Multimodal Recognition Methods for Maritime Vessel Identification in Complex Scenarios

Qiuyu Tian[1,*], Kun Wang[2], Hongwei Tang[1,3], Rui Zhu[2]
[1]Institute of Information Superbahn, Nanjing, Jiangsu, China
[2]Institute of Computing Technology Chinese Academy of Sciences, Beijing, China
[3]University of Chinese Academy of Sciences, Nanjing, Jiangsu, China
**Corresponding Author.*

Abstract: Maritime vessel recognition is a crucial task in maritime monitoring and traffic management, supporting various applications such as vessel tracking, operational safety, and anomaly detection. Traditional vessel detection and recognition processes rely heavily on manual inspection, which is constrained by environmental noise, low visibility, and the need for real-time performance, making high accuracy and reliability difficult to achieve. This study develops a multimodal classification method that effectively integrates image data with vessel identification numbers to improve the accuracy of automated vessel recognition. A comprehensive recognition system, integrating vessel detection and classification, has been successfully developed and deployed in a maritime monitoring center. With the inclusion of vessel identification numbers as auxiliary data, the system achieved an accuracy of 89% in practical applications. The innovation of this research lies in the application of a multimodal model to address the challenges of vessel recognition, significantly enhancing recognition accuracy and establishing an intelligent recognition system suitable for real-world maritime monitoring and analysis.

Keywords: Maritime Vessel Recognition; Multimodal Image Classification; Image Augmentation; Maritime Intelligence Analysis

1. Introduction

Accurately identifying maritime vessels is a critical task in maritime operations and intelligence gathering. Beyond its importance in assessing maritime security, vessel recognition also plays a key role in maritime traffic management and the protection of marine resources. However, the inherent complexities of the maritime environment—such as variations in lighting conditions, adverse weather, wave interference, and long-range imaging—pose significant challenges for recognizing vessels from real-world imagery. These challenges are further compounded by issues such as low image resolution and high levels of noise, which can degrade the performance of recognition systems.

In recent years, machine learning techniques, particularly convolutional neural networks (CNNs), have achieved significant advancements in image recognition tasks. While these methods have demonstrated success in controlled environments, their reliance on single-modality data, such as image-only inputs, limits their effectiveness in real-world scenarios. Maritime vessel recognition, in particular, suffers from the drawbacks of such approaches when faced with noisy, low-resolution, or incomplete data captured in complex maritime environments.

To address these limitations, this study proposes an innovative multimodal approach to maritime vessel recognition. By integrating diverse data sources, such as visual imagery and hull number information, the proposed method enhances recognition accuracy under challenging conditions. The approach also leverages advanced image processing techniques to improve the quality of real-world maritime images, enabling the model to extract more robust features from noisy data. The combination of multimodal data and enhanced image processing ensures that the system performs reliably, even in adverse environments.

This paper introduces the methodology and technical framework of the proposed multimodal recognition system, emphasizing

its potential to overcome the challenges of traditional single-modality approaches. The effectiveness of the method is demonstrated through its deployment in real-world scenarios, where it significantly improves the accuracy and reliability of maritime vessel recognition. By providing a practical and precise automated recognition solution, this work contributes to advancing the field of maritime intelligence gathering and maritime security monitoring, offering a robust foundation for future developments in complex maritime environments.

2. Related Works

2.1 CNN-Based Classification Model

In the field of automatic vessel recognition, image classification and object detection technologies play a crucial role. Convolutional Neural Networks (CNNs) have become the core technology for image classification tasks [1]. Leveraging their deep structures and complex feature extraction capabilities, CNNs have demonstrated remarkable performance in classifying maritime vessel images captured in complex environments. Notable architectures such as AlexNet [2] and VGGNet [3] have achieved significant success in image classification tasks.

Moreover, multimodal classifiers [4] are increasingly recognized for their value in automatic vessel recognition. These methods integrate information from multiple data sources, such as visual images, textual labels, or other sensor data, providing a more comprehensive and accurate perspective for classification. In vessel recognition tasks, multimodal approaches enhance classification accuracy and robustness by combining visual features of images with key information such as hull numbers.

