Journal of Southeast University (English Edition) Vol. 40, No. 1, pp. 33 – 40 Mar. 2024 ISSN 1003—7985

# Vision-based vessel detection for vessel-bridge collision warnings under complex scenes

Liao Ruixuan Wu Tong Zhang Yiming Mao Jianxiao Wang Hao

(Key Laboratory of Concrete and Prestressed Concrete Structures of Ministry of Education, Southeast University, Nanjing 211189, China)

**Abstract:** To enable accurate vessel recognition for bridge collision avoidance and early warning, an image dataset for vessels in bridge channels is established using cameras and data augmentation. This dataset includes complex scenarios such as long distances, multiple targets, and low visibility. Subsequently, the you-only-look-once version 5 (YOLOv5) model is employed as the basic detector, and several modifications are applied to its network structure. Key enhancements involve replacing C3 modules in the backbone network with C2f modules, integrating the squeeze-excitation attention mechanism into the feature fusion network, and optimizing the prior anchors of the dataset using the K-means++ clustering algorithm. Finally, the modified model undergoes training and validation using PyTorch as the deep learning framework. Results demonstrate that the mean average precision for crucial vessels in the modified YOLOv5 model reaches 99.4%, representing an 11.1% improvement compared to the original YOLOv5 model. Additionally, the inference speed is measured at 102 frame/s. The established YOLOv5 model is a reliable and efficient cornerstone for warning against vessel-bridge collisions in complex navigable scenes.

**Key words:** vessel detection; vessel-bridge collision; you-only-look-once version 5 (YOLOv5); squeeze-excitation attention mechanism; data augmentation

**DOI:** 10.3969/j.issn.1003 – 7985.2024.01.004

The rapid development of water transport has substantially increased the quantity and tonnage of vessels in channels. Thus, the navigation environment near bridges is becoming increasingly sophisticated, raising the risks of vessel-bridge collisions[1]. Protective structures are usually arranged around piers to avoid vessel-bridge collisions or reduce the impact force, but this passive approach cannot prevent collisions. Identifying and warning vessels with a high risk of colliding with a bridge is an active way to avoid this problem. Vessel detection is the primary step to warn of collisions. After the vessel targets are detected, their geographical coordinates can be obtained by homography[2], which provides information support for early warning. Therefore, the recognition accuracy and classification performance of vessel targets are crucial for the anti-collision measures of bridges and the navigation planning of vessels[3].

In general, vessel detection methods are classified into manual observation, shallow feature-based methods, and deep learning-based methods[4]. Particularly, manual observation involves reliability, and it is difficult to find distant vessels in time and accurately judge the characteristics of vessels. Shallow feature-based methods generally include three steps: region selection, feature extraction, and classification. In particular, feature extraction mainly refers to artificially designed features, such as edges, textures, and colors. The scale of these features remains constant during the detection process, resulting in the poor performance of shallow feature-based methods in sophisticated environments[5]. Recently, driven by pixel-level feature extraction of deep neural networks, deep learning-based target detection methods have begun to emerge[4]. Such methods can be divided into two categories: two-stage target detection algorithms, generally including R-CNN, Fast R-CNN, and Faster R-CNN, and single-stage target detection algorithms, mainly including SSD and YOLO series[5–7]. The two-stage detection algorithm has a long inference time due to the candidate region. The YOLO series has good accuracy and fast inference speed, which play an important role in one-stage detectors in object detection tasks. For instance, Shao et al.[8] built a large-scale dataset of vessels by monitoring cameras in a deployed coastline video surveillance system called SeaShips. The reliability of SeaShips was verified using YOLOv2, which advances research and applications on vessel detection. Li et al.[9] proposed an enhanced YOLOv3 tiny network for real-time vessel detection, and an attention module named CBAM was introduced into the backbone network. The detection accuracy of the proposed algorithm on the SeaShips dataset was better than that of the original YOLOv3 model. Lee et al.[10] constructed a virtual image-based dataset using Unity to overcome the difficulty of obtaining vessel images, and vessel detection was performed using the deep learning-based detection model. Thus, the entire detection

**Received** 2023-07-13, **Revised** 2023-10-26.

**Biographies:** Liao Ruixuan (1999—), male, Ph. D. candidate; Wang Hao (corresponding author), male, doctor, professor, wanghao1980@seu.edu.cn.

**Foundation items:** The National Natural Science Foundation of China (No. 51978155, 52208481).

**Citation:** Liao Ruixuan, Wu Tong, Zhang Yiming, et al. Vision-based vessel detection for vessel-bridge collision warnings under complex scenes[J]. Journal of Southeast University (English Edition), 2024, 40(1): 33 – 40. DOI: 10.3969/j.issn.1003 – 7985.2024.01.004.

showed a good performance. Existing studies generally use open datasets or self-built datasets to train deep-learning models for vessel detection. These datasets do not cover complex navigable scenarios in channels near bridges, so the accuracy and robustness of deep-learning models cannot meet the requirements of vessel-bridge collision warnings.

