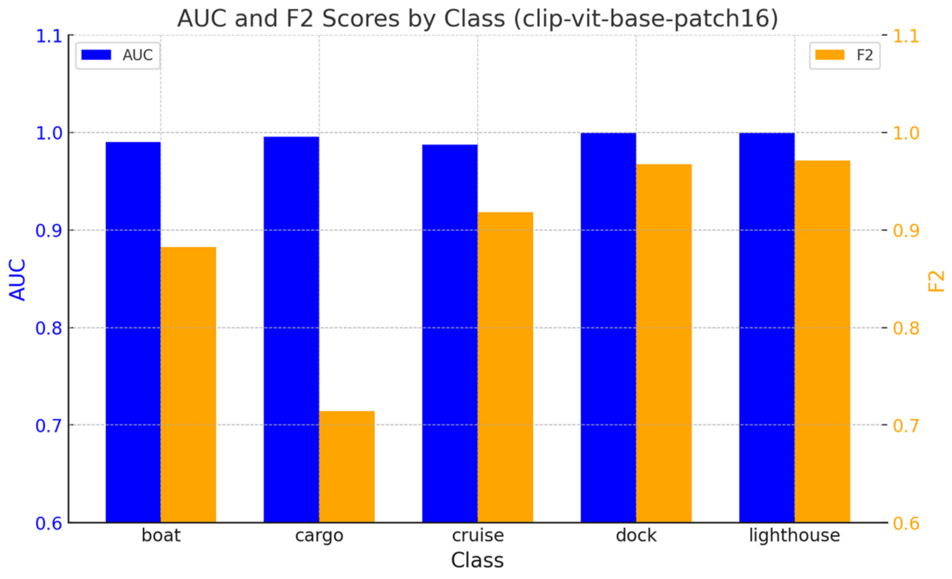


**Figure 4** AUC and F2 Scores for Various Classes Using the "clip-vit-base-patch16" Model

The results for the "clip-vit-base-patch32" model across five classes show that AUC values are consistently high across all classes, ranging from 0.991 for the "cargo" class to nearly 1.0 for the "lighthouse" class, as presented in Figure 5. This indicates that the model has an excellent ability to distinguish between true positives and false positives, effectively identifying correct classifications across different categories. However, the F2 scores reveal a more varied performance, with scores ranging from 0.620 for the "cargo" class to 0.970 for the "lighthouse" class. The lower F2 score for the "cargo" class suggests that the model may struggle with recall or precision in this category, leading to a higher incidence of either false negatives or false positives. On the other hand, the high F2 scores for the "boat," "cruise," "dock," and "lighthouse" classes demonstrate that the model performs well in these areas, balancing recall and precision effectively.