2.2 Object Detection Techniques

This study employs two advanced technologies at different stages of the automatic vessel recognition process: YOLO (You Only Look Once) [5] and Grounding DINO [6]. YOLO, a highly efficient real-time object detection method known for its speed and accuracy, is primarily utilized during the recognition process. It directly predicts the bounding boxes and class probabilities of targets from input images, making it particularly suited for handling maritime images with multiple objects or complex backgrounds. YOLO is used in the initial stage of the recognition pipeline to determine whether vessel targets exist in the image, laying the foundation for subsequent vessel classification tasks.

In the automated construction of training datasets, where the image quality is relatively high, zero-shot learning techniques that require no additional training are employed. Grounding DINO, a zero-shot object detection model, excels in recognizing objects beyond the training set by leveraging its deep understanding of language and visual content. This study uses Grounding DINO primarily for automated annotation and classification of training set images, particularly for filtering and identifying complex objects in maritime environments.

2.3 Image Quality Enhancement

Image clarity has a significant impact on classification accuracy in the field of automatic vessel recognition. The complexities of the maritime environment, such as weather conditions (e.g., fog, rain) and lighting issues (e.g., shadows, backlighting), as well as inherent image noise, motion blur, or low resolution, often degrade image quality. These factors adversely affect feature recognition and classification. Therefore, applying image quality enhancement techniques to mitigate these challenges is essential before image classification.

Image enhancement methods include denoising, deblurring, and super-resolution techniques. Over the past few decades, these methods have evolved from traditional filter-based approaches to deep learning-based techniques. For example, autoencoders (AE) [7] and Generative Adversarial Networks (GANs) [8] have achieved significant success in image deblurring and denoising.

This study utilizes Real-ESRGAN (Real Enhanced Super-Resolution Generative Adversarial Network) [9], which demonstrates superior performance in super-resolution tasks and is particularly suitable for processing vessel images. Real-ESRGAN enhances the resolution of images through deep learning, making fine details more discernible. This capability is crucial for processing vessel images, as such images often contain structural features and intricate details that are critical for

accurate vessel recognition. By improving the quality of these images, Real-ESRGAN provides a clearer visual foundation for subsequent object detection and classification.

2.4 Semantic Segmentation

Semantic segmentation is a key task in computer vision that aims to classify each pixel in an image into different object categories. Unlike traditional object detection, which identifies entire objects, semantic segmentation focuses on accurately delineating and distinguishing the boundaries of objects at the pixel level. In this domain, the U-Net model [10] has been widely applied to various image segmentation tasks due to its exceptional performance.

In recent years, the Segment Anything Model (SAM) [11] has brought innovative advancements to image segmentation tasks. With its efficiency and versatility, SAM can precisely segment objects in any image. Trained on extensive datasets, SAM possesses powerful recognition capabilities, enabling it to accurately identify vessel boundaries in diverse scenarios.

In this study, SAM is used in conjunction with object detection models like Grounding DINO and YOLO to accurately extract the hull regions in vessel images. The combination of SAM with Grounding DINO, referred to as the Grounded-Segment-Anything framework, leverages text prompts to achieve zero-shot object detection and semantic segmentation. This integrated approach effectively filters out background noise, such as waves, ripples, and weather interferences, enhancing the accuracy and robustness of vessel classification tasks. By combining SAM with object detection models, this study not only improves segmentation accuracy but also provides a clearer and more reliable foundation for subsequent vessel recognition and classification tasks.

3. Dataset Description

The quality and diversity of a dataset directly impact the performance of recognition algorithms. Larger and more comprehensive datasets tend to enhance the accuracy and robustness of machine learning models. However, there is no publicly available dataset that categorizes vessel images based on country, type, specific model, or hull number. To address this gap, this study constructed a vessel knowledge base and a corresponding image dataset. By systematically combining information from multiple sources and employing advanced data collection techniques, the resulting dataset provides a detailed and reliable foundation for vessel recognition and classification tasks.