In this study, to achieve the first step of vessel-bridge collision warnings, a dataset for the detection task of vessels in channels near bridges is augmented, and a modified YOLOv5-based detector is proposed. C2f modules and squeeze-excitation (SE) attention mechanisms are integrated into the detector. Finally, the accuracy and robustness of the modified YOLOv5 model are verified using the augmented dataset.

## 1 Methodology for Computer Vision-based Vessel Detection

### 1.1 Overview of the base detector YOLOv5

As the representatives of the YOLO series, YOLOv3, YOLOv4, and YOLOv5 have greatly improved the accuracy and speed of network detection by introducing residual network structure, data enhancement, and a focus module. Among these models, YOLOv5 has higher detection accuracy and detection speed and is easier for network deployment. Thus, YOLOv5 is selected as the base detector for vessel detection in this paper. The YOLOv5 network comprises four parts: the input, the backbone network, the neck network, and the head[11]. Its original architecture is shown in Fig. 1.

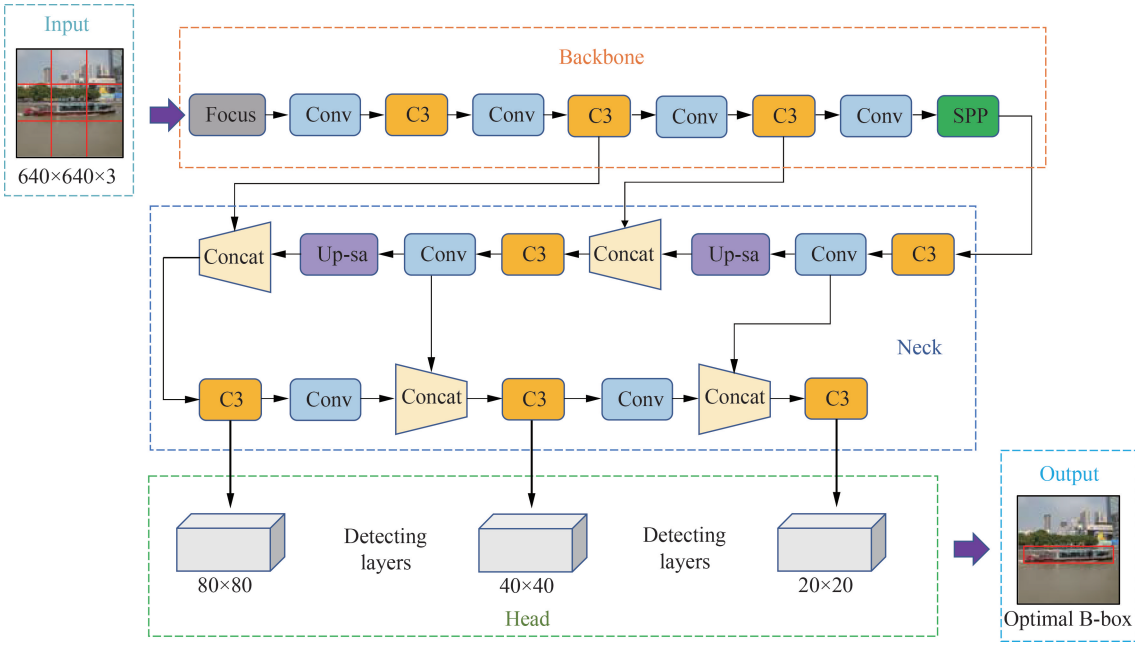



**Fig. 1** Architecture of the original YOLOv5 network

As shown in Fig. 1, the input images of different sizes are scaled to 640 × 640 pixels. The backbone network is used to extract image features. In the backbone, the focus module is used to slice the image and halve its pixel size. Then, the image features are extracted by Conv, C3, and SPP modules. The Conv module is used to double the number of channels in the feature map. The C3 module is a residual structure composed of the Conv and the Bottleneck. The C3 module adopts the idea of extraction and diversion of CSPNet, which can enhance feature extraction ability. The SPP module takes the maximum pooling of the image features with convolution kernel sizes of 5 × 5, 9 × 9, and 13 × 13, and then feature extraction efficiency is increased. The neck network is used to fuse the partially extracted features of the backbone network[6-7]. The detection head is based on the non-maximum suppression to output bounding boxes. There are three detecting layers in the head: 80 × 80 for small objects, 40 × 40 for medium objects, and 20 × 20 for large objects.