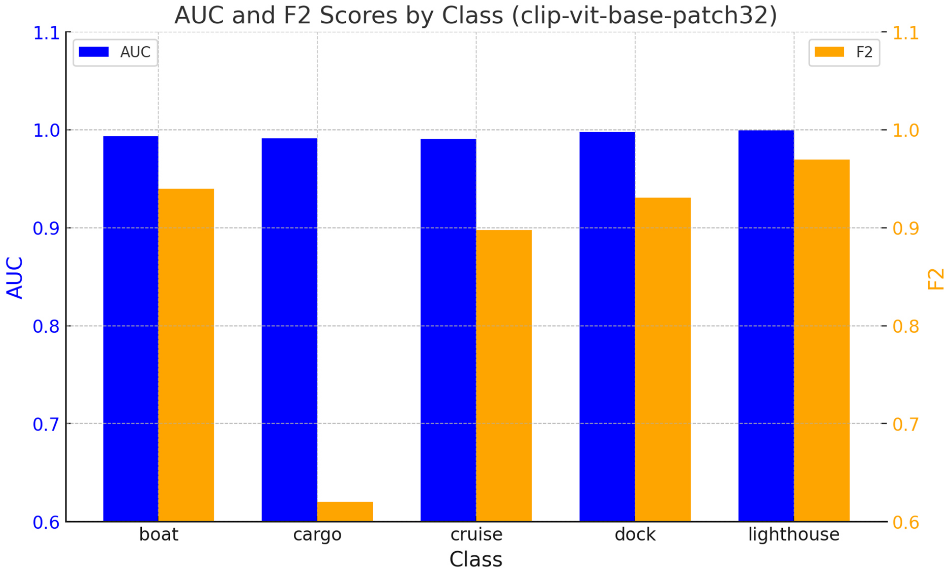


**Figure 5** AUC and F2 Scores for Various Classes Using the "clip-vit-base-patch32" Model

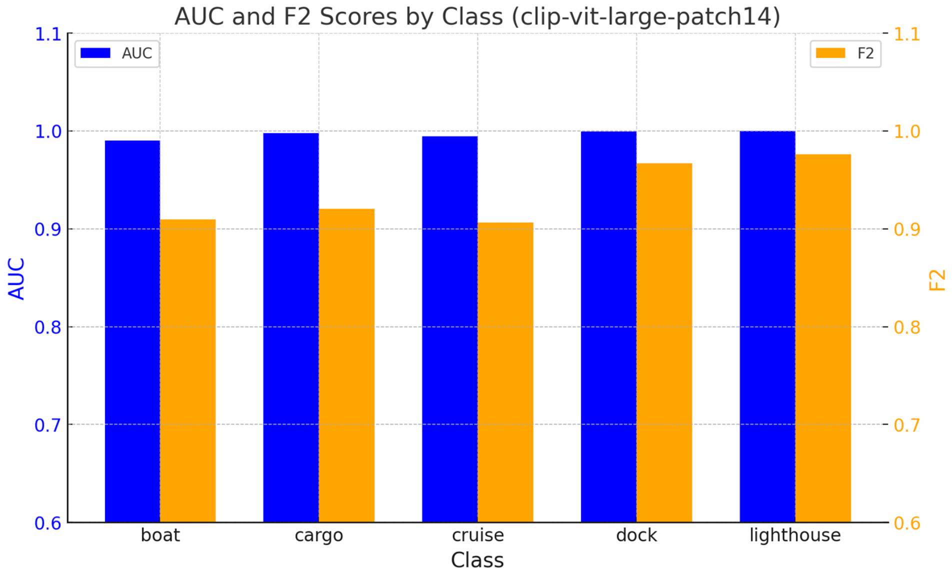


**Figure 6** AUC and F2 Scores for Various Classes Using the "clip-vit-large-patch14" Model

The AUC values are consistently close to 1 across all classes, indicating that the model has a very high capability to distinguish between the positive class (correct identification) and the negative class (incorrect identification). Specifically, the AUC values range from 0.990 for the "boat" class to 0.999 for the "lighthouse" class, suggesting that the model is highly effective in correctly classifying these classes. The F2 score, which places more emphasis on recall than precision, is also relatively high across all classes, though it exhibits slightly more variation than the AUC. The scores range from 0.906 for the "cruise" class to 0.976 for the "lighthouse" class, as presented in Figure 6. This indicates that while the model performs well in recalling true positives, there is some variability depending on the class, with the "cruise" class presenting slightly more challenges. Overall, the model demonstrates strong performance on both metrics, particularly in its ability to distinguish between different classes, as reflected by the high AUC values. The F2 scores suggest that the model maintains a good balance between precision and recall, with a slight emphasis on minimizing false negatives. The "lighthouse" class appears to be the easiest for the model to classify correctly, achieving the highest scores in both AUC and F2. In contrast, the "cruise" class, despite its high AUC, shows a slightly lower F2 score, indicating potential areas for improvement in recall or precision.