3.1 Construction of the Training Dataset

3.1.1 Comprehensive maritime vessel knowledge base

This study established a comprehensive knowledge base containing information on 205 distinct maritime vessel classes. The database encompasses a wide range of vessel categories, including surface combatants, aircraft carriers, amphibious ships, and electronic reconnaissance vessels. Detailed records of subclassifications, designations, hull numbers, and service statuses were included to ensure thorough data representation.

The knowledge base was developed through extensive analysis of 205 active vessel models and incorporated insights from historical and potentially relevant vessel models worldwide. Based on this analysis, 101 representative vessel models were selected to construct the training dataset. This process was guided by the need to create a robust foundation for future research and technical development.

3.1.2 Image collection and filtering

To support the knowledge base and subsequent applications, a web scraping technique was employed to gather vessel images from various sources, including popular search engines like Google and Bing. The goal was to acquire a diverse set of high-quality images suitable for tasks such as vessel recognition and classification.

During the image collection process, a stringent filtering mechanism was applied. Only high-quality images that clearly and completely depicted the side view of vessels were retained. Images missing critical portions of the hull, showing only top-down perspectives, or unrelated to the research objectives were excluded. This rigorous selection process ensured that the dataset contained precise and reliable data, enabling effective training of automatic recognition and classification algorithms.

3.2 Testing Dataset

The testing dataset used in this study consists of 2,482 real-world images of maritime vessels captured in open water. Each image is accompanied by metadata, including hull numbers, geographic coordinates, and vessel classification labels. These labels provide critical information for image analysis and vessel recognition tasks.

The images in the testing dataset predominantly feature side views of vessels captured under challenging maritime conditions. Issues such as low image clarity, lighting variations, and long-distance perspectives present significant challenges for recognition tasks. Additionally, some vessel colors closely match the blue hues of the ocean and sky, further complicating the visual distinction between vessels and their backgrounds.

This testing dataset provides a realistic and challenging application scenario for evaluating recognition models. It underscores the importance of robust image processing and recognition techniques capable of overcoming low-quality visuals and high background similarity. Successfully identifying vessels under such conditions demonstrates the practical applicability of the developed recognition system and highlights the critical role of a high-quality knowledge base and dataset in advancing maritime monitoring and intelligence capabilities.

4. Methodology

This section describes the methodology employed to optimize product assortment, encapsulated within a framework that is referred to as the Discrete Choice Model (DCM) [10]. This model comprises four essential components: Substitution Group Learning, Within Group Demand Model, Cross Group Complementarity Model, and Constrained Assortment Optimization. Together, these modules form an integrated approach to understanding customer behavior and making optimal product selection decisions. Each module is introduced below, with a more detailed explanation provided in the appendices.

4.1 System Architecture

This study proposes a comprehensive vessel classification system designed to achieve high accuracy and efficiency in vessel recognition and classification. As shown in Figure 1, the system consists of four key components, each performing specific functions to ensure a seamless workflow:

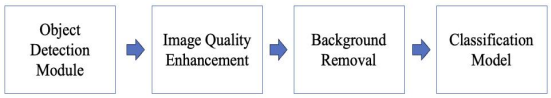

Figure 1. System Architecture

(1) Target detection module

This module employs the YOLO algorithm to detect vessels in real-world images. When a vessel is detected, the system extracts the bounding box region containing the vessel. If no target is identified, the module returns a "No Target" result. This step forms the foundation for subsequent processing, ensuring that only regions containing potential targets are analyzed further.

(2) Image quality enhancement module

The extracted regions from the target detection module are processed using the Real-ESRGAN technique to enhance image clarity. Improving the quality of these images is crucial for ensuring the effectiveness and reliability of the subsequent background segmentation module.

(3) Background segmentation module

This module leverages semantic segmentation models, such as the Segment Anything Model (SAM), to accurately segment the hull of the vessel from the image. The goal of this step is to filter out non-target elements like waves and the sky, thereby improving the precision and effectiveness of the classification model.

