Journal of Marine Science and Application (2026) 25: 600-616
https://doi.org/10.1007/s11804-025-00703-7

RESEARCH ARTICLE

A Comparative Analysis of Deep Learning Approaches for Visual Perception Benchmarks in Ship Navigation

Ruolan Zhang[1], Xingchen Ji[1,2], Jinichi Koue[2] and Katsutoshi Hirayama[2]

Received: 08 January 2025 / Accepted: 07 March 2025

Abstract
The establishment of a reliable benchmark for evaluating model performance is critical for advancing deep learning (DL), including its application in the recognition of the ship navigation environment. Despite the steady progress being made in object detection models across various tasks, maritime navigation presents unique challenges, such as long distances, miscellaneous objects, wide perception scales, and local conditions and features of water areas. Therefore, the improvement of DL approaches for this domain remains a significant challenge. Using a widely applicable offshore image dataset from the ship bridge, we evaluated the performance of the state-of-the-art object detection model from three perspectives: average precision, multiscale feature calculation, and intersection-over-union design, and explored the factors that may affect the model performance evaluation benchmark from the perspective of data quality, scale calculation, feature quantification, and object association. Our experiments have demonstrated that, in the context of object detection tasks within complex water surface traffic scenes, comprehensive model performance evaluation benchmarks are essential. Such benchmarks must incorporate multiple dimensions of the model.

Keywords Long-range perception; Visual navigation; Dataset; Multiscale detection; Vision benchmark

1 Introduction

The navigation environment of a ship is complex and dynamic, particularly regarding visual perception tasks, such as arrival and departure from ports and busy canal areas, which pose significant challenges. Autonomous navigation relies on vision as a crucial and valuable perception method. Artificial intelligence (AI) has garnered significant attention within the autonomous shipping industry, where intelligent navigation equipped with advanced perception is a critical scientific topic. Specifically, computer vision is the most promising direction for autonomous ship navigation. However, the implementation of visual perception and the utilization of disparate maritime datasets present numerous obstacles that impede this objective (Er et al., 2023).

Over the last few decades, computer vision has undergone tremendous growth owing to the widespread adoption of deep learning (DL) techniques. Among the fundamental tasks in computer vision, object detection has seen significant applications (Kaur and Singh, 2023; Manakitsa et al., 2024). In the autonomous vehicle industry, visual perception is of paramount importance (Li et al., 2023; Karas et al., 2023; Liu et al., 2023), where it economically resolves the issue of surrounding environment perception. Moreover, computer vision has transformed medical image detection, considerably enhancing the efficiency of computed tomography scans and X-rays (Lenka and Tripathy, 2024; Hussain et al., 2022; Khan et al., 2021). Computer vision is also increasingly utilized in precision agriculture, toy manufacturing, and textile quality inspection (Borkar et al., 2023; Islam et al., 2024; Ismail and Malik, 2022).

Computer vision approaches have been instinctively employed in various directions in the maritime industry, such as obstacle classification, navigational aid object detection, and object tracking. However, the utilization of computer vision in actual nautical practice remains limited. The applications of computer vision in the maritime industry can be categorized into surveillance for maritime managers and navigation for seafarers (Durlik et al., 2023). From the per-

Article Highlights

- This article emphasizes the necessity of precise evaluation benchmarks for visual model performance in complex maritime scenarios.
- Improved annotation strategies significantly enhance object detection accuracy in diverse and dynamic maritime environments.
- Advanced IoU loss functions are proposed to improve the detection performance of small and distant objects under challenging maritime conditions.

✉ Katsutoshi Hirayama
hirayama@maritime.kobe-u.ac.jp

[1] Navigation College, Dalian Maritime University, Dalian 116000, China

[2] Graduate School of Maritime Sciences, Kobe University, Kobe 6580022, Japan

spective of maritime surveillance, computer vision can significantly improve the efficiency of monitoring marine traffic and detecting transportation anomalies, illegal fishing, pollutants, and smuggling. From the perspective of autonomous ship navigation, computer vision is a promising solution for identifying abnormal behavior, avoiding collisions, and recognizing cross-encounter situations, despite the difficulties involved in obtaining information regarding the location, size, course, speed, and other relevant parameters of a ship. Hence, we aim to identify the limiting factors for performance benchmarks.

Computer vision models are designed to address different problems, and state-of-the-art (SOTA) models typically focus on achieving high performance in mean average precision (mAP) and frames per second (FPS). In the maritime community, numerous researchers have proposed different algorithms and assumptions for deployment in the maritime industry. For instance, Yan et al. (2021) presented a new tracking architecture that uses an encoder−decoder transformer as the key component, casting object tracking as a direct bounding box prediction problem without using any proposals or predefined anchors. Bochkovskiy et al. (2020) and Han et al. (2021) proposed an improved YOLOv4 detection model for ship detection, which introduces a new structure that can effectively improve the feature extraction effect of the model for ships of different scales and reduce the number of model parameters, thereby improving the inference speed and detection accuracy of the model. In addition, other algorithms, such as the Andrew Howard NAS+NetAdapt algorithm and latency-aware architecture search, have been proposed (Howard et al., 2019; Chen et al., 2020). However, no consolidated evaluation standard or dataset benchmark to confirm the efficacy of these distinct algorithms and assumptions has been established.

The deployment of improved models in real navigation environments is challenging. Several researchers focus solely on improving object detection accuracy, validating their enhanced models using public datasets. However, evidence reveals that detection methods can only identify ship classes. For instance, Lin et al. (2014) proposed a new dataset aimed at advancing the SOTA in object recognition by placing the question of object recognition in the broader context of scene understanding. Similarly, Idrees et al. (2018) proposed the University of Central Florida-Quality Normalized and Region Fusion dataset (UCF-QNRF), a large-scale crowd-counting dataset, which is currently the largest dataset available for training and evaluating large-scale dense crowd-counting models. Although these datasets are well-known public datasets, they are not specific to ship navigation. As such, detecting only ship classes falls far short of satisfying actual navigation applications.

In addition, different scenarios require different perspectives, such as on-shore and on-board, and each perspective has its specific applications. For instance, Shao et al. (2018) developed a new large-scale dataset of ships called SeaShips, which is designed for training and evaluating ship object detection algorithms. However, the dataset only labels six major types of ships limited to the Yangtze River Basin. Alternatively, Zhou et al. (2021) introduced the Water Surface Object Detection Dataset (WSODD), a high-quality annotated benchmark dataset for detecting different water surface objects. The dataset serves as a benchmark for various water surface object detection algorithms. Nonetheless, the complexity of the marine environment poses a challenge to the real-time monitoring of marine navigation.

To summarize the research, we present an image dataset from the bridge perspective that covers various ship navigation scenarios, along with the corresponding object detection benchmarks. We optimized the object classification and detection strategy for ship navigation environment recognition and adapted a real navigation scene dataset to train and evaluate the performance benchmark of milestone detection models. This paper is organized as follows: Section 2 compares different benchmarks and describes the dataset production process. Section 3 outlines different evaluation indicators and presents our standard. Section 4 discusses the training results of each object detection model and analyzes the reasons for these results. Finally, Section 5 provides the conclusion and future perspectives.

2 Related work

The benchmark is a crucial aspect of evaluating the performance of a given method in various disciplines. Cavegn et al. (2014) proposed the International Society for Photogrammetry and Remote Sensing and European Spatial Data Research (ISPRS-EuroSDR) benchmark for high-density aerial image matching to evaluate dense matching methods for oblique aerial images. Hackel et al. (2017) introduced a new 3D point cloud classification benchmark dataset with over four billion manually labeled points to be used as input for data-hungry (deep) learning algorithms. These benchmarks are essential, with applications in robotics, augmented reality, and urban planning. 3D point cloud classification is a critical task that finds applications in these fields. Recent advancements in machine learning and computer vision demonstrate that large training datasets are necessary for training classifiers to tackle complex real-world tasks.

2.1 Dataset design

The data quality sets the upper limit of the objective. For DL-based image-understanding tasks, standard configurations require large-scale benchmark datasets with millions of images. Deng et al. (2009) proposed the ImageNet

dataset, which has 12 subtrees with 5 247 synsets and a total of 3.2 million images. The ImageNet dataset has been a benchmark for evaluating the performance of image classification algorithms. Everingham et al. (2010) proposed the PASCAL Visual Object Classes (VOC) challenge as a benchmark in visual object category recognition and detection, providing standard datasets of images and annotations and standard evaluation procedures for the computer vision and machine learning communities. The PASCAL VOC challenge and its datasets have been the foundation for many excellent computer vision tasks, such as classification, localization, detection, segmentation, and action recognition.

Table 1 shows the commonly used open-source datasets and their evaluation metrics. The general norm is the basic evaluation index of computer vision, whereas the special norm gives special status to different datasets. Cityscapes is a benchmark suite that includes an evaluation server for developers to upload their research results and obtain rankings for different tasks, including pixel-level, instance-level, and panoptic semantic labeling, as well as 3D vehicle detection (Cordts et al., 2016). Cityscapes has a special benchmark, i.e., instance-level intersection-over-union (iIoU), and evaluates semantic labeling using an iIoU metric, with a detection score calculated for each label class. KITTI is an evaluation dataset for computer vision algorithms in autonomous driving scenarios, focusing on the highly ill-posed problem of scene flow estimation from two temporally consecutive images to obtain 3D structure and 3D motion (Menze and Geiger, 2015). Places2 is a benchmark used for scene recognition, scene fix, and super-resolution reconstruction (Zhou et al., 2017). Finally, CelebA is used for face-related training, with special evaluation benchmarks for pose variations (Liu et al., 2018).

