Hierarchical Ship Classification Employing a Multi-Level Framework Based on YOLOv11: Enhancing Accuracy in Similar Class Differentiation

Eduardo H. Teixeira, Samuel B. Mafra, and Felipe A. P de Figueiredo

***Abstract*— This work proposes a hierarchical approach for ship classification in optical images, aiming to improve the separation between classes with high visual similarity. The strategy employs two stages based on the YOLOv11 architecture, with the first stage using a generalist detector that groups similar classes to reduce classification complexity and the second stage applying a specialized classifier to distinguish between subcategories. Experiments conducted with the InaTechShips dataset demonstrated that the hierarchical framework increased the mAP from 96.3% to 98.6% and improved classification accuracy for similar classes from 83.46% to 91.03%.**

***Keywords*— YOLOv11, Hierarchical Ship Classification, Detection.**

I. Introduction

The detection of ships in optical images plays a crucial role in applications such as maritime security, environmental monitoring, and port logistics [1]. In many of these scenarios, it is not enough to locate the presence of ships. It is also necessary to identify their class to enable specific monitoring, response, or control actions. For example, oil tankers are monitored more consistently than recreational yachts because they carry cargo that can cause significantly greater environmental impact. However, the presence of visually similar classes poses substantial challenges, even when using modern object detection architectures based on convolutional neural networks (CNNs) [2].

Recent studies have shown that state-of-the-art detection and classification models still struggle to adequately separate certain ship categories, highlighting the inherent limitations of approaches based on a single model [3]. These difficulties directly impact the reliability of automated surveillance and maritime traffic analysis systems, particularly in operational scenarios that demand high precision [4].

Convolutional neural networks have become the primary tool for object classification and detection tasks due to their ability to extract multiscale and adaptive representations of visual patterns [5]. However, even advanced architectures encounter difficulties when class separation relies on very subtle visual variations. In this context, hierarchical approaches emerge as an effective alternative by organizing the decision process into multiple levels of granularity [6].

By dividing the classification problem into sequential stages, each focused on subsets of classes or specific features, it is possible to reduce the complexity faced by each model. This strategy favors the identification of subtle patterns, mitigates class overlap, and increases system robustness when dealing with visual ambiguities. Additionally, the modularity provided by the hierarchical structure facilitates system adaptation to new scenarios, allowing the incorporation of additional specialists without requiring a complete redesign of the original architecture [7].

To address these limitations, this work proposes a hierarchical solution based on the YOLOv11 architecture, structured into two distinct stages. In the first stage, classes with higher visual similarity are grouped, such as TANKER and OIL PRODUCTS TANKER in Figure 1, adapted from [3]. In the second stage, a specialized classifier performs fine-grained differentiation between the corresponding subcategories. This modular framework aims to simplify decision-making during inference and enhance system robustness when handling visually ambiguous cases [8]. The choice of the TANKER classes was made due to the high confusion rate between them, as presented in [3], but this approach can be extended to any classes with high visual similarity.

This paper is organized as follows. Section II presents related work. Section III describes the proposed methodology. Section IV details the experiments and discusses the results. Finally, Section V concludes the paper and suggests directions for future research.

II. Related Works

Classifying visually similar objects remains one of the most persistent challenges in computer vision, especially in domains where small visual variations correspond to distinct semantic categories [9]. Even with modern convolutional neural network architectures, traditional models exhibit limitations when handling classes that share overlapping visual and contextual features [2]. For instance, in [3], 68 models are evaluated for ship classification, and all of them exhibit a similar error pattern, especially between 2 of the 10 evaluated classes. In this context, hierarchical structures have been adopted as a strategy to increase sensitivity and modularize the decision-making process, reducing the complexity that a single model would face in distinguishing all classes simultaneously.

This work was partially funded by CNPq (403612/2020-9, 311470/2021-1, and 306199/2025-4), by FAPEMIG (PPE-00124-23, APQ-04523-23, APQ-05305-23, and APQ-03162-24), by the Brasil 6G project (01245.020548/2021-07), supported by RNP and MCTI, and by the projects XGM-AFCCT-2024-2-5-1 and XGM-AFCCT-2024-9-1-1 supported by xGMobile - EMBRAPII-Inatel Competence Center on 5G and 6G Networks, with financial resources from the PPI IoT/Manufatura 4.0 from MCTI grant number 052/2023, signed with EMBRAPII.

Fig. 1: Examples of the 10 ship classes addressed in this study.