(4) Vessel classification model

Finally, the system utilizes a multimodal classification model with dual input streams (image and hull number) to categorize the vessel. Based on the input data, the model returns the probabilities for each vessel category. In practical applications, the system outputs the top three categories with the highest probabilities.

This architecture integrates advanced techniques in object detection, image enhancement, semantic segmentation, and multimodal classification to create a robust and reliable system for automatic vessel recognition and classification.

4.2 Image Processing and Enhancement

4.2.1 Image augmentation during training

(1) Image segmentation

Given the high quality of training images and the prominence of vessels within these images,

this study employed the Grounded-Segment-Anything framework for semantic segmentation. This framework integrates the zero-shot object detection model GroundingDINO, eliminating the need for complex data preprocessing or additional training. By using "ship" as the prompt to activate the GroundingDINO model, vessels were effectively detected. Subsequently, the Segment Anything Model (SAM) was applied for precise and rapid segmentation, generating accurate masks to filter out non-relevant elements such as oceans, waves, and skies, leaving only the hull of the vessel. This process enhances vessel recognition accuracy by reducing environmental noise, as illustrated in Figure 2.

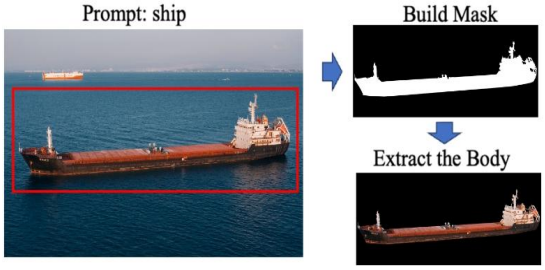

Figure 2. Prompt-Driven Image Segmentation Process

(2) Blur simulation

Real-world maritime images often exhibit varying degrees of clarity due to external factors such as distance or weather conditions. To simulate these real-world scenarios, extracted vessel images underwent median blur augmentation with varying kernel sizes (e.g., 3×3, 7×7, up to 31×31 pixels). This method, illustrated in Figure 3, enhances the model's robustness to real-world conditions, particularly for long-distance or weather-affected images.

Figure 3. Examples of Blur Effects with Different Kernel Sizes

(3) Image padding

For neural network training, input images must be standardized to a square format. Direct resizing can distort vessel proportions, negatively affecting classification accuracy. To preserve the original proportions of the vessel, this study employed image padding. Vessel-only images were padded at the edges to create square images while maintaining the vessel's visual integrity, as demonstrated in Figure 4.

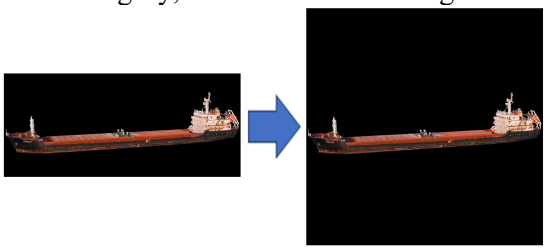

Figure 4. Example of Image Padding

(4) Additional augmentation techniques

The training process incorporated additional augmentations, including random affine transformations such as slight rotations and translations to simulate perspective variations. Color adjustments, such as modifications to brightness, contrast, saturation, and hue, simulated diverse lighting conditions. Random horizontal flipping was also applied to improve robustness to directional changes. These combined techniques significantly enhanced the model's performance in diverse maritime environments.

4.2.2 Image processing during recognition

For real-world vessel recognition tasks, a series of key image processing steps were implemented, including contrast adjustment, target detection, image enhancement, segmentation, and padding. To address the challenges of real-world test images, the previously unified Grounded-Segment-Anything framework was split into separate modules for target detection and semantic segmentation.

(1) Contrast adjustment

To address the common challenge of visual similarity between vessels and their ocean or sky backgrounds, contrast adjustment was applied to enhance the color difference. This step facilitates clearer differentiation of the vessel from its background during subsequent recognition stages.