The results for the "clip-vit-large-patch14-336" model show that the AUC values are consistently high, ranging from 0.992 for the "boat" class to nearly 1.0 for the "lighthouse" class, as presented in Figure 7. These values indicate that the model has an exceptional ability to distinguish between correct and incorrect classifications across all classes, reflecting its overall strong discriminative power. The F2 scores are also notably high, with values ranging from 0.914 for the "boat" class to 0.976 for the "lighthouse" class. These scores suggest that the model is highly effective at maintaining a balance between precision and recall, with a particular emphasis on minimizing false negatives. The "lighthouse" class again stands out with the highest F2 score, indicating that it is the easiest class for the model to classify correctly. Meanwhile, the "boat" class, despite having a slightly lower F2 score, still performs well, showing a minor gap between precision and recall.

When the results are compared, we can see that the "clip-vit-large-patch14-336" model consistently achieves the highest average scores in both AUC and F2, as presented in Figure 8. Presented results are indicating superior overall performance across classification tasks. The AUC scores, represented in blue, show that all models are highly effective in distinguishing between true and false positives, with minimal variation. However, the F2 scores, shown in orange, reveal more variability, particularly in how well each model balances precision and recall. The "clip-vit-large-patch14-336" model stands out for its balanced performance, while the other models, such as "clip-vit-base-patch16" and "clip-vit-base-patch32," show lower average F2 scores, suggesting potential areas for improvement in recall or precision. This comparison highlights the relative strengths of each model, with "clip-vit-large-patch14-336" emerging as the most robust option.

While the CLIP models demonstrated strong performance in classifying maritime objects, several limitations and potential biases should be considered, particularly when applying these models to real-world maritime object detection tasks.