Several studies have explored the benefits of this structure. The HW-CNN [10] introduced hierarchical loss functions based on the Wasserstein distance, allowing the model to prioritize separation between semantically related class groups. The HD-CNN [11] proposed a two-level classifier architecture that combines an initial generalist stage with specialists focused on confounding classes, achieving substantial error reductions in benchmarks such as CIFAR-100 and ImageNet. Wang and Lu [12] further demonstrated that in tasks such as handwritten character recognition, hierarchical division into broad groups followed by specialized classifiers enables the extraction of structural nuances typically overlooked by flat models. Complementarily, Tazuddin et al. [13] developed an automated method for hierarchy construction using confusion matrices to group visually similar classes based on statistical error, a particularly effective approach in scenarios involving small or ambiguous objects.

Hierarchical structures have also shown success in semantic segmentation tasks. Chen et al. [14] proposed a hybrid model that combines convolutional networks with Markov Random Fields (MRFs) to capture spatial relationships across multiple semantic levels. The system refines predictions across hierarchical layers through an iterative self-learning module, increasing spatial consistency between segmented regions. This method proved robust in complex urban environments where multiple overlapping classes coexist.

In the maritime domain, adopting hierarchical models is a relatively recent but promising approach. Zhu et al. [15] proposed a hierarchical attention-based architecture for ship detection in SAR images, integrating global and local attention mechanisms to capture both scene-level context and detailed ship features. This strategy significantly reduced false positives in high-noise backgrounds. Similarly, Sun et al. [16] implemented a multiscale regional feature fusion method to enhance the distinction between civilian and military ships in high-resolution SAR data. Although based on different sensing modalities, both studies highlight the effectiveness of hierarchical modeling in resolving visual ambiguities in ship classification tasks.

More recently, the AMEFRN model [17] incorporated multiscale feature representations and auxiliary attributes into a hierarchical architecture tailored for fine-grained ship classification. This approach, evaluated on domain-specific benchmarks, achieved notable gains in both accuracy and modularity by enabling the specialization of subcomponents in critical class subsets.

Despite these contributions, a clear gap remains in applying such strategies to ship classification in high-resolution optical imagery, where fine details and visual similarity across classes pose additional challenges. This work addresses this gap by proposing a two-stage hierarchical framework based on the YOLOv11 architecture, which integrates a generalist detector and a specialized classifier to enhance classification performance in visually ambiguous scenarios while maintaining inference efficiency.

III. Methodology

The experiments conducted in this work were performed on a system equipped with an AMD Ryzen 9 5950X 16-core processor of 3.40 GHz, 128 GB of RAM, and an NVIDIA RTX 3090 GPU with 24 GB of memory. The PyTorch library was employed to train and run the models. The training and validation images were obtained from the InaTechShips dataset[1], which comprises 20,000 optical images of ships categorized into 10 classes, as shown in Figure 2. The dataset was split into 15,000 images for training and 5,000 for validation, following a 75–25% distribution. All images were resized to 640×640 pixels during inference to ensure consistency across comparisons. To ensure statistical robustness, the models were trained using 10 repetitions of stratified hold-out with random shuffling before each training iteration. The results were reported based on aggregated mean values.

Four different approaches based on the detector YOLOv11 architecture were evaluated. In the first approach, referred to as the YOLO Normal model, the network was trained using the 10 original classes with an image resolution of 640×640 pixels. The second approach, called YOLO with higher resolution, used 1280×1280 pixel images during training while maintaining inference at 640×640, to assess whether higher input resolution during learning would improve the distinction of visually similar classes [18]. The third approach, YOLO with more images, was validated using the same images as the previous models to ensure a fair comparison, but doubled the number of training samples for the two most frequently confused classes, TANKER and OIL PRODUCTS TANKER. The aim was to investigate whether the statistical