(2) Target detection and bounding box extraction

The YOLOv8 model, specifically enhanced with Coordinate Attention (CA), was employed for vessel detection. This model was fine-tuned using a training dataset of 3,277 images, with 100 representative images carefully annotated with bounding boxes to provide high-quality supervision. These annotations captured vessels in various real-world scenarios, including challenging

conditions like cluttered backgrounds and low visibility.

During the detection process, YOLOv8-CA identified vessel locations within images and extracted bounding box regions containing the hulls. This step ensured that only relevant portions of the images were passed on for subsequent processing. The integration of the CA mechanism significantly improved the model's ability to focus on vessel features amidst distracting maritime elements such as waves, skies, and reflections. An example detection is illustrated in Figure 5, showing the precise localization of vessels under real-world conditions.

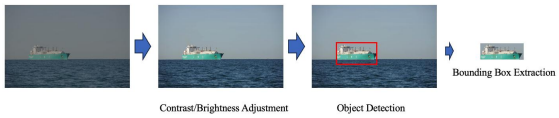

Figure 5. Object Detection Module (Contrast Adjustment and Target Detection)

(3) Image enhancement

Once the bounding box regions were extracted, the Real-ESRGAN super-resolution model was applied to address the common issue of vessels occupying small, low-resolution regions within images. This model enhanced the clarity of these regions by reconstructing finer details, such as structural lines and edges, while resizing the longest edge to 512 pixels. This resizing ensured a standardized input size suitable for subsequent segmentation and classification tasks, maintaining the balance between computational efficiency and image quality.

The enhanced images were then processed using the Segment Anything Model (SAM). SAM utilized semantic segmentation to distinguish vessels from background elements with high precision. This step involved generating hull masks that excluded irrelevant features like waves, skies, and reflections, isolating the vessel for detailed analysis. These masks provided a clean and accurate representation of the vessel, ensuring the effectiveness of downstream classification tasks.

(4) Image padding

As illustrated in Figure 6, padding was applied to maintain the vessel's original proportions, ensuring the preservation of key visual features, after segmentation. This padding strategy significantly improved the overall accuracy and robustness of the recognition process.

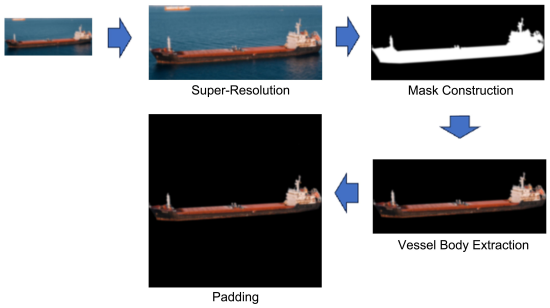

Figure 6. Semantic Segmentation Module and Subsequent Operations

These techniques collectively form a robust image processing pipeline, ensuring that the recognition system can handle diverse real-world challenges while maintaining high accuracy and efficiency.

4.3 Multimodal Classification Model

The multimodal classification model is the final stage in the vessel recognition framework, designed to integrate image features with hull number information to enhance classification accuracy. The overall model structure is shown in Figure 7. This model effectively addresses challenges such as low-resolution images and incomplete data by combining visual and textual inputs through a carefully structured pipeline. The main components of the model include the following:

(1) Image feature extraction

The model uses a ResNeXt architecture enhanced with the Convolutional Block Attention Module (CBAM). This component captures high-level features from vessel images, with CBAM focusing on critical regions, such as vessel structures, while suppressing irrelevant background elements like waves or skies.

(2) Hull number embedding

Hull numbers are processed through an embedding layer that maps alphanumeric sequences into a high-dimensional vector space. This representation captures the semantic and structural information of the hull numbers, providing auxiliary input to complement the image features.

(3) Feature fusion

The extracted image features and hull number embeddings are concatenated into a unified vector representation. This fusion step ensures that the model can simultaneously leverage both modalities, improving robustness even when one modality (e.g., image data) is less informative due to noise or low resolution.