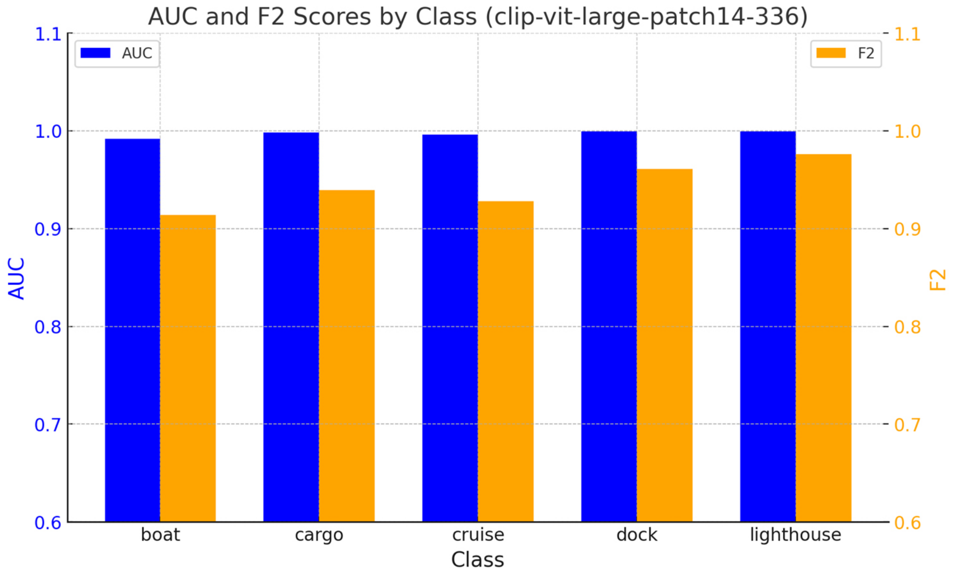


**Figure 7** AUC and F2 Scores for Various Classes Using the “clip-vit-large-patch14-336” Model

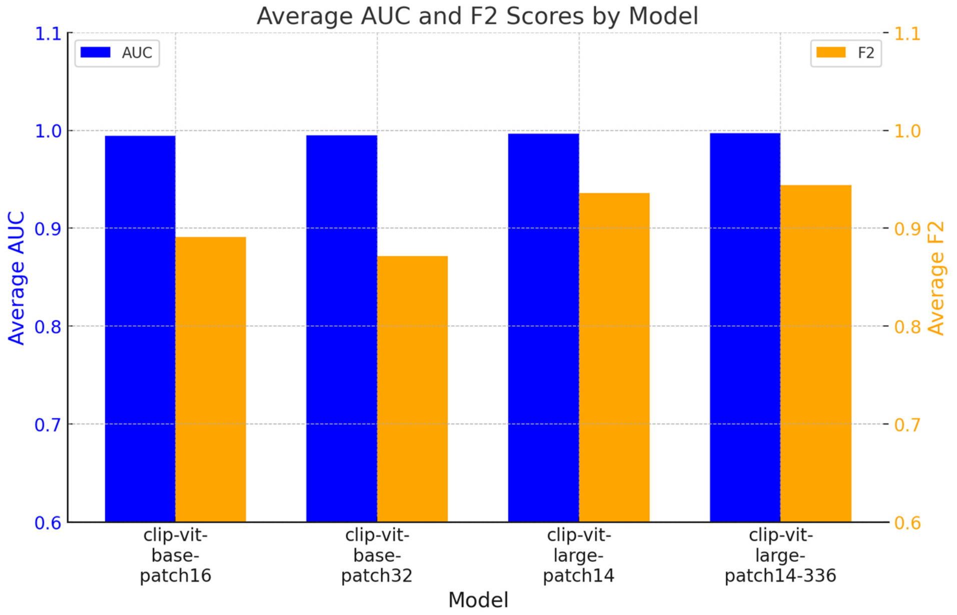


**Figure 8** Average AUC and F2 Scores by Model

The performance of CLIP models is heavily influenced by the quality and diversity of the dataset used for training and evaluation. The curated dataset in this study, while comprehensive, may still lack the full variability encountered in maritime environments. For instance, certain classes, such as “cruise” and “cargo,” may exhibit inherent biases due to the more frequent and visually distinct representation of these objects in available data. This could lead to a model that performs well on these more common classes but struggles with rarer or more ambiguous categories, such as small boats or floating debris, which might not be as well-represented.

Furthermore, the reliance on web-scraped images introduces another layer of potential bias. Images found online are often taken in favorable conditions and from aesthetically pleasing angles, which may not reflect the challenging real-world conditions of maritime environments, such as poor lighting, rough seas, or occlusions caused by other objects. This could limit the model’s robustness when deployed in less controlled settings.

The zero-shot learning capabilities of CLIP allow it to generalize to unseen classes, yet the performance can still be inconsistent when dealing with rare or ambiguous classes. For instance, the relatively lower F2 scores in classes like “cargo” suggest that the model may struggle with distinguishing these objects, especially when they share visual similarities with other classes (e.g., large boats vs. small cargo ships). The challenge lies in the semantic and visual overlap between classes, which could lead to misclassification or reduced confidence in the predictions for these less distinct categories.