[1] https://www.github.com/EduardoHT/InaTechShips/

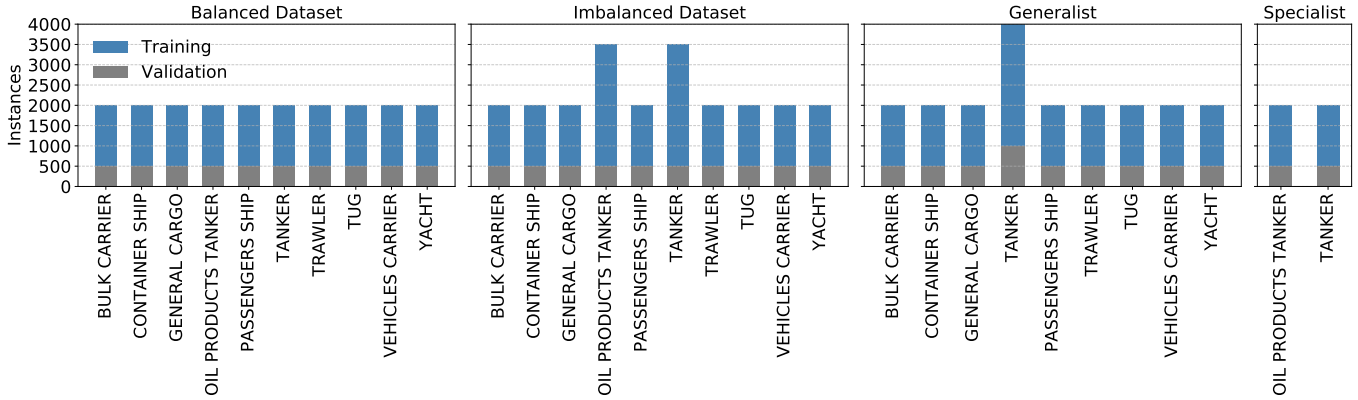

Fig. 2: Distribution of training and validation samples across the four experimental setups. From left to right: balanced dataset with uniform class distribution, imbalanced dataset with doubled instances for TANKER and OIL PRODUCTS TANKER, dataset used in the generalist detector with 9 classes, and dataset used in the specialist classifier with only the 2 classes.

reinforcement of these categories would reduce misclassification errors. Finally, the fourth approach proposed a hierarchical two-stage structure, reorganizing the decision process across complementary levels.

In the hierarchical approach, the system leverages a two-level structure, as illustrated in Figure 3. The first level consists of a YOLOv11 model trained to detect 9 classes, where the TANKER and OIL PRODUCTS TANKER categories are merged into a generic TANKER class. Detections corresponding to this generic class are cropped from the original images and resized to 640×640 pixels. These crops are then passed to a second model, which reuses the YOLOv11 architecture as a specialized classifier trained exclusively to distinguish between the two originally merged subcategories. This second model performs a binary classification task, enabling more focused and accurate representation learning for subtle visual differences.

The YOLOv11 architecture used in the proposed system comprises three main blocks: the backbone, responsible for visual feature extraction; the neck, which reorganizes these features across multiple scales; and the head, which generates final predictions based on the learned representations. For detection tasks, this comprehensive structure is essential for localizing ships of varying sizes and orientations. The neck facilitates multiscale fusion, enhancing the detection of small or partially occluded objects, while the head produces bounding boxes and corresponding class labels. For classification, YOLOv11 is used in a modified version, referred to as YOLOv11-cls. This variant retains the backbone for feature extraction but removes the neck, as multiscale preservation is unnecessary for categorical decisions on single instances. The head is simplified into a binary classification