Table 1 Specific evaluation benchmarks required for measuring model performance in different scenarios and tasks

Name	Field	Type	General norm	Special norm
Cityscapes	Scene understanding	Video	IoU/FPS/mAP	iIoU/DS
KITTI	Autonomous driving	Video	IoU/FPS/mAP	Scene flow
Places2	Scene recognition	Image	IoU/FPS/mAP	Scene fix
CelebA	Face attributes	Image	IoU/FPS/mAP	Pose variations

In the maritime domain, the limited availability of evaluation benchmarks and datasets has hindered research on ship navigation perception tasks. To address this issue, Iancu et al. (2021) compiled a dataset consisting of images of maritime vessels called ABOships, which provides a solution to the inconsistencies in image annotation that arise from manual labeling. However, given the complex maritime environment of ship navigation and the diversity of obstacles and aids to navigation, the existing open-source dataset benchmarks are inadequate. With the increase in the application and development of intelligent transportation systems, such as autonomous ship navigation, maritime surveillance, and navigation facility deployment, a widely accepted and standardized large-scale marine object verification dataset is needed. This work presents a dataset that facilitates the deployment of vision models for ship navigation and monitoring.

Ship navigation perception is a complex issue influenced by various factors, such as equipment vibration, model dependency, and shore light interference (Cai et al., 2024). Although several DL-based marine object detection models have been proposed, the absence of common evaluation criteria makes it challenging to compare different improved models (Zhang et al., 2022). Navigation safety is impacted by multiple factors during a voyage, such as ship shaking, extreme weather conditions, different waterways, visibility, fog, and nighttime environment. Consequently, the dataset must include diverse navigation scenarios.

To address this issue, we propose a ship navigation benchmark and dataset called "ShipNav" to assess the performance of vision scenarios and tasks. The ShipNav dataset is based on 12 ship bridge acquisition classes. This study classifies the ShipNav dataset into two types of objects that arise during practical ship navigation at sea: navigational aid and obstacle objects. Navigational aid objects comprise "work boat", "cargo ship", "own body", "unidentified ship", "unidentified ship-N", "packet", "ship bow", and "island". Obstacle objects include "gantry crane", "lighthouse", "bridge", and "ship lock".

2.2 DL-based object detection model

Over the past decade, the progress of computer vision tasks has been closely linked to the growth of data and the application of convolutional neural networks (CNNs). CNNs have achieved remarkable success in computer vision applications, including face recognition, object detection, vision in robotics, and self-driving cars (Voulodimos et al., 2018). DL, a dominant branch of AI, has extended with diversified network structures, which enables it to capture the features of big data automatically and efficiently (Chai et al., 2021).

Several classical networks will be introduced in this subsection. For instance, Simonyan and Zisserman proposed the VGGNet that thoroughly evaluates networks with increasing depth using an architecture with small (3 × 3) convolution filters. The use of a smaller 3 × 3 convolution kernel and a deeper network can reduce the number of parameters while increasing the learning capability of CNN for features through more nonlinear transformations (Simonyan and Zisserman, 2014). Then, He et al. (2016) introduced ResNet, which contains a residual network with 152 layers, 8 times deeper than VGGNet, ResNet50, and ResNet101 with dif-

ferent depths. EfficientNet was introduced by Tan and Le (2019) to balance network depth, width, and resolution to achieve better performance.

Specifically, we verify that the use of a deeper architecture often involves employing smaller convolution filters (e.g., 3 × 3) in succession. Two consecutive 3 × 3 convolutional layers can achieve an effective receptive field equivalent to a single larger filter (e.g., 5 × 5 or 7 × 7) while requiring significantly fewer parameters (because 3 × 3 convolutions have only 9 parameters per filter versus 25 or 49 parameters for larger kernels). In addition, deeper networks frequently utilize architectural techniques, such as bottleneck layers and residual connections, which help reduce the overall parameter count by compressing the feature space and enabling efficient feature reuse. These strategies contribute to building deeper networks that are both computationally efficient and capable of learning complex representations.

Different object detection methods have been developed based on various backbone networks. The current mainstream approaches are divided into single-stage methods, such as SSD, RetinaNet, and YOLO (Liu et al., 2016; Lin et al., 2017; Redmon et al., 2016; Long et al., 2020), and two-stage methods, such as the R-CNN series (Girshick et al., 2014; He et al., 2015; Girshick, 2015; Ren et al., 2015). YOLOv5 is a typical network architecture in the YOLO series and has achieved high detection accuracy and fast inference speed (Liu et al., 2020; Jocher, 2020). In addition, EfficientDet and CenterNet have exhibited good performance in object detection (Tan et al., 2020; Duan et al., 2019). We have opted to utilize YOLOv5 in our research paper despite the availability of YOLOv11 for several reasons. YOLOv5 is a well-recognized and commonly employed object detection model in the research community. YOLOv5 has undergone rigorous testing and benchmarking on a diverse array of datasets, and its effectiveness is well-documented in the literature.

To achieve outstanding performance in evaluating DL-based object detection models, we provide an image dataset containing various ship navigation scenarios, along with the corresponding object detection benchmarks and optimized object classification strategies. This dataset effectively demonstrates the superior performance of SOTA object detection models in real-world scenarios and promotes the industrial deployment of visual perception for autonomous ships.

3 Methodology and experiment

3.1 Conventional evaluation benchmarks

The conventional and most important evaluation metric for computer vision tasks is mAP (Henderson and Ferrari, 2016). mAP is widely used to analyze the performance of classification, object detection, and segmentation. Average precision (AP) measures how well the learned model performs in each category, whereas mAP measures the overall performance across all categories. After obtaining the AP for each category, the mAP can be calculated by taking the average value across all categories. Many object detection models, including Faster R-CNN, MobileNet, SSD, and YOLO, use mAP to evaluate model performance. Given that one image may have multiple labels, single-label classification criteria cannot be used as the evaluation standard. mAP is also used in several benchmarks, such as PASCAL VOC, COCO, and other open-source datasets.

We employ the standard performance evaluation metrics, namely, precision, recall, accuracy, and mAP, to assess the performance of our detection model. In this context, TP denotes true positive, FN denotes false negative, FP denotes false positive, and TN denotes true negative. The formulas for these metrics are expressed in Eqs. (1), (2), (3), and (4):

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}} \tag{3}$$

$$\text{mAP} = \frac{\sum_{i=1}^{K} \text{AP}_i}{K} \tag{4}$$

3.2 Multiscale feature learning

The process of multiscale feature learning involves transforming raw data into features that better express the essence of the problem. By applying these features to prediction models, the accuracy of the predictions of the model for unseen data can be improved (Yu et al., 2021). Alternatively, feature engineering involves identifying features that have a significant impact on the dependent variable *Y*, which is usually referred to as the independent variable *X*. The goal of feature engineering is to identify the most important features in the data that directly impact the predicted model and the results obtained. The predicted outcome depends on the available data, the prepared features, and the choice of the model.

Figure 1 illustrates a typical multiscale fusion network structure, which has a small feature receptive field suitable for processing small objects. In this network, the features of the last residual block layer of conv2, conv3, conv4, and conv5 are selected as the features of FPN. The C5 layer first undergoes a 1 × 1 convolution to obtain the M5 features, which are then upsampled. After the 1 × 1 convolution, the features of the C4 layer are added to obtain M4. This process is repeated two more times to obtain M3 and M2. Then, the features of the M layers are subjected to a 3 × 3

convolution to obtain the final P2, P3, P4, and P5 layer features. Multiscale feature learning is a promising approach for accurately extracting useful objects in complex backgrounds. In this study, we propose a powerful multiscale feature learning module, i.e., a hierarchical visual transformer via a shifted window, as one of the backbones in the maritime object extraction network. This module helps address the problem of object omission.

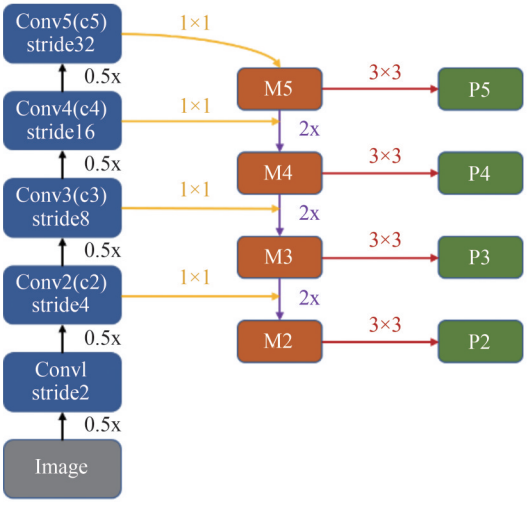

Figure 1 An example of a multiscale feature engineering structure

The quality of detection in computer vision heavily relies on the labeling level of the dataset. Different dataset labels are classified in varying ways and may have inconsistent definitions because of varying interpretations by different people, which is particularly true in the maritime industry, where labeling inconsistencies arise because of several factors, such as distance. For example, a boat may appear as a bright spot from far away but can be identified as a fishing boat when it comes closer. Furthermore, different application scenarios and classes have diverse classification standards, which currently lack a universal benchmark in the maritime research community. To address this issue, we propose a common evaluation benchmark for the maritime domain.