### 1.2 Modifications of YOLOv5 for vessel detection

YOLOv5 has achieved good results on open datasets, e. g., PASCAL VOC[11] and COCO[12]. These datasets only include a few ship targets, underlying the performance of YOLOv5 for detecting vessels under sophisticated scenarios with long distance, multi-target detection, and low visibility[12]. Therefore, some modifications of YOLOv5's architecture are proposed to enhance the generalization ability and detection performance of the network. The modifications are as follows: Five C3 modules in the backbone and neck networks are replaced by C2f modules, and two SE attention mechanisms are embedded in the backbone and neck networks.

The C2f module combines the idea of ELAN structure based on the principle of the C3 module, which can make the backbone network obtain more abundant gradient flow information while ensuring its lightweight[9]. The SE attention mechanism is essentially a squeeze-excitation network (SENet). The structure of SENet is shown in Fig. 2[13].

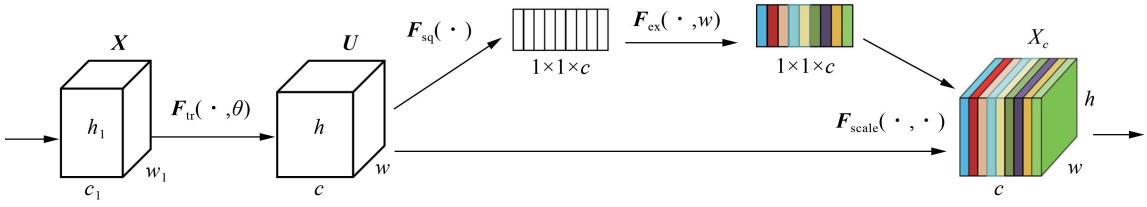



**Fig. 2** Structure of SENet

In Fig. 2, $\boldsymbol{F}_{tr}$ represents the convolution operation; $\boldsymbol{F}_{sq}$ represents the compression of features; $\boldsymbol{F}_{ex}$ represents feature extraction; $\boldsymbol{F}_{scale}$ represents the recalibration of features; $\boldsymbol{X}$ represents the input feature map; $\boldsymbol{U}$ represents a new feature map; $h_1$, $w_1$, and $c_1$ represent the length, width, and number of channels of the input feature map, respectively; $h$, $w$, and $c$ represent the length, width, and number of channels of the output feature map, respectively; and $\boldsymbol{X}_c$ represents the output feature map.

First, the input feature map $\boldsymbol{X}$ is transformed to generate a new feature map $\boldsymbol{U}$. Then, $\boldsymbol{U}$ is squeezed, and the global spatial information is squeezed into a channel descriptor.

$$z_c = \boldsymbol{F}_{sq}(u_c) = \frac{1}{hw}\sum_{i=1}^{h}\sum_{j=1}^{w} u_c(i,j) \tag{1}$$

where $z_c$ represents the global descriptor for the $c$-th channel, and $u_c$ represents the value of each point on the characteristic graph channel.

Second, the complexity degree of the model is reduced and the generalization ability of the model is enhanced. The calculation of the excitation factor can be expressed as

$$s_c = \boldsymbol{F}_{ex}(z_c, w) = \sigma[W_2\delta(W_1 z_c)] \tag{2}$$

where $s_c$ represents the excitation factor for the $c$-th channel; $\sigma$ and $\delta$ are the ReLU and sigmoid activation functions, respectively; and $W_1$ and $W_2$ are the weights of the dimension reduction and dimension increase, respectively.

The final output of the block is obtained by rescaling $\boldsymbol{U}$ with the activations $s_c$:

$$\boldsymbol{X}_c = \boldsymbol{F}_{scale}(u_c, s_c) = s_c u_c \tag{3}$$

The addition of the C2f and SENet can improve the performance of convolutional neural networks, which is conducive to improving the detection accuracy of small targets and the generalization ability of the model. Consequently, the modifications of YOLOv5's architecture are shown in Fig. 3.