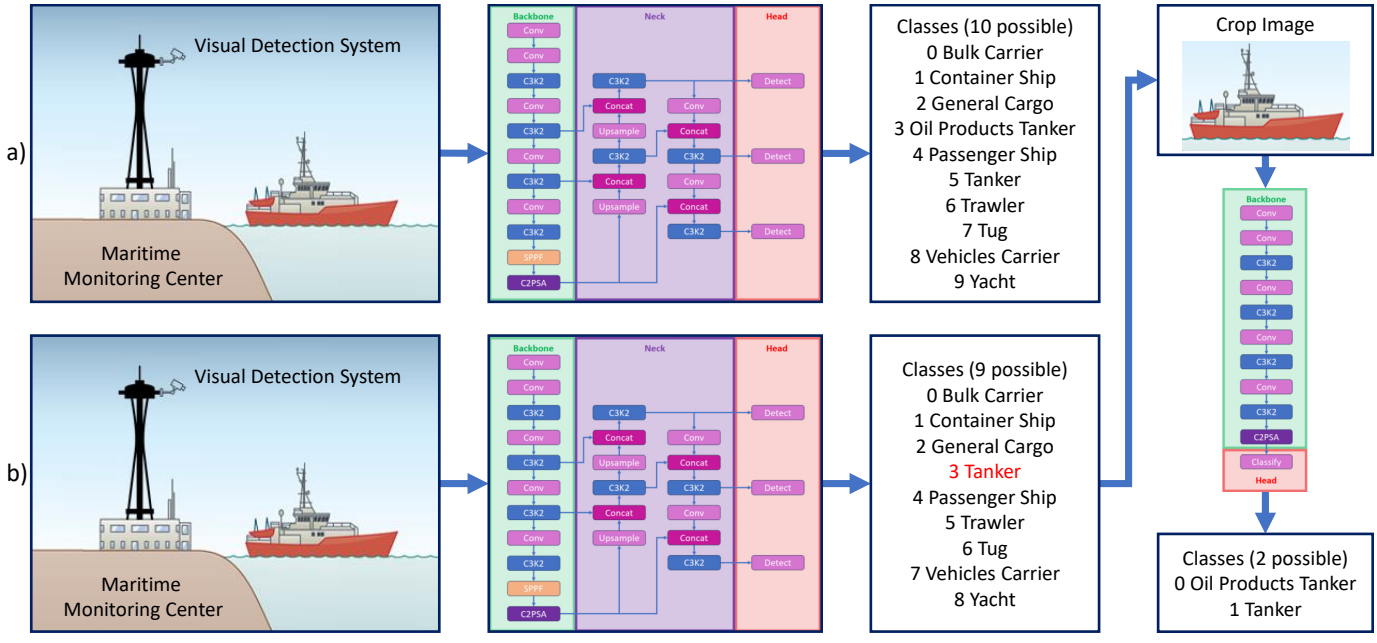

Fig. 3: Comparison between a) the original YOLOv11-based detection approach using only the detector, and b) the proposed hierarchical framework combining the YOLOv11 detector with an additional classification module.

block. Additionally, the Spatial Pyramid Pooling Fast (SPPF) module, originally positioned at the end of the backbone in the detection version to summarize multiscale information, is removed. The Cross-Stage Partial with Spatial Attention (C2PSA) spatial attention block, already part of the original architecture, is retained and plays a key role in highlighting discriminative regions of the image, optimizing fine-grained classification between visually overlapping categories [19].

This functional separation between generalist detection and specialized classification represents the core of the proposed hierarchical strategy. By delegating more complex decisions to a dedicated and focused module, the system can mitigate inter-class confusion, particularly between categories with highly similar visual features. The comparison among the four approaches enables an isolated analysis of the effects of resolution, sample reinforcement, and hierarchical structuring on overall system performance, providing a detailed assessment of the key factors that enhance accuracy in the classification of visually similar ships [20].

The classification accuracy between the TANKER and OIL PRODUCTS TANKER classes was computed as shown in Equation 1, which expresses the proportion of correct predictions over the total number of detections, both correct and incorrect, restricted to those two classes:

$$\text{Accuracy} = \frac{\sum_{i\in\{T\}} \text{correct}_i + \sum_{i\in\{O\}} \text{correct}_i}{\sum_{i\in\{T,O\}} (\text{correct}_i + \text{incorrect}_i)} \times 100 \quad (1)$$

where

- $i \in \{T\}$ and $i \in \{O\}$ correspond to the TANKER and OIL PRODUCTS TANKER classes, respectively
- correct_i denotes the number of detections correctly assigned to class i
- incorrect_i denotes the number of detections belonging to class i that were misclassified either as the other tanker class or as one of the remaining eight classes

IV. Experiments and Results

To evaluate the effectiveness of the proposed hierarchical strategy, four variations of the YOLOv11 architecture were compared using 10 repetitions of stratified hold-out with random shuffling before each training. The evaluated models were: YOLOv11s, YOLOv11s High Resolution, YOLOv11s Imbalanced (with the number of training samples duplicated for the TANKER classes), and YOLOv11s Hierarchical, which incorporates two levels of decision-making. Table I summarizes the architectural statistics of the evaluated models.

TABLE I: Architectural complexity of YOLOv11 models

Model	Layers	Parameters	GFLOPs	Classes
YOLOv11s	238	9,416,670	21.3	10
YOLOv11s High Resolution	238	9,416,670	21.3	10
YOLOv11s Imbalanced	238	9,416,670	21.3	10
YOLOv11s Hierarchical (det)	238	9,416,283	21.3	9
YOLOv11s Hierarchical (cls)	112	5,436,690	12.0	2

Table II presents the results for the first detection level of each model. The first three retain the original set of 10 classes, while the hierarchical approach groups TANKER and OIL PRODUCTS TANKER into a single generic class, reducing the number of categories to 9 at this stage.