3.3 E-IoU with YOLOv5

Regression of bounding boxes is a crucial step in one-stage object detection. In object detection, the IoU expressed in Eq. (5) is the most commonly used indicator that possesses scale invariance, nonnegativity, identity, symmetry, and triangle inequality properties. IoU optimization does not exponentially increase the model complexity and is more conducive to training small models than adding convolutional layers. However, when the two boxes do not intersect, the distance between them cannot be determined. To address this issue, He et al. (2021) proposed a new power IoU loss function, called $\alpha-$IoU, which unifies the exponentiation of existing losses based on IoU for accurate box regression and object detection. The choice of α ($\alpha > 1$) can improve the loss and gradient adaptive weighted box regression accuracy for high IoU targets. This method is proven to improve the detection accuracy of small objects at long distances.

$$\mathcal{L}_{\mathrm{IoU}} = 1 - \mathrm{IoU} \Rightarrow \mathcal{L}_{\alpha-\mathrm{IoU}} = 1 - \mathrm{IoU}^{\alpha} \tag{5}$$

Reducing this distance helps the predicted bounding box more rapidly converge to the ground truth box in terms of position. The GIoU (Eq. (6)) approach introduces a minimum bounding box to address the issue of zero loss when there is no overlap between the detection and actual boxes, leveraging the characteristics of IoU, where the squared Euclidean distance between the center points of the two boxes B^{gt}. However, GIoU reduces to IoU in cases where the ground truth and detection frames are inclusive, and convergence in the horizontal and vertical directions is slow when the two frames intersect, resulting in decreased accuracy in detection.

$$\begin{aligned}\mathcal{L}_{\mathrm{GIoU}} &= 1 - \mathrm{IoU} + \frac{\left|C \setminus \left(B \cup B^{\mathrm{gt}}\right)\right|}{|C|} \Rightarrow \mathcal{L}_{\alpha-\mathrm{GIoU}} \\ &= 1 - \mathrm{IoU}^{\alpha} + \left(\frac{\left|C \setminus \left(B \cup B^{\mathrm{gt}}\right)\right|}{|C|}\right)^{\alpha}\end{aligned} \tag{6}$$

Based on the properties of IoU and given the limitations of GIoU, DIoU (Eq. (7)) calculates the Euclidean distance between the center points of two boxes directly to speed up convergence. Where C in $\mathcal{L}_{\mathrm{GIoU}}$ denotes the smallest convex shape enclosing B and B^{gt}; b and b^{gt} in $\mathcal{L}_{\mathrm{DIoU}}$ denote central points of B and B^{gt} with $\rho(\cdot)$ being the Euclidean distance and c being the diagonal length of the smallest enclosing box. However, the aspect ratio of the bounding box is not considered during the regression process, and there is still room for improving accuracy.

$$\begin{aligned}\mathcal{L}_{\mathrm{DIoU}} &= 1 - \mathrm{IoU} + \frac{\rho^2\left(b, b^{\mathrm{gt}}\right)}{c^2} \Rightarrow \mathcal{L}_{\alpha-\mathrm{DIoU}} \\ &= 1 - \mathrm{IoU}^{\alpha} + \frac{\rho^{2\alpha}\left(b, b^{\mathrm{gt}}\right)}{c^{2\alpha}}\end{aligned} \tag{7}$$

CIoU (Eq. (8)) is designed to improve the loss function of the detection frame scale by augmenting DIoU, which increases the loss of both length and width to align the predicted frame with the ground truth frame. However, the aspect ratio of CIoU is described in relative terms, making it challenging to converge during the training process and to balance the difficulty of sample detection.

$$\begin{aligned}\mathcal{L}_{\mathrm{CIoU}} &= 1 - \mathrm{IoU} + \frac{\rho^2\left(b, b^{\mathrm{gt}}\right)}{c^2} + \beta v \Rightarrow \mathcal{L}_{\alpha-\mathrm{CIoU}} \\ &= 1 - \mathrm{IoU}^{\alpha} + \frac{\rho^{2\alpha}\left(b, b^{\mathrm{gt}}\right)}{c^{2\alpha}} + (\beta v)^{\alpha}\end{aligned} \tag{8}$$

The IoU approach has continued to evolve, and a recent advancement is EIoU (Zhang et al., 2022). This metric calculates the difference between the width and height of the predicted bounding box based on CIoU instead of using the aspect ratio. In addition, EIoU introduces focal loss to address the issue of unbalanced difficult and easy samples. In our work, we integrate the EIoU module into the YOLOv5 algorithm to improve performance. EIoU (Eq. (9)) is defined as follows:

$$\begin{aligned}\mathcal{L}_{\mathrm{EIoU}} &= \mathcal{L}_{\mathrm{IoU}} + \mathcal{L}_{\mathrm{dis}} + \mathcal{L}_{\mathrm{asp}} \\ &= 1 - \mathrm{IoU} + \frac{\rho^2(b, b^{\mathrm{gt}})}{(w^c)^2 + (h^c)^2} + \frac{\rho^2(w, w^{\mathrm{gt}})}{(w^c)^2} + \frac{\rho^2(h, h^{\mathrm{gt}})}{(h^c)^2}\end{aligned} \tag{9}$$

where h^w and h^c are the width and height of the smallest enclosing box covering the two boxes. Based on the aforementioned IoU design ideas, the approach that we present can considerably improve the detection effect of the night object.

We assessed the annotation quality of the ShipNav dataset using Tesla V100 × 4 graphics cards and trained it on the SOTA object detection model. We partitioned the dataset in a ratio of 8 : 2 to create separate training and validation sets. To obtain the output results, we experimented with different image input sizes and backbone networks of various models.

4 Results and discussion

The results of our experiment are presented in Figure 2, which illustrates the relationship between the number of labeled classes and the detection accuracy in different ship navigation scenarios. Overall, we observe a positive correlation between the number of labels and the accuracy of the model. In cases where the labeled classes exhibit clear features and scales, the accuracy of the model can achieve the expected results even with a small number of labels. However, because ship navigation relies heavily on precise visual perception, the training of computer vision models requires a significant amount of data. As illustrated in Figure 2, the accuracy of the model is influenced by the design and distribution of the data labels.

As shown in Figure 2, the left vertical axis indicates the number of labels, the right vertical axis indicates the training results of the model, and the horizontal axis indicates the label category. The histogram presents the number of labels, whereas the line graph presents the mAP values of different SOTA models. Despite having only a few labels, the "ship lock", "bridge", and "island" classes exhibit obvious external features, enabling the detection accuracy to exceed 70%. Moreover, "ship lock" and "bridge" are multiscale features in our dataset. Because of the difficulty in capturing recognizable features of objects detected at night,

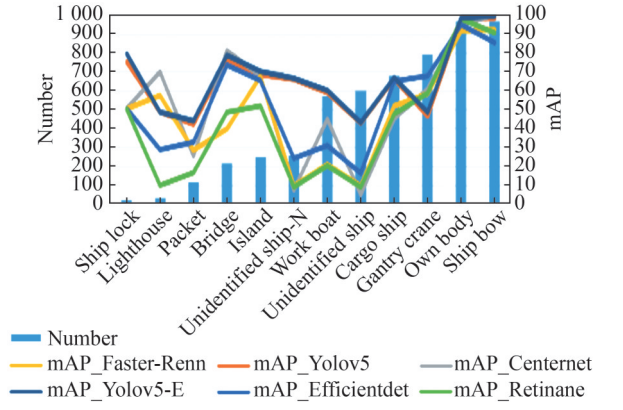

Figure 2 Relationship between the number of different labeling classes and the detection accuracy in different ship navigation scenarios

the accuracy of other SOTA models, except for "YOLOv5" and "CenterNet", is significantly lower than the average, despite the "unidentified ship-N" class being tagged with 280 labels. Based solely on mAP, the "YOLOv5" model has better detection performance than other SOTA models.

Table 2 shows the comparison between the original YOLOv5 and the improved version. YOLOv5-E is an enhanced method that incorporates EIoU and multiscale features, resulting in a 3.4% increase in mAP and a 1.5 FPS improvement. YOLOv5-E assigns a larger loss to the superior regression object, which leads to an improvement in the regression accuracy. The subsequent comparative analysis of our YOLO model is based on YOLOv5-E.

Table 2 Comparison of the improved YOLOv5 and the original version

Model	Input size	Precision	Recall	mAP50 (%)	FPS
YOLOv5-E	554 × 554	69.2	64.8	66.8	53.6
YOLOv5	554 × 554	64.9	61.3	63.4	52.1

4.1 Object detection speed performance criteria

The object detection speed is a critical aspect of real-time navigation applications. Hence, FPS serves as a key evaluation metric in environment perception models. Furthermore, the performance of the hardware device and the model itself, as well as the FPS, is influenced by the input size of the training samples, as demonstrated in this study.

4.1.1 Comparison of different SOTA models and input size training results

Table 3 presents the detection accuracy and speed of different models under various input sizes. We compared the results of training five detection models using different pretrained weights, input sizes, and backbone networks (i. e., VGG16, ResNet50, and ResNet101). The YOLOv5m model with an input size of 544 × 544 achieved the best detection results, with a mAP of 66.8%. The detection accuracy of

the RetinaNet series models was not positively correlated with the input size; however, RetinaNet(net50) achieved relatively good results. The EfficientDet, Faster R-CNN, and CenterNet models all exhibited improved detection accuracy with the increase in input size. Regarding FPS, the YOLOv5s model with an input size of 448 × 448 had the best performance, with a score of 62.8, whereas the EfficientDet model with an input size of 640 × 640 had the worst performance, with a score of 7.4. In selecting a ship navigation perception model, multiple dimensions need to be considered, and large-scale SOTA model tests can be conducted to screen appropriate models for further optimization.