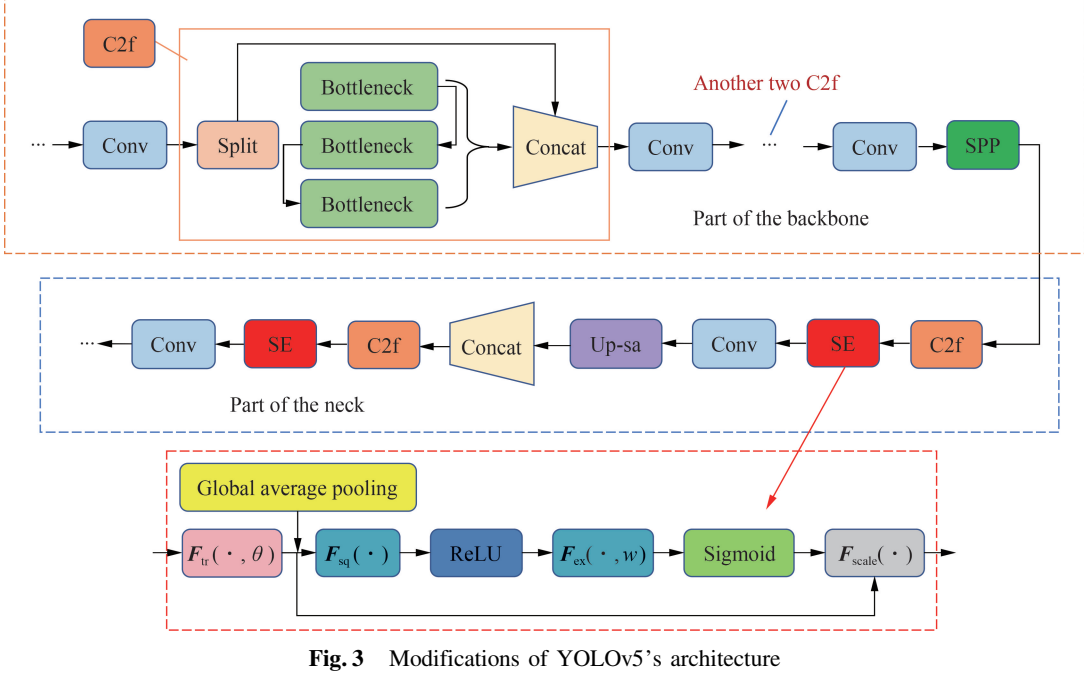



**Fig. 3** Modifications of YOLOv5's architecture

### 1.3 Data collection and augmentation

A total of 568 images are selected from an open dataset provided by IC-SHM 2022[14], as well as from captured images. These images mainly include four common types of ships: fishing ships, cargo vessels, sand vessels, and cruise vessels. Fishing ships are small vessels, whereas the other ships have large tonnage. Their impact on a bridge will cause great structural response and even affect structural safety, so they are crucial objects for detection. However, the amount of the collected data is relatively small, which may lead to overfitting problems in the process of model training. Therefore, the data augmentation technique is used to increase the diversity and complexity of the data. The data augmentation methods include translation, clipping, rotation, mirroring, changing brightness, increasing noise, and cutout[15]. A total of 3 700 images are selected to build the dataset, of which 2 220 are used for training, 740 for validation, and 740 for testing. The dataset covers all possible image changes, e. g., backgrounds, angles, and times. All images are (RGB) three-channel images and 1 920 × 1 080 pixels in size. Example images of four ship types in the dataset are shown in Fig. 4.

(a) (b) (c) (d)

**Fig. 4** Example images of four ship types. (a) Fishing ships; (b) Cargo vessels; (c) Sand vessels; (d) Cruise vessels

Then, LabelImg software, a commonly used image annotation tool, is used to annotate the dataset. Consequently, 4 674 vessel targets are annotated in the dataset.

### 1.4 Optimization of priori anchors

The priori anchor in YOLOv5 can be defined as the most likely width and height of an object to detect objects concentrated in agrid[11]. This definition adds another dimension to the output label. The numbers and sizes of anchors affect the detection speed and precision. The anchors' sizes of the original YOLOv5 model are obtained based on the clustering of the COCO dataset. Although the COCO dataset includes many categories, the anchors are more suitable for indoor objects and less relevant to vessels[12], which may lead to a deviation in the positioning accuracy of targets. To make anchors more suitable for vessel targets and improve positioning accuracy, the K-means++ algorithm is applied to automatically determine the priori anchors in the training dataset. Cluster analysis results based on K-means++ are shown in Fig. 5.

The anchors' sizes of each detecting layer before and after clustering are shown in Table 1. Then, the sizes of anchors after clustering are used to correct the previous anchors.

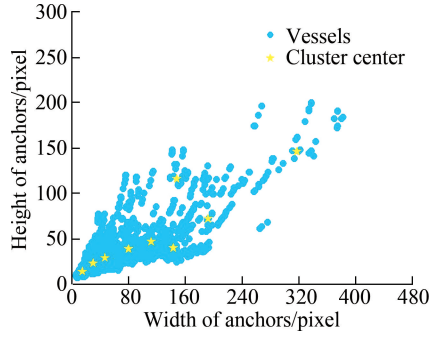



**Fig. 5** Cluster analysis results based on K-means++

**Table 1** Anchors before and after clustering pixel

| Detecting layers | Anchors before clustering ($W$, $H$) | Anchors after clustering ($W$, $H$) |
|---|---|---|
| 80 × 80 | (10, 13), (16, 30), (33, 23) | (15, 14), (30, 23), (47, 29) |
| 40 × 40 | (30, 61), (62, 45), (59, 119) | (80, 39), (112, 47), (143, 40) |
| 20 × 20 | (116, 90), (156, 198), (373, 326) | (192, 72), (148, 116), (317, 146) |

Note: ($W$, $H$) represents the width and height of anchors.