TABLE II: Detection performance of YOLOv11 models (first level)

Model	Precision	Recall	mAP@50	Time (ms)
YOLOv11s	0.898	0.946	0.963 ± 0.0084	10.4
YOLOv11s High Resolution	0.894	0.915	0.933 ± 0.0093	10.5
YOLOv11s Imbalanced	0.922	0.920	0.954 ± 0.0113	10.4
YOLOv11s Hierarchical	0.971	0.973	0.986 ± 0.0062	10.3

The high-resolution training model did not significantly improve over the baseline, suggesting that training with increased input size alone does not benefit detection at 640×640 inference resolution. The model with the number of training samples duplicated in critical classes resulted in a slight precision gain, from 0.898 to 0.922, but the mAP remained nearly unchanged. In contrast, the hierarchical strategy yielded improvements across all metrics, with a mAP of 98.6%, representing a 2.3-point gain over YOLOv11s. The increase in average inference time by 2.2 ms, due to adding the classification model at a second level, raised the total inference time from 10.3 ms to 12.5 ms. This computational overhead introduced by using two models can be considered acceptable given the performance gain.

A dedicated evaluation was conducted using 1,000 balanced samples, 500 images from each class, to assess the distinction between TANKER and OIL PRODUCTS TANKER. All detection models were applied to this set, and the predictions corresponding to the two target classes were evaluated either directly by the single-stage models or at two levels for the hierarchical approach. The outcomes of this focused evaluation are presented in Table III.

TABLE III: Classification accuracy for TANKER classes

Model	Detections	Correct	Errors	Accuracy (%)
YOLOv11s	955	797	158	83.46
YOLOv11s High Resolution	955	794	161	83.14
YOLOv11s Imbalanced	980	825	155	84.18
YOLOv11s Hierarchical	981	893	25 + 63	91.03

The results indicate that oversampling the critical classes had only a marginal effect, increasing accuracy from 83.46% to 84.18%. The hierarchical approach, however, achieved a significant improvement, reaching 91.03% final accuracy. This figure accounts for 25 misclassifications at the first level (where predictions were incorrectly assigned to one of the remaining 8 classes) and 63 second-level classification errors between the two merged classes. Separating responsibilities between levels proved essential in reducing local ambiguities that single-model approaches could not resolve.

Notably, the hierarchical model already performed better at the detection stage, even before second-level refinement. This suggests that merging visually similar categories simplifies the initial decision space and improves the overall detection performance.

These findings confirm the benefits of using a modular decision strategy based on two levels, particularly in scenarios involving visually ambiguous categories. The proposed structure enhances class-level discrimination and overall system robustness, offering a viable solution for maritime monitoring

scenarios, where classification errors can lead to operational consequences.

V. Conclusion and Future Works

This work presented a hierarchical ship classification strategy for optical images, structured into two levels based on the YOLOv11 architecture. The system combines a generalist detector in the first level, where visually similar classes are merged, with a specialist classifier in the second level, responsible for refined differentiation between critical subcategories. This modular design reduces classification complexity and enhances precision when dealing with overlapping visual patterns.

Four different approaches were compared: the baseline YOLOv11s model, a version trained with higher-resolution images, an imbalanced configuration with increased samples in the most confused classes, and the proposed hierarchical model. While resolution and oversampling led to marginal improvements, the hierarchical approach consistently outperformed the others, achieving an mAP of 98.6% and 91.03% classification accuracy for the challenging class pair. Additionally, it was observed that detection performance also improved in the hierarchical version, even before second-level classification.

The proposed architecture proved effective in resolving visual ambiguities and maintaining a balanced trade-off between accuracy and computational cost. Notable gains in robustness and precision offset the slight increase in inference time.

Future work includes extending the hierarchical structure to additional low-separability class groups, enabling multi-level decision hierarchies. Furthermore, the use of model compression and quantization techniques will be explored to support deployment on embedded devices. The integration with complementary sources, such as AIS signals, also represents a promising direction for monitoring scenarios involving potential AIS spoofing or going dark events, thereby enhancing decision reliability. Finally, adopting active learning strategies could enhance the model's adaptability in real-world environments with unlabeled data.