Table 3 Detection results on the ShipNav dataset

Model	Input size	mAP50 (%)	FPS
YOLOv5-Es	448 × 448	60.8	62.5
YOLOv5-Es	544 × 544	63.1	58.9
YOLOv5-Es	640 × 640	65.7	55.5
YOLOv5-Em	448 × 448	64.7	57.4
YOLOv5-Em	544 × 544	66.8	53.6
YOLOv5-Em	640 × 640	65.1	50.1
RetinaNet(net101)	544 × 544	37.7	11.5
RetinaNet(net50)	640 × 640	43.2	15.7
RetinaNet(net101)	640 × 640	35.8	12.3
RetinaNet(net50)	544 × 544	34.5	18.0
EfficientDet	512 × 512	43.2	9.0
EfficientDet	640 × 640	54.2	7.4
Faster R(vgg)	544 × 544	50.0	34.7
Faster R(net50)	640 × 640	51.1	8.2
Faster R(vgg)	640 × 640	44.2	29.6
Faster R(net50)	544 × 544	57.7	13.2
CenterNet(net50)	544 × 544	57.0	23.2
CenterNet(net50)	640 × 640	64.5	21.2

Our training results demonstrate the effectiveness of our improved model in detecting long-distance and small objects. As presented in Table 3, the YOLOv5 model with our proposed modifications outperforms the other models across different input sizes, including 448 × 448, 544 × 544, and 640 × 640 pixels. When using the EfficientDet and Faster R-CNN detectors with the VGG16 backbone network, increasing the input image size to 544 × 544 and 640 × 640 pixels significantly improves the mAP performance. Thus, in addition to using an optimized IoU, increasing the input image size is also a viable strategy to enhance the detection accuracy of long-distance and small objects.

Additionally, as shown in Table 3, the mAP of RetinaNet101 decreases as the input scale increases, and the mAP of YOLOv5-Em performs suboptimally at the largest input size. This is because when the input resolution increases significantly, the training complexity and GPU memory usage also increase, often leading to a smaller batch size that can make the training process less stable if hyperparameters (e.g., learning rate and batch size) are not tuned accordingly. In addition, the default anchor configurations and feature pyramid scales of RetinaNet may become suboptimal at higher resolutions, limiting the capability of the network to accurately regress bounding boxes and ultimately leading to reduced mAP. Moreover, beyond a certain resolution, the incremental gains in detection performance may plateau or even decline if the dataset does not provide sufficient high-frequency details to justify such large images. At the same time, the use of larger input sizes can lead to a smaller batch size, negatively affecting gradient estimation. Coupled with the fact that different input scales often require carefully adjusted hyperparameters (e.g., learning rate schedules and momentum), these factors can cause the performance of YOLOv5-Em to deteriorate rather than improve at the highest resolution.

4.1.2 FPS analysis

In video capture and playback, the number of consecutive images displayed each second is measured using FPS, which is a common unit and metric. FPS is also a common measure for evaluating model performance. Although EfficientDet achieves considerable effects using an end-to-end training method similar to Faster R-CNN in terms of model detection speed, its real-time performance improvement is not significant. In our experimental comparison, YOLOv5-Em outperforms other detectors in terms of detection speed after using the improved EIoU, with an FPS of 62.5. This finding is especially relevant for practical engineering applications, such as ship navigation environment perception, which require real-time detection.

4.2 Accuracy performance evaluation criteria of object detection

Our classification strategy aims to address the complexity of navigation waterways by dividing objects that may affect ship navigation into two classes: navigable and surrounding coastal areas. Instead of classifying ships based on traditional ship classes, we classify objects in the navigable area into seven classes: “workboat”, “cargo ship”, “own body”, “unidentified ship”, “unidentified ship-N”, “packet”, “ship bow”, and “island”. The “unidentified ship-N” category represents ships that are difficult to distinguish at night, which poses a challenge for current DL models because of their large feature scale. For objects that may affect model recognition in the surrounding shore area, we classify them based on the following common main objects: “gantry crane”, “lighthouse”, “bridge”, and “ship lock”. These classes are easy to identify and can help the model distinguish between objects in navigable areas and objects on the shore.

4.2.1 Comparison of object detection mAP among different models

In this subsection, we investigate the impact of input size

on the detection accuracy of each class. Common sense dictates that long-distance and small objects require larger input images to achieve better detection results. Table 4 compares the detection results of YOLOv5 using different training weights. The detection accuracy of three classes: "own body", "bridge", and "ship bow" (highlighted in bold), is generally better than that of the other classes. In our dataset, "lighthouse" has large variations in distance and scale, and its appearance characteristics in different navigation areas considerably differ, leading to low detection accuracy. As "lighthouse" is an important navigational aid object in ship navigation, the volume of data for different navigation areas needs to be increased and more subdivision classes of "lighthouse" need to be added to improve its detection accuracy.

In our classification strategy, "workboat" is another typical object class. Most of the "workboats" are easily identifiable in the image data obtained from the bridge perspective because the background of the object is mostly a pure color. By optimizing the model and increasing the number of convolutional layers, we can achieve higher detection accuracy. Alternatively, the "gantry crane" has obvious color characteristics and often appears in busy port areas. The detection of such objects is positively correlated with the size of the model and the number of labels used.

Table 5 compares the detection results of two SOTA models with ResNet as the backbone network, which is categorized into ResNet50 and ResNet101, and the SOTA models include RetinaNet and EfficientDet. Although the comprehensive mAP of our training results is lower than that of the YOLO series, the RetinaNet model outperforms YOLOv5 in the "gantry crane" class with 19.2% higher detection accuracy. However, the detection accuracy of the four classes "unidentified ship", "unidentified ship-N", "lighthouse", and "ship lock" is significantly lower than that of the YOLOv5 series. This is due to the extremely imbalanced ratio of positive and negative samples labeled in this dataset, where complex samples dominate the vast majority, and the gradient is dominated by simple samples, causing the simplest samples to dominate the loss calculation process. Understanding this reason, we can further optimize the YOLO series model.

Table 4 mAP comparison of various classes under the detector of YOLOv5 series

Type	YOLOv5-Es (448 × 448)	YOLOv5-Es (544 × 544)	YOLOv5-Es (640 × 640)	YOLOv5-Em (448 × 448)	YOLOv5-Em (544 × 448)	YOLOv5-Em (544 × 544)
Gantry crane	45.2	47.6	49.4	50.6	51.1	48.7
Work boat	49.0	54.2	56.3	60.5	61.6	60.3
Island	72.7	69.7	73.0	69.1	73.5	70.3
Cargo ship	62.1	61.2	62.4	67.2	67.6	67.0
Own body	99.5	99.5	99.3	99.5	98.3	98.6
Unidentified ship	38.6	39.4	39.3	37.8	42.6	43.2
Unidentified ship-N	51.1	49.1	60.7	63.5	61.2	66.6
Lighthouse	24.8	49.5	49.5	24.8	49.5	24.8
Bridge	82.2	81.8	79.7	80.4	74.1	78.7
Ship bow	99.5	99.5	99.5	99.5	99.5	99.5
Packet	39.6	44.1	38.6	44.3	47.2	44.0
Ship lock	65.2	61.4	79.5	79.5	79.5	79.5

Table 5 mAP comparison of various classes under the detectors of RetinaNet and EfficientDet

Type	YOLOv5-Es (448 × 448)	YOLOv5-Es (544 × 544)	YOLOv5-Es (640 × 640)	YOLOv5-Em (448 × 448)	YOLOv5-Em (544 × 448)	YOLOv5-Em (544 × 544)
Gantry crane	64.4	60.8	60.2	53.4	64.4	65.8
Work boat	18.0	20.2	17.4	15.5	20.8	30.3
Island	43.7	51.9	51.3	42.3	64.6	69.6
Cargo ship	49.4	45.8	38.1	40.5	56.5	67.6
Own body	100	99.9	100	99.5	98.5	100
Unidentified ship	3.3	5.3	3.2	0.7	3.6	15.3
Unidentified ship-N	9.2	7.7	6.6	5.5	6.0	14.5
Lighthouse	3.1	10.9	4.1	1.0	2.0	28.1
Bridge	41.8	47.1	40.0	41.5	75.5	75.4
Ship bow	100	100	100	99.5	100	100
Packet	11.6	17.5	8.0	9.3	23.7	33.2
Ship lock	8.3	50.8	1.3	1.9	1.0	50.0

The difference between Fast R-CNN and Faster R-CNN lies in the use of a special region proposal method to create region proposals. Instead, Faster R-CNN trains a region proposal network that takes a feature map as the input and outputs the region proposals. Table 6 presents the training results of Faster R-CNN. CenterNet, proposed in 2019, is an anchor-free object detection network that has advantages in speed and accuracy compared with the anchor detection networks used in YOLO, SSD, and Faster R-CNN.

The detailed detection accuracies for each class are presented in Table 6. The detection results of "ship lock", "unidentified ship", and "unidentified ship-N" are lower than those achieved by the YOLO series models, with the detection results of "ship lock" partially zero. These results confirm that an imbalance in the proportions of positive and negative samples can lead to abnormal detection accuracy. Therefore, we need to consider the sample proportions more when designing datasets and selecting labeling strategies. Although the mAP of these two models is no longer dominant compared with the latest version of the YOLO model, their approach to addressing the slow speed of selecting search candidate boxes can provide us with further directions for optimization.