## 2 Case Study

### 2.1 Experimental configuration and evaluation indicators

The operating system is Windows 10, the GPU model is NVIDIA GeForce RTX 1080Ti, the compiler language is Python 3.8.3, the deep learning framework is PyTorch 1.7.0, and the CUDA version is 10.2. The gradient descent algorithm is used to train 160 epochs throughout the process.

In this experiment, the average precision (AP) and mean average precision (mAP) are selected as the evaluation indexes of model detection ability. The AP comprehensively reflects the detection accuracy of one category. According to the different IOU thresholds, two types of AP indicators are determined: $AP_{0.5}$ and $AP_{0.5:0.95}$. The mAP reflects the detection accuracy of all categories, and there are also the $mAP_{0.5}$ and $mAP_{0.5:0.95}$.

### 2.2 Model training and verification

Four groups of models are trained: modified YOLOv5 + AUG + anchors, modified YOLOv5 + AUG, original YOLOv5 + AUG, and original YOLOv5. "+ AUG" represents the use of data augmentation in the model, and "+ anchors" represents the use of the K-means + + method to optimize the anchors. The mAP and loss curves for the four groups of models are shown in Fig. 6.

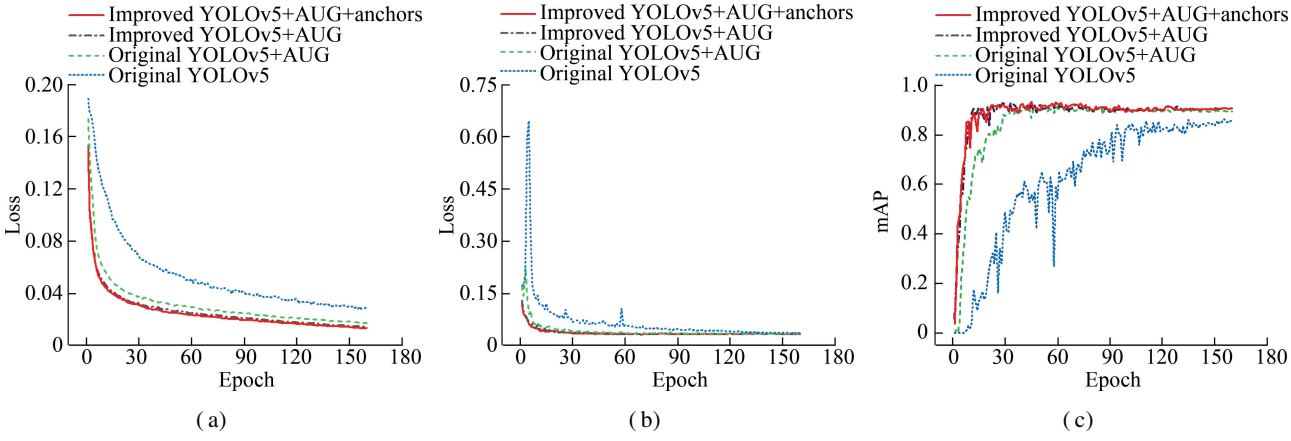



**Fig. 6** Loss curves and mAP for four groups of models. (a) Loss of training process; (b) Loss of validation process; (c) mAP

The loss curves demonstrate that the four models converge in the same way during the training and validation processes. Fig. 6(a) shows that in the training process, with different degrees of model modifications, the losses of the models become increasingly smaller, and the loss of the original YOLOv5 model decreases substantially after data augmentation. This result shows data augmentation, network architecture modifications, and priori anchors' optimization help enhance detector performance, and data augmentation is the most useful approach when the data volume is small. From Fig. 6(b), in the validation process, the modified YOLOv5 model converges faster and has smoother loss curves than the original YOLOv5 model. The main reason is the introduced C2f and SE modules strengthen the models' generalization capability. Fig. 6(c) shows the value and growth rate of the mAP of the original YOLOv5 model is the lowest. The mAP curves of the other three models are generally consistent. The modified YOLOv5 + AUG + anchors and modified YOLOv5 + AUG models take approximately 20 epochs to converge, and the original YOLOv5 + AUG model takes approximately 40 epochs to converge, indicating that the former models require fewer resources for training and are more lightweight than the latter model.