References

[1] M. Cruz, E. H. Teixeira, S. B. Mafra, and F. A. P. de Figueiredo, "A multi-faceted approach to maritime security: Federated learning, computer vision, and iot in edge computing," in *XLI Brazilian Symposium on Telecommunications and Signal Processing (SBrT2023)*, 2023.

[2] E. Teixeira, B. Araujo, V. Costa, S. Mafra, and F. Figueiredo, "Literature review on ship localization, classification, and detection methods based on optical sensors and neural networks," *Sensors*, vol. 22, no. 18, 2022.

[3] E. H. Teixeira, S. B. Mafra, and F. A. De Figueiredo, "Inatechships: A validation study of a novel ship dataset through deep learning-based classification and detection models for maritime applications," *Ocean Engineering*, vol. 326, p. 120823, 2025.

[4] V. Costa, E. Teixeira, S. Mafra, and F. Figueiredo, "Pré-processamento de imagens de baixa resolução utilizando deep learning baseado em um autoencoder," in *XL Simpósio Brasileiro de Telecomunicações e Processamento de Sinais (SBrT2022)*, 2022.

[5] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks: Analysis, applications, and prospects," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 12, pp. 6999–7019, 2022.

[6] C. Zhu, H. Zhou, R. Wang, and J. Guo, "A novel hierarchical method of ship detection from spaceborne optical image based on shape and texture features," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 9, pp. 3446–3456, 2010.

[7] H. Zheng, Z. Hu, J. Liu, Y. Huang, and M. Zheng, "Metaboost: A novel heterogeneous dcnns ensemble network with two-stage filtration for sar ship classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[8] X. Zhang, Y. Lv, L. Yao, W. Xiong, and C. Fu, "A new benchmark and an attribute-guided multilevel feature representation network for fine-grained ship classification in optical remote sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 1271–1285, 2020.

[9] M. R. Cruz, F. A. P. de Figueiredo, S. Mafra, D. M. de Oliveira, and E. H. Teixeira, "Evaluating computer vision architectures for ship classification: A comparative study," in *XLI Brazilian Symposium on Telecommunications and Signal Processing (SBrT2023)*, 2023.

[10] Y. Liu, C. Y. Suen, Y. Liu, and L. Ding, "Scene classification using hierarchical wasserstein cnn," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 5, pp. 2494–2509, 2019.

[11] Z. Yan, H. Zhang, R. Piramuthu, V. Jagadeesh, D. DeCoste, W. Di, and Y. Yu, "Hd-cnn: Hierarchical deep convolutional neural networks for large scale visual recognition," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 2740–2748.

[12] Q. Wang and Y. Lu, "Similar handwritten chinese character recognition using hierarchical cnn model," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 01, 2017, pp. 603–608.

[13] A. M. Tazuddin, A. Abdullah, and Z. R. Mahayuddin, "Hierarchical cnn automated hierarchy creation using output prediction analysis with misprediction matrix," in *2023 International Conference on Electrical Engineering and Informatics (ICEEI)*, 2023, pp. 1–6.

[14] Y. Chen, L. Wang, J. Li, and C. Zheng, "Hierarchical self-learning knowledge inference based on markov random field for semantic segmentation of remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–18, 2024.

[15] C. Zhu, D. Zhao, Z. Liu, and Y. Mao, "Hierarchical attention for ship detection in sar images," in *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, 2020, pp. 2145–2148.

[16] Z. Sun, B. Xiong, Y. Lei, X. Leng, and K. Ji, "Ship classification in high-resolution sar images based on cnn regional feature fusion," in *2021 CIE International Conference on Radar (Radar)*, 2021, pp. 1445–1449.

[17] X. Zhang, Y. Lv, L. Yao, W. Xiong, and C. Fu, "A new benchmark and an attribute-guided multilevel feature representation network for fine-grained ship classification in optical remote sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 1271–1285, 2020.

[18] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, "Deep high-resolution representation learning for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3349–3364, 2021.

[19] Z. Zhou, "Traffic accident detection based on yolov11," in *2024 IEEE 2nd International Conference on Electrical, Automation and Computer Engineering (ICEACE)*, 2024, pp. 363–369.

[20] A. M. Carrington, D. G. Manuel, P. W. Fieguth, T. Ramsay, V. Osmani, B. Wernly, C. Bennett, S. Hawken, O. Magwood, Y. Sheikh, M. McInnes, and A. Holzinger, "Deep roc analysis and auc as balanced average accuracy, for improved classifier selection, audit and explanation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 329–341, 2023.