After comparing five SOTA models, several conclusions can be drawn. First, when comparing models, the dataset and labeling strategy need to be considered, as differences can affect performance evaluation beyond just the mAP score. Second, ship navigation environment perception has high demands for object detection, especially in certain tasks, such as autonomous navigation and intelligent perception. As a result, many challenges and bottlenecks still need to be overcome in this area.

The study presents three issues that affect object detection in the subdivision field of ship navigation environment perception. First, as the object moves, changes in object scale and features can lead to object detection failures. Second, long-distance and small objects often have low resolution and blurry appearance, resulting in low learning efficiency or detection failure. Third, small objects are less tolerant to bounding box perturbations than large objects at close range, making it difficult to fit the training. This issue is particularly pronounced when detecting distant ambiguous objects in open seas or busy waterway areas with complex backgrounds. For more intuitive image comparisons, please refer to the next subsection. The use of clear language and short sentences improves readability, whereas a more straightforward description of the three issues increases clarity.

4.2.2 Evaluation criteria for object detection performance by comparing scenarios and models

This subsection presents the detection results of typical scenarios in our ShipNav dataset. To enhance the robustness of the object detection model in complex environments, this study considers the challenges of situational perception in various areas under different weather and sea conditions during ship navigation. Specifically, we propose performance criteria for object detection to guide the design of new architecture.

Most ship navigation involves international trade, and each country and region has its unique features and conditions, including various types of waterway maintenance facilities and objects. The ship situational awareness model must meet the high demands for robustness. Therefore, the dataset used in this study not only covers the main scenarios encountered by various ships but also includes a significant amount of data from the major waterways of the world. The dataset is sourced from Mitsui O.S.K. Lines, Ltd., Japan, and the Internet to ensure wide coverage. To evaluate object detection performance, we use the widely accepted SOTA models in recent years and analyze the detection results under various meteorological and sea conditions.

Figure 3 illustrates the visual detection performance of different models in port operation scenarios. The abscissa shows specific port situations, the ordinate shows different detection models, and the bottom row shows the original

Table 6 mAP comparison of various classes under the detectors of Faster R-CNN and CenterNet

Type	Fast(vgg) (544 × 544)	YOLOv5-Es (640 × 640)	YOLOv5-Es (640 × 640)	YOLOv5-Em (544 × 544)	YOLOv5-Em (544 × 448)	YOLOv5-Em (640 × 640)
Gantry crane	56.4	61.8	57.7	56.9	64.8	68.9
Work boat	26.7	24.7	21.2	20.1	49.1	50.7
Island	58.3	62.3	68.0	70.4	75.9	67.9
Cargo ship	51.8	47.1	51.1	47.1	61.9	53.2
Own body	99.7	100	96.4	99.0	100	100
Unidentified ship	10.8	10.0	8.6	8.9	48.6	47.1
Unidentified ship-N	10.5	13.0	7.5	12.0	53.1	59.6
Lighthouse	49.2	35.7	56.4	21.4	31.2	68.8
Bridge	77.9	71.9	67.1	68.3	78.5	80.9
Ship bow	100	100	100	100	100	100
Packet	20.0	19.1	28.9	26.2	20.6	26.4
Ship lock	0	50.0	50.0	0	0	50.0

image. Figure 3(a) depicts the state of the ship when entering the port. Most of the port scenes encountered by ships on nonscheduled routes are dissimilar, necessitating the improvement of the low-density object detection capability of the model. Figure 3(b) exhibits object detection in complex port areas, where operating ship trajectories are unpredictable and high-density traffic flow is present, thus requiring a robust model. Figure 3(c) illustrates the identification of islands, reefs, and berthing areas during navigation. Figure 3(d) shows that the visibility of ship navigation poses a major challenge to visual perception, as poor visibility reduces recognition accuracy. Figure 3(e) shows that, during berth, visual monitoring of the surroundings of the ship is equally crucial, and camera deployment optimization can address this issue.

Regarding the detection model, in all subdivision scenarios, YOLOv5 and CenterNet exhibit better overall detection performance than the other models. The detail processing capability of the RetinaNet model is uncertain for long-distance complex objects. Moreover, the EfficientDet and Faster R-CNN models cannot deal with low-light scenes. For partially occluded objects, innovative object association approaches are necessary.

Ships sailing across intercontinental oceans, through straits or long-distance canals, encounter various challenges. Figure 4 depicts the object detection results of different models in canal scenarios. Each column shows a different canal situation, and each row shows the detection outcomes for distinct models. Figure 4(a) illustrates the detection of breakwater obstacles, which require detection in a wide and large-scale range. Figure 4(b) describes the detection of bridges on canal waterways by the visual detection model. Objects appearing above the field of view can interfere with normal detection results, and special attention is required during navigation. Figure 4(c) presents the comparative analysis of tugboat detection, and different models have different sensitivities to long-distance and small objects, which can diminish because of line-of-sight occlusion. Figure 4(d) shows the detection results of five models for high-speed ships, and the inference speed of the model is critical for detecting high-speed moving objects. Therefore, the one-stage model is preferred as the backbone model. Figure 4(e) compares the detection results in twilight scenes, where illumination changes pose a significant challenge to object detection algorithms. Intensity mapping can be utilized to preprocess images and improve detection accuracy.

The dazzle phenomenon can compromise human vision and endanger their safety while driving or exercising outdoors. Similarly, vision-based object detection can face similar challenges. In backlight situations, such as the bow of a ship facing the sun or city lights on the shore, the object detection capability of the camera can be limited. Figure 5 illustrates the object detection results in dazzle scenarios. Figure 5(a) illustrates that the water object appears dim because of strong sunlight during sunrise. Figure 5(b) shows that YOLOv5 can detect several small objects in the complicated coastal environment influenced by light, whereas Faster R-CNN repeatedly detects the “packet”. Figure 5(c) shows that EfficientDet fails to detect the “bridge” because