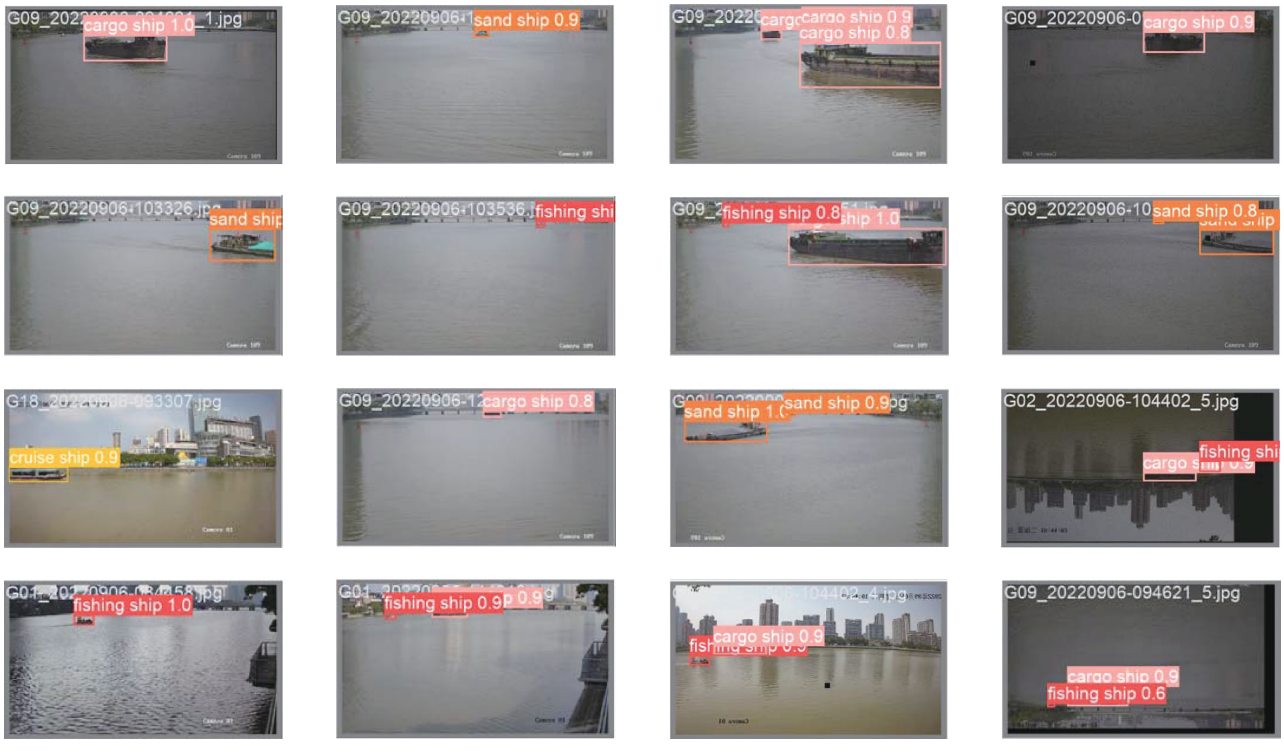
When evaluating the trained models on the dataset, all models perform differently for different classes, according to the AP values and the mAP values, as shown in Table 2.

**Table 2** AP and mAP of different models %

| Detection models | Fishing ship | | Cargo vessel | | Sand vessel | | Cruise vessel | | Crucial vessel | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $AP_{0.5}$ | $AP_{0.5:0.95}$ | $AP_{0.5}$ | $AP_{0.5:0.95}$ | $AP_{0.5}$ | $AP_{0.5:0.95}$ | $AP_{0.5}$ | $AP_{0.5:0.95}$ | $mAP_{0.5}$ | $mAP_{0.5:0.95}$ |
| Original YOLOv5 | 68.5 | 44.3 | 95.8 | 69.4 | 97.7 | 71.8 | 71.4 | 63.1 | 88.3 | 68.1 |
| Original YOLOv5 + AUG | 62.1 | 39.5 | 98.1 | 71.8 | 99.1 | 72.5 | 99.5 | 75.2 | 98.8 | 73.2 |
| Modified YOLOv5 + AUG | 69.6 | 46.8 | 98.4 | 73.5 | 99.2 | 73.2 | 99.5 | 78.6 | 99.0 | 75.1 |
| Modified YOLOv5 + AUG + anchors | 72.7 | 49.5 | 99.2 | 75.1 | 99.5 | 74.7 | 99.5 | 80.6 | 99.4 | 76.8 |

With the data augmentation, the addition of the C2f and SE, and the optimization of anchors, the precision of all categories has been promoted, as shown in Table 2. The final $mAP_{0.5}$ and $mAP_{0.5:0.95}$ for cargo vessels, sand vessels, and cruise vessels of the modified YOLOv5 + AUG + anchors model reach 99.4% and

76.8%, respectively. Compared with the original YOLOv5 model, the values are improved by 11.1% and 8.7%. The results of the modified YOLOv5 + AUG + anchors-based vessel detection for vessel-bridge collision warnings under complex navigable scenes are shown in Fig. 7. From Fig. 7, under complex scenes such as a far distance of targets, a small size of targets, multiple targets, and low visibility of targets, the precision for crucial vessels of the modified YOLOv5 + AUG + anchors remains at a high level, which can meet the requirements of high precision-based vessel detection.