(4) Classification layer

The unified feature vector is passed through a fully connected layer that outputs probabilities for each vessel category. In practical scenarios, the model also provides the top-3 predictions, accommodating cases where multiple classifications are relevant.

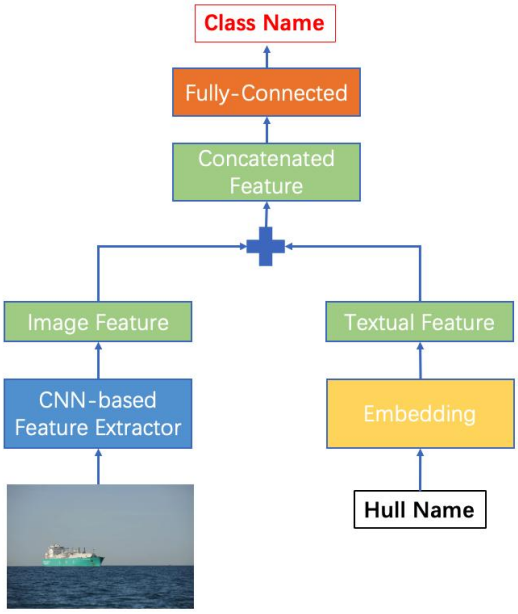

Figure 7. Multi-Modal Classification Model

This design enables the multimodal classification model to achieve superior performance by leveraging the complementary strengths of image features and hull number information. Experimental results in Section 5 validate the effectiveness of this approach, showing significant accuracy improvements when incorporating CBAM and hull number embeddings, particularly in challenging maritime environments.

5. Experiments and Results

This section presents the experimental setup, training details, and results of the proposed framework for vessel detection and classification. The evaluation focuses on the YOLOv8 detection models and multimodal classification models, with detailed analysis of the impact of attention mechanisms and super-resolution techniques.

5.1 Training Details

The training process utilized the datasets described in Section 4.1, comprising 3,277 high-resolution images for training and 655 low-resolution real-world images for testing. Images were preprocessed using Real-ESRGAN for super-resolution, YOLO-CA for object detection, and SAM for background segmentation. Hull numbers were either manually annotated or randomly selected based on the vessel knowledge base.

(1) Data augmentation

To enhance the generalization capability of the models, various augmentation techniques were applied, including random horizontal flipping, cropping, rotation, brightness adjustment, saturation adjustment, and Gaussian noise. All images were resized to a fixed 512×512 resolution for consistency in training.

(2) Optimization and loss function

The Adam optimizer was employed for both the detection and classification models, with a learning rate scheduler ensuring stable convergence. Cross-entropy loss was used for classification tasks, incorporating label smoothing to prevent overfitting.

(3) Pretraining and transfer learning

The YOLOv8 models were initialized with pre-trained weights from the COCO dataset, fine-tuned on the vessel detection dataset to adapt to the specific requirements of maritime scenarios.

5.2 Object Detection Results

The performance of the YOLOv8 models was evaluated using precision, recall, and mean average precision (mAP) metrics. Table 1 summarizes the results of the YOLOv8L and YOLOv8X models, with and without the inclusion of the Coordinate Attention (CA) mechanism.

Table 1. YOLOv8 Performance with and without Coordinate Attention

Model	Precision (%)	Recall (%)	mAP@0.5 (%)
YOLOv8L	90.5	88.7	90.1
YOLOv8L-CA	91.6	90.4	91.1
YOLOv8X	94.2	93.3	95.4
YOLOv8X-CA	95.9	93.4	98.3

The YOLOv8X-CA model achieved the highest detection performance, demonstrating the effectiveness of the CA mechanism in improving vessel localization and feature extraction. The incorporation of CA resulted in notable improvements in mAP, particularly under challenging conditions such as complex backgrounds.