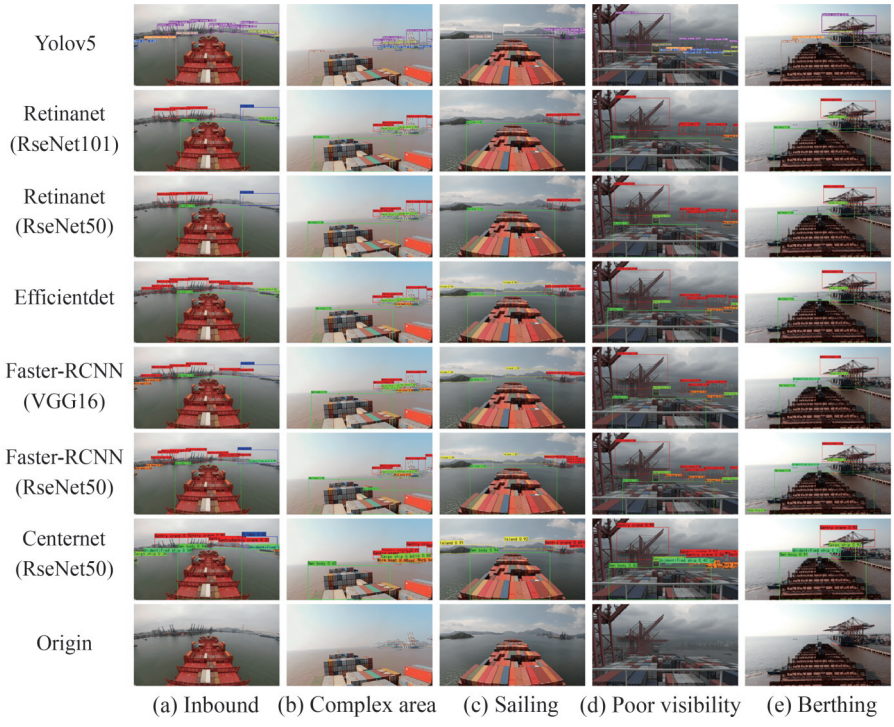

Figure 3 Detection performance of different models in port operation scenarios

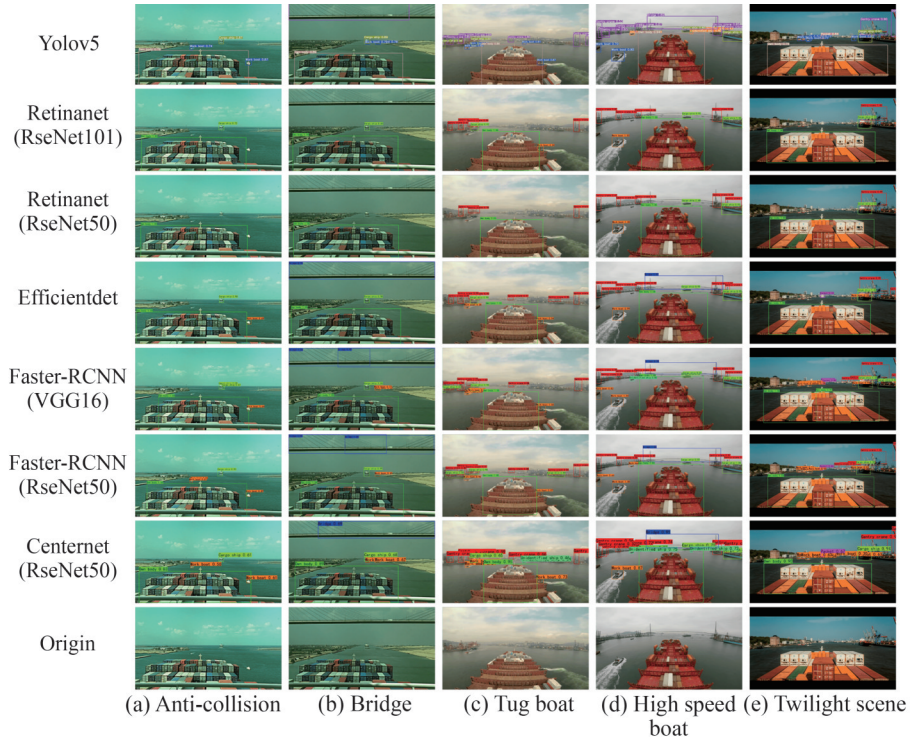

Figure 4 Detection performance of different models in canal navigation scenarios

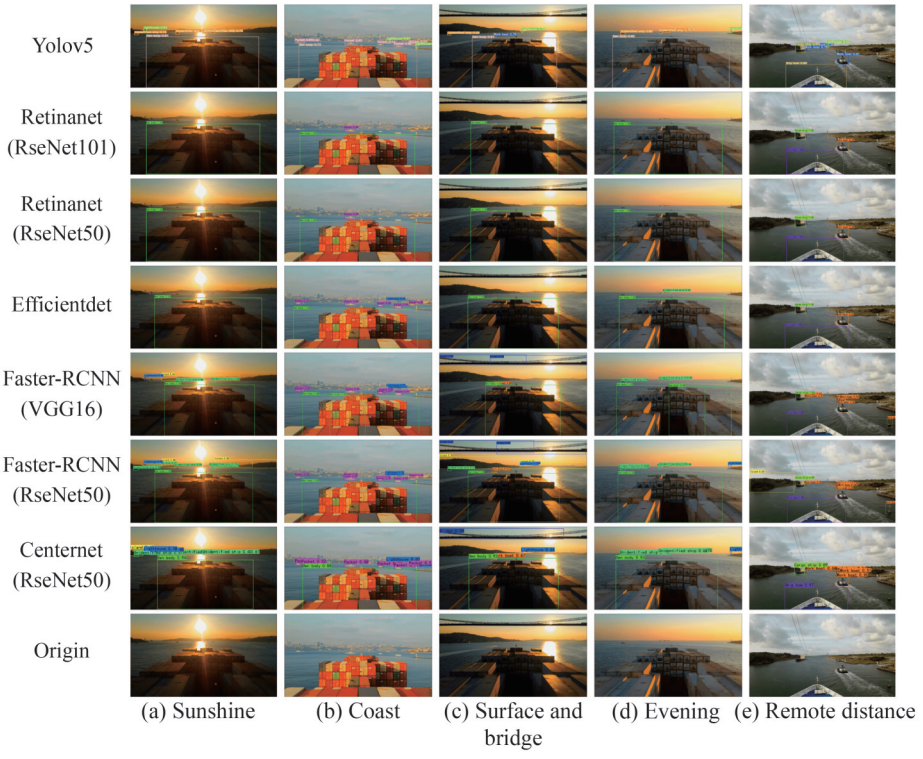

Figure 5 Comparison of object detection results in dazzle conditions

of the strong light. Figure 5(d) indicates that Faster R-CNN outperforms other models at night by detecting more small objects. Figure 5(e) shows that YOLOv5 excels at detecting distant objects, whereas Faster R-CNN and CenterNet perform better in detecting occluded objects, such as the "workboat".

Figure 6 presents the results of object detection in five different night scenarios during ship navigation. Figure 6(a)

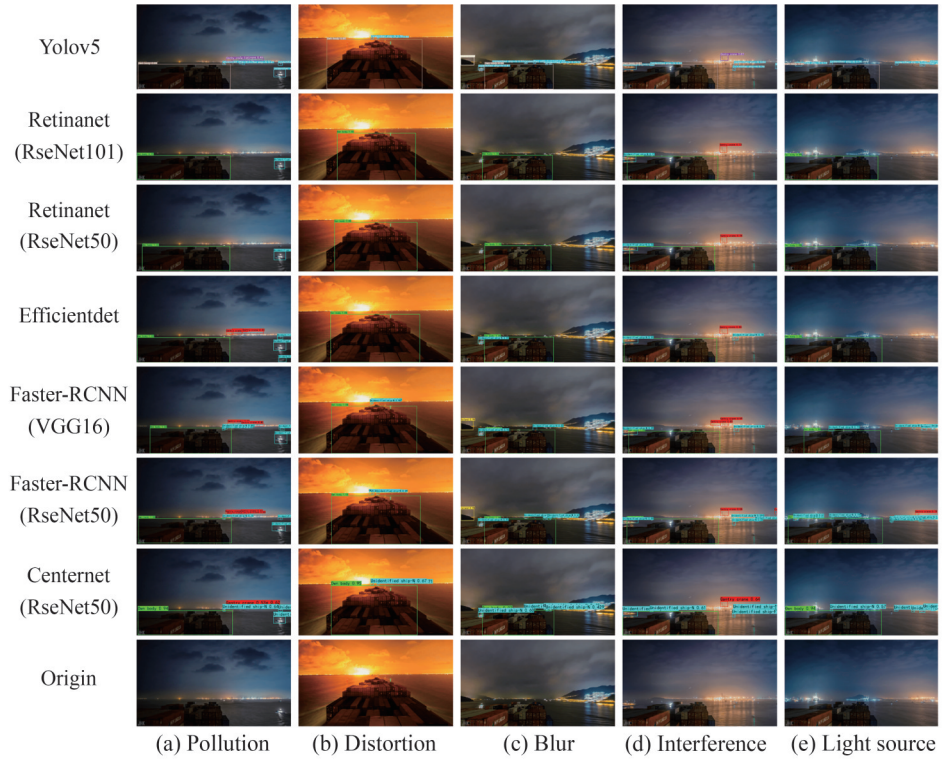

Figure 6 Comparison of the detection results of navigation objects at night

illustrates the impact of light pollution from operating fishing boats, which emit strong lights and can make it challenging for ships to rely on their navigation lights to confirm their position. Figure 6(b) depicts the detection of objects in distorted images obtained when the sun sets, requiring a high-performance camera. The YOLOv5 and Faster R-CNN models with the VGG16 backbone exhibit better detection results. Figure 6(c) compares the recognition results of each model in the presence of blurred images, and both the YOLOv5 and Faster R-CNN models with the VGG16 backbone perform better than the other detectors. Figure 6(d) shows the influence of land light sources, where high light spots from street lamps and vehicle lights on the shore cause fluctuations in illumination and decrease visual salience, leading to difficulty in distinguishing objects. Figure 6(e) further confirms this point.

Numerous studies have proposed different methods to improve the detection of small and long-distance objects. This study demonstrates that the improved IoU can enhance the detection performance of small objects. Figure 7 presents the results of detecting small and long-distance objects in five different environments. Figure 7(a) shows that YOLOv5 significantly outperforms the other models in detecting objects in foggy weather. Figure 7(b) illustrates that only CenterNet can detect more objects because of the unclear features of the unknown ship; however, its overall accuracy is inferior to that of YOLOv5. Figure 7(c) shows the influence of the complex environment of the shore and port on the detection effect of ships entering the port, with RetinaNet exhibiting poor detection results for small objects. Figure 7(d) compares the detection performance of small objects at sea after the ship leaves the port, and YOLOv5 and CenterNet perform better than the other models. Figure 7(e) compares the detection performance of small objects under strong light interference, and Faster R-CNN outperforms the other detectors. Because of the influence of ship navigation, the scale of long-distance and small objects changes significantly, which proves the performance of different SOTA models in multiscale feature calculation.

Based on the results of our comparative experiments and case studies, we determined that no single SOTA model can handle all ship navigation scenarios effectively. In general, different models have their strengths in specific subdivision scenarios. However, our experiments have also demonstrated that the detection performance can decrease significantly in some special scenarios. Therefore, more detailed evaluation benchmarks for object detection models, including dataset preparation, appropriate labeling strategies, model optimization, and evaluation of detection performance in subdivided scenes, need to be established. Widely accepted benchmarks are needed to facilitate the development of more effective detection models.

4.3 Visual perception performance benchmarks

After analyzing the mAP and interpreting the experimental

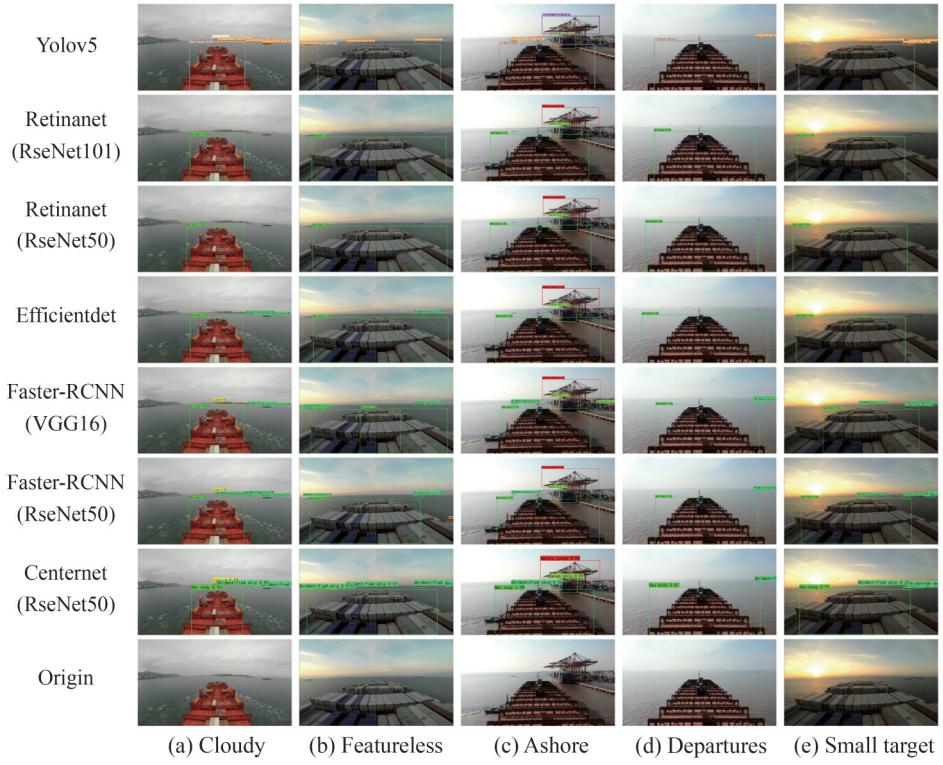

Figure 7 Comparison of the detection results of long-distance and small objects in different scenarios

results, we conclude that data quality, scale calculation, feature quantification, and object association are the benchmarks for visual perception performance in solving the challenges faced by situational awareness in ship navigation.