## 7 Conclusions

The comparative analysis of the models—”clip-vit-base-patch16,” “clip-vit-base-patch32,” “clip-vit-large-patch14,” and “clip-vit-large-patch14-336”—reveals several key insights into their performance across various classes, namely boat, cargo, cruise, dock, and lighthouse, using AUC and F2 score metrics. The AUC values across all models are consistently high, particularly in the “dock” and “lighthouse” classes, where they approach or reach 1.0. This indicates that all models exhibit a strong ability to correctly distinguish between true positive and false positive classifications, highlighting their effectiveness in identifying correct classifications across different categories. However, when examining the F2 scores, which emphasize the balance between precision and recall, more variability is observed. The “clip-vit-base-patch16” model, while performing well in terms of AUC, shows significant variability in F2 scores, particularly in the “cargo” class, where the F2 score drops to 0.714. This suggests that this model may struggle with either recall or precision in certain categories, leading to a higher incidence of false negatives or false positives. The “clip-vit-base-patch32” model similarly demonstrates high AUC values but exhibits even greater variability in F2 scores, with the “cargo” class again being a weak point. The F2 score of 0.620 for this class indicates a notable challenge in maintaining an effective balance between recall and precision. On the other hand, the “clip-vit-large-patch14” model performs more consistently, with both AUC and F2 scores being high across most classes. Despite a slightly lower F2 score in the “cruise” class, this model shows a good balance between precision and recall, with minimal gaps in performance across different categories. The “clip-vit-large-patch14-336” model emerges as the strongest performer among the four, consistently achieving the highest average scores in both AUC and F2. This model not only excels in distinguishing between true and false positives but also maintains a robust balance between precision and recall across all classes, with particularly high scores in the “lighthouse” class. When the results are compared, it is evident that the “clip-vit-large-patch14-336” model offers the most balanced and reliable performance, making it the most suitable option for tasks that require high accuracy and a strong balance between recall and precision. The other models, while effective in certain areas, show varying degrees of performance across different classes, indicating potential areas for refinement. The findings suggest that for applications where both AUC and F2 are critical metrics, the “clip-vit-large-patch14-336” model would be the preferred choice.

One important direction is improving model robustness in real-world maritime environments, particularly by expanding the dataset to include more diverse and challenging conditions, such as poor lighting, occlusions, and rough seas, which are not well-represented in the current dataset. Another valuable research avenue is the exploration of hybrid models that combine ZSL with semi-supervised or few-shot learning techniques, potentially reducing the performance variability in rare or ambiguous classes like “cargo” or “small boats.” Additionally, future studies could focus on refining the textual embeddings to better capture the semantic differences between visually similar classes, thus improving classification accuracy. Finally, research could investigate the deployment of such models in real-time systems, addressing computational challenges and the scalability required for maritime surveillance and security applications. These advancements could further enhance the applicability of zero-shot learning in the maritime domain.

**Funding:** The research presented in the manuscript did not receive any external funding.

**Author Contributions:** Research, Ivan Lorencin; Writing, Ivan Lorencin and Domagoj Frank; Data Collection, Damir Vusić; Methodology, Domagoj Frank; Formal Analysis, Domagoj Frank and Damir Vusić; Data Curation, Damir Vusić; Supervision, Domagoj Frank; Validation, Domagoj Frank; Review and Editing, Domagoj Frank and Damir Vusić; Final Approval, Ivan Lorencin.

## References

[1] Wang, N., Wang, Y., & Er, M. J. (2022). Review on deep learning techniques for marine object recognition: Architectures and algorithms. Control Engineering Practice, 118, 104458. (DOI: 10.1016/j.conengprac.2020.104458)

[2] Lorencin, I., Anđelić, N., Mrzljak, V., & Car, Z. (2019). Marine objects recognition using convolutional neural networks. NAŠE MORE: znanstveni časopis za more i pomorstvo, 66(3), 112-119. (DOI: https://doi.org/10.17818/NM/2019/3.3)

[3] Marin, I., Mladenović, S., Gotovac, S., & Zaharija, G. (2021). Deep-feature-based approach to marine debris classification. Applied Sciences, 11(12), 5644. (DOI: https://doi.org/10.3390/app11125644)

[4] Duarte, M. M., & Azevedo, L. (2023). Automatic detection and identification of floating marine debris using

multispectral satellite imagery. IEEE Transactions on Geoscience and Remote Sensing, 61, 1-15. (DOI: 10.1109/TGRS.2023.3283607)

[5] Nixon, M., & Aguado, A. (2019). Feature extraction and image processing for computer vision. Academic press.