5.3 Classification Results

The classification performance of three

ResNeXt-based models was evaluated:
(1) ResNeXt**:** A baseline model for image classification.
(2) ResNeXt-CBAM**:** Incorporates the CBAM attention mechanism to focus on relevant spatial and channel information.
(3) ResNeXt-CBAM-Hull**:** Enhances ResNeXt-CBAM with an additional hull number input, utilizing multimodal data for classification.
Table 2 compares the accuracy of these models on the test set.

Table 2. Classification Model Performance

Model	Test Accuracy (%)
ResNeXt	58.2
ResNeXt-CBAM	61.6
ResNeXt-CBAM-Hull	91.6

The results indicate that the CBAM module improves feature extraction, boosting accuracy by 3.4% compared to the ResNeXt baseline. The integration of hull number embeddings in the ResNeXt-CBAM-Hull model further enhances accuracy, achieving a significant improvement of 30% over ResNeXt. This underscores the effectiveness of multimodal data in distinguishing vessels with similar visual features but different textual identifiers.

5.4 Impact of Super-Resolution

To investigate the impact of super-resolution, the Real-ESRGAN model was applied to the test set images prior to classification. The results, presented in Table 3, show consistent improvements across all models, with the ResNeXt-CBAM-Hull model benefiting the most.

Table 3. Performance on Super-Resolution Test Set

Model	W/O Super-Resolution Accuracy (%)	With Super-Resolution Accuracy (%)
YOLOv8-CA	17.3	21.4
ResNeXt	58.2	62.1
ResNeXt-CBAM	61.6	65.7
ResNeXt-CBAM-Hull	91.6	97.4

The ResNeXt-CBAM-Hull model achieved an accuracy of 97.4% on the super-resolution test set, demonstrating that enhancing image quality can significantly improve classification performance, particularly in low-resolution scenarios. This result highlights the value of integrating advanced image processing techniques into the vessel recognition pipeline.

6. Conclusion

In this study, the integration of visual and textual information, particularly through the use of hull numbers, has demonstrated significant potential for multimodal classification models in vessel recognition tasks. Despite the challenges posed by image quality and incomplete information, the proposed model has achieved notable advancements in recognition accuracy. Compared to existing research, this study offers a new theoretical perspective and practical evidence of its application value.
Future research will focus on continuously updating and expanding the maritime vessel knowledge base while increasing the diversity of vessel types in the training dataset to enhance the model's generalization capability. Optimizing the model architecture, particularly by incorporating advanced attention mechanisms, will further improve its ability to capture critical features. Additionally, advancements in image enhancement techniques will aim to address low-resolution and high-noise maritime images more effectively.
Exploration of the model's broader applications in maritime surveillance tasks will also be a key area of focus, especially in enhancing real-time data processing capabilities. These improvements are expected to drive significant technological innovation and performance enhancement in the field of maritime monitoring, providing more effective solutions to the complexities and challenges of oceanic environments.

Acknowledgements

This research was supported by The Research of Key Technologies and Industrialization for Intelligent Computing in Large-Scale Video Scenarios under Digital Social Governance of Henan, China (grant number 241100210100).

Reference

[1] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521 (7553), 436–444.
[2] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with

deep convolutional neural networks. Advances in Neural Information Processing Systems, 25.

[3] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. International Conference on Learning Representations (ICLR), 2015.

[4] Sleeman IV, W. C., Kapoor, R., & Ghosh, P. (2022). Multimodal classification: Current landscape, taxonomy and future directions. ACM Computing Surveys, 55 (7), 1–31.

[5] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 779–788).

[6] Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., ... & Zhang, L. (2024, September). Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In European Conference on Computer Vision (pp. 38–55). Cham: Springer Nature Switzerland.

[7] Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. Science, 313 (5786), 504–507.

[8] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. In Advances in Neural Information Processing Systems (pp. 2672–2680).

[9] Wang, X., Xie, L., Dong, C., & Shan, Y. (2021). Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 1905–1914).

[10] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention (MICCAI) (pp. 234–241). Springer, Cham.

[11] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., ... & Girshick, R. (2023). Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 4015–4026)