Data quality. The quality of training data plays a crucial role in the performance of object detection models. To ensure sufficient dataset quantity and balanced classes, we evaluate data quality based on the label consistency and accuracy of annotated data.

"Label consistency" refers to the degree of agreement between different annotators when labeling the same dataset or when the same annotator labels the same class in different scenarios. The labeling results should be consistent to reduce the random noise in labeled data.

The term "accuracy" refers to the degree of proximity between the label and the ground truth. The ground truth data, which is a sample subset of the training data, are labeled by domain experts or data scientists to assess the accuracy of the annotators. The accuracy can be measured using widely recognized benchmarks, which enable data scientists to oversee the quality of their data. To gain insights into the accuracy of the annotators' work and to identify and resolve potential quality declines, the challenges and navigational aids encountered during ship navigation in various scenarios need to be examined.

Scale calculation. The significant variation in the scale of detected objects can result in reduced model accuracy. As ships move, the scales of objects encountered during navigation also change, which can result in missed detections. As accurate object recognition is crucial for ensuring safe ship navigation, the achievement of precise object detection and tracking is a daunting challenge. Therefore, the scale calculation of objects within the same class needs to be considered.

The CNN-based backbone network for object detection utilizes multilayer convolution to extract in-depth information, generate multilayer feature maps, and process further for positioning and classification based on the deep feature maps. To solve the problem of different object scales, several methods, such as "feature image pyramid" and "data augmentation" can be utilized.

The feature image pyramid method creates a series of images with varying resolutions from the same image. As the level of the pyramid changes, the bounding boxes labeled in the image also produce multiple scales ranging from large to small. By inputting these pyramid images of different scales into various SOTA models, the object scale that the model is capable of processing can always be found in a certain layer of the pyramid, even if the model is only proficient at recognizing objects within a specific scale.

Data augmentation is a technique that artificially expands the training dataset by generating additional equivalent data from limited data. Data augmentation effectively reduces overfitting of the model, particularly for small sample size datasets and the detection of low-density scene objects. However, the generated data may introduce noise because of

differences with real data. To evaluate model performance benchmarks in various low-density scene samples of ship navigation, different methods, including image processing, geometric and photometric conversion, and DL, need to be used.

Feature quantification. The use of DL for object detection involves end-to-end training, which includes image input, adaptive object feature extraction, classification, and regression. Despite varying appearances of channel objects in different regions, seafarers can identify them based on navigation materials or experience. Thus, during the feature extraction stage, the features should be quantified to enhance the robustness of the model. To ensure reproducibility, researchers should provide a detailed description of the internal structure of their model, including its network structure, loss function, and activation function.

Object association. In ship navigation, object detection is a fundamental task. Other essential requirements include object tracking, object state estimation, and more. However, object association is a prerequisite for object tracking, which requires detecting all objects of interest in each video frame and correlating them with the objects detected in the previous frame for tracking. The object association benchmarks are evaluated and measured by calculating the degree of association between objects. This association calculation is based on object similarity, texture similarity, and local color similarity, and matching is performed based on different dimensions: overall and local features, regional features, and long-term features.

5 Discussion

5.1 Detailed introduction and comparison of the datasets

Our ShipNav dataset is designed specifically for visual perception tasks in ship navigation, capturing diverse scenarios, such as different waterways, weather conditions, day/night illumination, and a variety of vessel types. This focus on real-world complexity sets it apart from many existing maritime datasets, which may concentrate on single perspectives (e. g., aerial or port-centric) or narrower geographic regions.

Nonetheless, a comprehensive comparison with other open-source maritime datasets (e. g., SeaShips, ABOships, and WSODD) would indeed be valuable. These datasets often differ in their labeling standards, object categories, image resolutions, and domain coverage. For instance, SeaShips labels only a handful of vessel categories, whereas ShipNav covers a broader set of classes, including navigational aids and shore-based infrastructure relevant to ship maneuvers. A direct, quantitative comparison is challenging because of inconsistent annotation guidelines and class definitions; however, future work could attempt to unify labeling schemes and performance metrics to enable a more rigorous benchmark across multiple datasets.

5.2 Combining ShipNav with other related datasets

Another related question is whether combining ShipNav with external datasets could further improve algorithm performance. In principle, the aggregation of multiple datasets can diversify training samples, improve generalization, and bolster the robustness of DL models. For example, incorporating images from other regions or featuring different vessel types can help address domain shift problems that arise when deploying a model in new environments.

However, practical challenges remain. Label inconsistency, class definition mismatches, and variance in annotation quality may necessitate extensive relabeling or domain adaptation techniques to ensure that the merged data are truly beneficial. Moreover, differences in resolution, sensor characteristics, and viewpoint can introduce further complexity when harmonizing datasets. Addressing these issues will require careful data curation and possibly advanced domain adaptation methods.

Despite these challenges, we believe that ShipNav provides an essential foundation for analyzing complex ship navigation scenarios, particularly from the bridge perspective. In future work, we plan to explore semiautomatic relabeling strategies and domain adaptation techniques to combine ShipNav more seamlessly with other maritime datasets, thereby expanding the coverage of diverse conditions and vessel types. This approach has the potential to create a unified, large-scale dataset that can serve as a more comprehensive benchmark for maritime visual perception research.

6 Conslusions

In this study, we present a dataset for visual object detection in ship navigation, which includes images of important navigational waters worldwide. We propose evaluation benchmarks for various subdivision scenarios and tasks related to ship navigation. We evaluated the performance of four widely used object detection algorithms with different backbones (i. e., YOLOv5, Faster R-CNN, CenterNet, and EfficientDet) and improved YOLOv5. Our experiments showed that YOLOv5-E with enhanced EIoU outperforms the other models in terms of comprehensive performance. However, different improvement methods may compromise other performance metrics in specific subdivision scenarios. The results showed that the SOTA object detection models exhibit uneven performance in specific real-world scenarios. Therefore, we explore factors that may affect

model performance evaluation benchmarks from the perspective of data quality, scale calculation, feature quantification, and object association. These factors should be considered in the formulation of performance evaluation benchmarks with depth and breadth in actual autonomous ship navigation practice.

Abbreviations

ShivNav	ShivNav benchmark
SOTA	State-of-the-art
COCO	Common objects in context
PASCAL	Pattern Analysis, Statistical modeling, and ComputAtional Learning
VOC	Visual Object Classes
Conv	Convolution
IoU	Intersection-over-union
mAP	Mean average precision
net101	ResNet101
net50	ResNet50
vgg	VGG16
AI	Artificial intelligence

Acknowledgement This work was partially funded by the International Association of Maritime Universities (IAMU) and The Nippon Foundation in Japan. The authors would like to acknowledge the support of the International Association of Maritime Universities (Research Project Number 20240201), The authors also gratefully acknowledge the support from the China Scholarship Council (Grant No. CXXM2209260070).

Competing interest The authors have no competing interests to declare that are relevant to the content of this article.

References

Bochkovskiy A, Wang CY, Liao HYM (2020) Yolov4: Optimal speed and accuracy of object detection. Computer Science, Computer Vision and Pattern Recognition, arXiv preprint arXiv: 2004.10934. https://doi.org/10.48550/arXiv.2004.10934

Borkar S, Ghutke P, Patil W, Joshi S, Sorte S (2023) A review of pick and place robots for the pharmaceutical industry. 11th International Conference on Emerging Trends in Engineering & Technology-Signal and Information Processing (ICETET-SIP), IEEE, Nagpur, India, 1-6. DOI: 10.1109/ICETET-SIP58143.2023.10151652

Cai J, Chen G, Yin J, Ding C, Suo Y, Chen J (2024) A Review of Autonomous Berthing Technology for Ships. Journal of Marine Science and Engineering 12(7): 1137. https://doi.org/10.3390/jmse12071137

Cavegn S, Haala N, Nebiker S, Rothermel M, Tutzauer P (2014) Benchmarking high density image matching for oblique airborne imagery. The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences 40(3): 45. https://doi.org/10.5194/isprsarchives-XL-3-45-2014

Chai J, Zeng H, Li A, Ngai EW (2021) Deep learning in computer vision: A critical review of emerging techniques and application scenarios. Machine Learning with Applications 6: 100134. https://doi.org/10.1016/j.mlwa.2021.100134