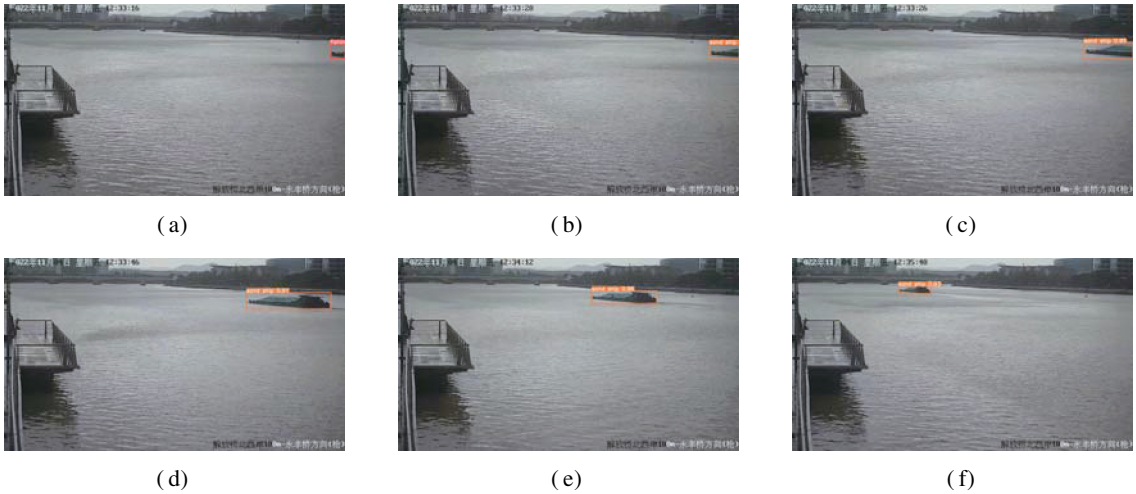



(a) (b) (c) (d)

**Fig. 7** Typical examples of detected vessels on the testing dataset. (a) Short distance; (b) Long distance; (c) Multiple objects; (d) Low visibility

## 2.3 Real-time detection

The detector for vessel-bridge collision warnings should not only be equipped with high precision but also have a fine inference speed. Therefore, the best weight of the trained modified YOLOv5 + AUG + anchors is extracted to detect vessels in a video of 25 frame/s continuously. Examples of real-time detection are shown in Fig. 8.



(a) (b) (c)

(d) (e) (f)

**Fig. 8** Typical examples of detected vessels on a video. (a) First frame; (b) 100th frame; (c) 250th frame; (d) 600th frame; (e) 1 000th frame; (f) 2 000th frame

During the test, the inference speed of the model is 102 frame/s, capable of the video detection task. As Fig. 8 shows, when a very small part of the sand vessel appears in the video, it is mistakenly detected as a fishing ship. Instantaneously, the model makes a correct recognition, and the AP is up to 96%. With the departure of the vessel, until the vessel is close to the bridge, the AP can still reach 93%. These results show that the model proposed in this paper is satisfactory for real-time detection performance and robustness.

## 3 Conclusions

1) With the addition of the C2f and SE modules, the modified YOLOv5 model has better generalization capability than the original YOLOv5 model and is more lightweight for model training.

2) The computer vision-based model proposed in this paper has high precision. The final $mAP_{0.5}$ and $mAP_{0.5:0.95}$ for crucial vessel targets reach 99.4% and 76.8%, respectively, 11.1% and 8.7% higher than that of the original YOLOv5 model.

3) In a real-time video scenario, the AP values are between 93% and 96%, and the inference speed is 102 frame/s, which agrees with the requirement for precision, reliability, and real-time vessel detection. The established model can provide information for subsequent anti-collision warnings of bridges under complex navigable scenes.

4) In future research, the geographical coordinates of detected vessels will be further determined by homography, and the dataset will continue to be explored to test the applicability of the model in case of inclement weather.

## References


[1] Yang Y D, Wang X F, Pan J J. Improved CNN and its application in ship identification[J]. *Computer Engineering and Design*, 2018, **39**(10): 3228 − 3233. DOI: 10.16208/j.issn1000-7024.2018.10.039. (in Chinese)

[2] Vagale A, Oucheikh R, Bye R T, et al. Path planning and collision avoidance for autonomous surface vehicles Ⅰ: A review[J]. *Journal of Marine Science and Technology*, 2021, **26**(4): 1292 − 1306. DOI: 10.1007/s00773-020 − 00787-6.