[6] Uemura, T., Lu, H., & Kim, H. (2020). Marine organisms tracking and recognizing using yolo. In 2nd EAI International Conference on Robotic Sensor Networks: ROSENET 2018 (pp. 53-58). Springer International Publishing. (DOI: 10.1007/978-3-030-17763-8_6)

[7] Zhong, J., Li, M., Qin, J., Cui, Y., Yang, K., & Zhang, H. (2022). Real-time marine animal detection using YOLO-based deep learning networks in the coral reef ecosystem. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 46, 301-306. (DOI: 10.5194/isprs-archives-XLVI-3-W1-2022-301-2022)

[8] Glučina, M., Anđelić, N., Lorencin, I., & Car, Z. (2023). Detection and classification of printed circuit boards using YOLO algorithm. Electronics, 12(3), 667. (DOI: https://doi.org/10.3390/electronics12030667)

[9] Gašparović, B., Lerga, J., Mauša, G., & Ivašić-Kos, M. (2022). Deep learning approach for objects detection in underwater pipeline images. *Applied artificial intelligence*, *36*(1), 2146853.

[10] Mihel, A. M., Lerga, J., & Krvavica, N. (2024). Estimating water levels and discharges in tidal rivers and estuaries: Review of machine learning approaches. *Environmental modelling & software*, 106033.

[11] Pourpanah, F., Abdar, M., Luo, Y., Zhou, X., Wang, R., Lim, C. P., ... & Wu, Q. J. (2022). A review of generalized zero-shot learning methods. IEEE transactions on pattern analysis and machine intelligence, 45(4), 4051-4070.

[12] Bucher, M., Herbin, S., & Jurie, F. (2017). Generating visual representations for zero-shot classification. In Proceedings of the IEEE International Conference on Computer Vision Workshops (pp. 2666-2673)

[13] Vilas, M. G., Schaumlöffel, T., & Roig, G. (2024). Analyzing Vision Transformers for image classification in class embedding space. Advances in neural information processing systems, 36.

[14] Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3156-3164).

[15] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2015). Vqa: Visual question answering. In Proceedings of the IEEE international conference on computer vision (pp. 2425-2433).

[16] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In International conference on machine learning (pp. 8748-8763). PMLR.

[17] Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in neural information processing systems, 32.

[18] Guo, J., Rao, Z., Guo, S., Zhou, J., & Tao, D. (2024). Fine-grained zero-shot learning: Advances, challenges, and prospects. arXiv preprint arXiv:2401.17766.

[19] Vishniakov, K., Shen, Z., & Liu, Z. (2023). Convnet vs transformer, supervised vs clip: Beyond imagenet accuracy. arXiv preprint arXiv:2311.09215.

[20] Wang, S., Yan, Y., Yang, X., & Huang, K. (2023, March). CRA: Text to Image Retrieval for Architecture Images by Chinese CLIP. In 2023 7th International Conference on Machine Vision and Information Technology (CMVIT) (pp. 29-34). IEEE. (DOI: 10.1109/CMVIT57620.2023.00015)

[21] Zhang, H., Cheng, D., Jiang, H., Liu, J., & Kou, Q. (2024). Task-like training paradigm in CLIP for zero-shot sketch-based image retrieval. Multimedia Tools and Applications, 83(19), 57811-57828. (DOI: 10.1007/s11042-023-17675-x)

[22] Tu, W., Deng, W., & Gedeon, T. (2024). A closer look at the robustness of contrastive language-image pre-training (clip). Advances in Neural Information Processing Systems, 36.

[23] Lin, Z., Geng, S., Zhang, R., Gao, P., De Melo, G., Wang, X., ... & Li, H. (2022, October). Frozen clip models are efficient video learners. In European Conference on Computer Vision (pp. 388-404). Cham: Springer Nature Switzerland.

[24] Edstedt, J., Berg, A., Felsberg, M., Karlsson, J., Benavente, F., Novak, A., & Pihlgren, G. G. (2022, August). Vidharm: A clip based dataset for harmful content detection. In 2022 26th International Conference on Pattern Recognition (ICPR) (pp. 1543-1549). IEEE.

[25] Sun, Z., Fang, Y., Wu, T., Zhang, P., Zang, Y., Kong, S., ... & Wang, J. (2024). Alpha-clip: A clip model focusing on wherever you want. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 13019-13029).