Chen B, Ghiasi G, Liu H, Lin TY, Kalenichenko D, Adam H, Le QV (2020) MnasFPN: Learning latency-aware pyramid architecture for object detection on mobile devices. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 13607-13616

Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R, Franke U, Roth S, Schiele B (2016) The Cityscapes dataset for semantic urban scene understanding. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 3213-3223

Deng J, Dong W, Socher R, Li LJ, Li K, Li FF (2009) ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, USA, 248-255. DOI: 10.1109/CVPR.2009.5206848

Duan K, Bai S, Xie L, Qi H, Huang Q, Tian Q (2019) CenterNet: Keypoint triplets for object detection. Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 6569-6578

Durlik I, Miller T, Cembrowska-Lech D, Krzemińska A, Złoczowska E, Nowak A (2023) Navigating the sea of data: a comprehensive review on data analysis in maritime IoT applications. Applied Sciences 13(17): 9742. https://doi.org/10.3390/app13179742

Er MJ, Chen J, Zhang Y, Gao W (2023) Research challenges, recent advances, and popular datasets in deep learning-based underwater marine object detection: A review. Sensors 23(4): 1990. https://doi.org/10.3390/s23041990

Everingham M, Van Gool L, Williams CK, Winn J, Zisserman A (2010) The PASCAL Visual Object Classes (VOC) challenge. International Journal of Computer Vision 88(2): 303-338. https://doi.org/10.1007/s11263-009-0275-4

Girshick R (2015) Fast R-CNN. Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 1440-1448

Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, USA, 580-587

Hackel T, Savinov N, Ladicky L, Wegner JD, Schindler K, Pollefeys M (2017) Semantic3d. net: A new large-scale point cloud classification benchmark. Computer Science, Computer Vision and Pattern Recognition, arXiv preprint arXiv:1704.03847. https://doi.org/10.48550/arXiv.1704.03847

Han X, Zhao L, Ning Y, Hu J (2021) ShipYolo: an enhanced model for ship detection. Journal of Advanced Transportation 2021(1): 1090182. https://doi.org/10.1155/2021/1060182

He J, Erfani S, Ma X, Bailey J, Chi Y, Hua XS (2021) α-IoU: A family of power intersection over union losses for bounding box regression. 35th Conference on Neural Information Processing Systems, 1-13

He K, Zhang X, Ren S, Sun J (2015) Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 37(9): 1904-1916. DOI: 10.1109/TPAMI.2015.2389824

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 770-778

Henderson P, Ferrari V (2016) End-to-end training of object class detectors for mean average precision. Asian Conference on Computer Vision, Springer, Cham, 198-213. https://doi.org/10.1007/978-3-319-54193-8_13

Howard A, Sandler M, Chen B, Wang W, Chen LC, Tan M, Chu G, Vasudevan V, Zhu Y, Pang R, Adam H, Le Q (2019) Searching for mobilenetv3. Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 1314-1324

Hussain M, Saher N, Qadri S (2022) Computer vision approach for liver tumor classification using CT dataset. Applied Artificial Intelligence 36(1): 2055395. https://doi.org/10.1080/08839514.2022.2055395

Iancu B, Soloviev V, Zelioli L, Lilius J (2021) ABOships—An inshore and offshore maritime vessel detection dataset with precise annotations. Remote Sensing 13(5): 988. https://doi.org/10.3390/rs13050988

Idrees H, Tayyab M, Athrey K, Zhang D, Al-Maadeed S, Rajpoot N, Shah M (2018) Composition loss for counting, density map estimation and localization in dense crowds. Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 532-546

Islam MA, Mobarak MH, Rimon MIH, Al Mahmud MZ, Ghosh J, Ahmed MMS, Hossain N (2024) Additive manufacturing in polymer research: Advances, synthesis, and applications. Polymer Testing 132: 108364

Ismail N, Malik OA (2022) Real-time visual inspection system for grading fruits using computer vision and deep learning techniques. Information Processing in Agriculture 9(1): 24-37. https://doi.org/10.1016/j.polymertesting.2024.108364

Jocher G (2020) YOLOv5 by Ultralytics (Version 7.0). Computer software. https://doi.org/10.5281/zenodo.3908559

Karas V, Schuller DM, Schuller BW (2023) Audiovisual affect recognition for autonomous vehicles: Applications and future agendas. IEEE Transactions on Intelligent Transportation Systems 25(6): 4918-4932. DOI: 10.1109/TITS.2023.3333749

Kaur R, Singh S (2023) A comprehensive review of object detection with deep learning. Digital Signal Processing 132: 103812. https://doi.org/10.1016/j.dsp.2022.103812

Khan W, Zaki N, Ali L (2021) Intelligent pneumonia identification from chest x-rays: A systematic literature review. IEEE Access 9: 51747-51771. DOI: 10.1109/ACCESS.2021.3069937

Lenka AK, Tripathy HK (2024) 5 Computer vision for medical diagnosis and surgery. Healthcare Big Data Analytics: Computational Optimization and Cohesive Approache, De Gruyter, Berlin, 101-124. https://doi.org/10.1515/9783110750942-005

Li Y, Moreau J, Ibanez-Guzman J (2023) Emergent visual sensors for autonomous vehicles. IEEE Transactions on Intelligent Transportation Systems 24(5): 4716-4737. DOI: 10.1109/TITS.2023.3248483

Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft COCO: Common objects in context. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T (Eds.). Computer Vision–ECCV 2014 (ECCV 2014). Springer, Cham, 740-755. https://doi.org/10.1007/978-3-319-10602-1_48

Lin TY, Goyal P, Girshick R, He K, Dollár P (2017) Focal loss for dense object detection. Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 2980-2988

Liu S, Gao C, Chen Y, Peng X, Kong X, Wang K, Xu R, Jiang W, Ma J, Wang M (2023) Towards vehicle-to-everything autonomous driving: A survey on collaborative perception. Computer Science, Computer Vision and Pattern Recognition, arXiv preprint arXiv: 2308.16714

Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, Berg AC (2016) SSD: Single shot multibox detector. European Conference on Computer Vision, Springer, Cham, 21-37. https://doi.org/10.1007/978-3-319-46448-0_2

Liu Y, Lu B, Peng J, Zhang Z (2020) Research on the use of YOLOv5 object detection algorithm in mask wearing recognition. World Scientific Research Journal 6(11): 276-284. DOI: 10.6911/WSRJ.202011_6(11).0038

Liu Z, Luo P, Wang X, Tang X (2018) Large-scale celebfaces attributes (celeba) dataset. Retrieved August 15(2018): 11

Long X, Deng K, Wang G, Zhang Y, Dang Q, Gao Y, Wen S (2020) PP-YOLO: An effective and efficient implementation of object detector. arXiv preprint arXiv:2007.12099. https://doi.org/10.48550/arXiv.2007.12099

Manakitsa N, Maraslidis GS, Moysis L, Fragulis GF (2024) A review of machine learning and deep learning for object detection, semantic segmentation, and human action recognition in machine and robotic vision. Technologies 12(2): 15. https://doi.org/10.3390/technologies12020015

Menze M, Geiger A (2015) Object scene flow for autonomous vehicles. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 3061-3070

Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: Unified, real-time object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 779-788

Ren S, He K, Girshick R, Sun J (2015) Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence 39(6): 1137-1149. DOI: 10.1109/TPAMI.2016.2577031

Shao Z, Wu W, Wang Z, Du W, Li C (2018) Seaships: A large-scale precisely annotated dataset for ship detection. IEEE Transactions on Multimedia 20(10): 2593-2604. DOI: 10.1109/TMM. 2018.2865686

Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. Computer Science, Computer Vision and Pattern Recognition, arXiv preprint arXiv:1409.1556

Tan M, Le Q (2019) Efficientnet: Rethinking model scaling for convolutional neural networks. Proceedings of the 36th International Conference on Machine Learning, Long Beach, USA, 6105-6114

Tan M, Pang R, Le QV (2020) EfficientDet: Scalable and efficient object detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 10781-10790

Voulodimos A, Doulamis N, Doulamis A, Protopapadakis E (2018) Deep learning for computer vision: A brief review. Computational Intelligence and Neuroscience 2018(1): 7068349. https://doi.org/10.1155/2018/7068349

Yan B, Peng H, Fu J, Wang D, Lu H (2021) Learning spatio-temporal transformer for visual tracking. Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, Canada, 10448-10457

Yu J, Zhang C, Wang S (2021) Multichannel one-dimensional convolutional neural network-based feature learning for fault diagnosis of industrial processes. Neural Computing and Applications 33(8): 3085-3104. https://doi.org/10.1007/s00521-020-05171-4

Zhang R, Ji X, Pan M (2022) Diversified assessment benchmark of vision dataset-based perception in ship navigation scenario. Proceedings of the 2022 5th International Conference on Signal

Processing and Machine Learning, Dalian, China, 282-287. https://doi.org/10.1145/3556384.3556427

Zhang YF, Ren W, Zhang Z, Jia Z, Wang L, Tan T (2022) Focal and efficient IOU loss for accurate bounding box regression. Neurocomputing 506: 146-157. https://doi.org/10.1016/j.neucom.2022.07.042

Zhou B, Lapedriza A, Khosla A, Oliva A, Torralba A (2017) Places: A 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 40(6): 1452-1464. DOI: 10.1109/TPAMI.2017.2723009

Zhou Z, Sun J, Yu J, Liu K, Duan J, Chen L, Chen CP (2021) An image-based benchmark dataset and a novel object detector for water surface object detection. Frontiers in Neurorobotics 15: 723336. https://doi.org/10.3389/fnbot.2021.723336