[3] Zhang B, Xu Z F, Zhang J, et al. A warning framework for avoiding vessel-bridge and vessel-vessel collisions based on generative adversarial and dual-task networks [J]. *Computer-Aided Civil and Infrastructure Engineering*, 2022, **37**(5): 629 − 649. DOI: 10.1111/mice.12757.

[4] Cui Z Y, Wang X Y, Liu N Y, et al. Ship detection in large-scale SAR images via spatial shuffle-group enhance attention[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2021, **59**(1): 379 − 391. DOI: 10.1109/TGRS.2020.2997200.

[5] Ren S Q, He K M, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, **39**(6): 1137 − 1149. DOI: 10.1109/TPAMI.2016.2577031.

[6] Liu W, Anguelov D, Erhan D, et al. SSD: Single shot-multibox detector[C]//*European Conference on Computer Vision*. Berlin, Germany, 2016: 21 − 37. DOI: 10.1007/978-3-319 − 46448-0_2.

[7] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*). Las Vegas, NV, USA, 2016: 779 − 788. DOI: 10.1109/CVPR.2016.91.

[8] Shao Z F, Wu W J, Wang Z Y, et al. SeaShips: A large-scale precisely annotated dataset for ship detection[J]. *IEEE Transactions on Multimedia*, 2018, **20**(10): 2593 − 2604. DOI: 10.1109/TMM.2018.2865686.

[9] Li H, Deng L B, Yang C, et al. Enhanced YOLOv3 tiny network for real-time ship detection from visual image [J]. *IEEE Access*, 2021, **9**: 16692 − 16706. DOI: 10.1109/ACCESS.2021.3053956.

[10] Lee W J, Roh M I, Lee H W, et al. Detection and tracking for the awareness of surroundings of a ship based on deep learning[J]. *Journal of Computational Design and Engineering*, 2021, **8**(5): 1407 − 1430. DOI: 10.1093/jcde/qwab053.

[11] Ni Y H, Mao J X, Wang H, et al. Toward high-precision crack detection in concrete bridges using deep learning[J]. *Journal of Performance of Constructed Facilities*, 2023, **37**(3): 04023017. DOI: 10.1061/jpcfev.cfeng − 4275.

[12] Zhou J C, Jiang P, Zou A R, et al. Ship target detection algorithm based on improved YOLOv5[J]. *Journal of Marine Science and Engineering*, 2021, **9**(8): 908 − 922. DOI: 10.3390/jmse9080908.

[13] Hu J, Shen L, Sun G. Squeeze-and-excitation networks [C]//2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT, USA, 2018: 7132 − 7141. DOI: 10.1109/CVPR.2018.00745.

[14] Xia Y, Chen L M, Wang J J, et al. Single shot multibox detector based vessel detection method and application for active anti-collision monitoring[J]. *Journal of Hunan University*: *Natural Science*, 2020, **47**(3): 97 − 105. DOI: 10.16339/j.cnki.hdxbzkb.2020.03.012. (in Chinese)

[15] Ni Y H, Lu H, Ji C, et al. Comparative analysis on bridge corrosion damage detection based on semantic segmentation [J]. *Journal of Southeast University* (*Natural Science Edition*), 2023, **53**(2): 201 − 209. DOI: 10.3969/j.issn.1001 − 0505.2023.02.003. (in Chinese)

# 用于船桥碰撞预警的复杂场景下视觉船舶检测

廖睿轩　吴　同　张一鸣　茅建校　王　浩

(东南大学混凝土及预应力混凝土结构教育部重点实验室,南京 211189)

**摘要**:为准确识别航道船舶,实现桥梁防撞预警,结合实拍图像和数据增强建立了针对桥梁航道船舶的图像数据集,包括远距离、多目标以及可视度低等复杂场景. 然后,以 YOLOv5 模型作为基本检测器并对其网络结构进行改进,主要改进包括将主干网络中的 C3 模块替换为 C2f 模块,在特征融合网络中嵌入 SE 注意力机制,采用 K-means + + 聚类算法对数据集的先验框进行优化. 最后,以 PyTorch 为深度学习框架对改进 YOLOv5 模型进行训练和验证. 结果表明,改进 YOLOv5 模型对重点检测船舶的平均精度达到 99. 4%,较原始 YOLOv5 模型提高了 11. 1%,检测速度达到 102 帧/s,可为复杂通航场景下船桥碰撞预警提供可靠、高效的支撑.

**关键词**:船舶检测;船桥相撞;YOLOv5;SE 注意力机制;数据增强

**中图分类号**:U447; U69