

Enhancing Container Damage Detection with improved YOLOv5 Model: Integrating Swin Transformer

Jiahao Chen, Chen Dong[✉], Yuxuan Wan

School of Computer Science and Engineering, Tianjin University of Technology,
Tianjin 300380, China
dongc@tjut.edu.cn

Abstract. Container damage is diverse and includes small-scale object damage (e.g., holes, dents, scratches). This paper proposes an improvement to the YOLOv5 model based on the Transformer self-attention mechanism for container damage detection. To effectively capture global relationships in damage images, two layers of Swin Transformer blocks are incorporated into the backbone network of YOLOv5. The PANet in YOLOv5 Neck has been optimized to BiFPN. Enhanced ability to fuse multi-scale features in damaged images while reducing computational complexity. Furthermore, use the Focaler-IoU Loss Function to improve the balance of features extracted from different samples in the dataset. Experimental results on the COCO and Tianjin Port official container damage datasets validate that the improved model achieves an mAP of 95.4%, demonstrating superior performance compared to commonly used object detection algorithms such as YOLOv5 and YOLOv8.

Keywords: Container damage detection, Improved YOLOv5, Transformer, BiFPN, Focaler-IoU Loss

1 Introduction

Containers are an essential part of modern logistics, serving as cargo carriers. Due to the continuous increase in global container transportation volume, containers are often

[✉] Corresponding author

subjected to impacts, compression, friction, and adverse weather conditions during long-distance transportation, loading, and stacking processes. Surface damage, such as holes, dents, cracks, and corrosion, can severely affect the safety, stability, and usability of containers [1-3]. Therefore, timely detection of container damage is crucial for ensuring the safe transportation of goods. However, there are currently issues with container damage detection technology, including poor detection accuracy and high missed detection rates for small-scale object damage like holes, dents. This paper discusses the challenges in container damage detection tasks and proposes an Improved YOLOv5 model using Transformer self-attention mechanism for container damage detection. Experimental results demonstrate that the proposed model can achieve promising results. The main contributions can be summarized as follows:

1. To effectively capture the global and long-range relationships of damaged container images, we introduce the SwinT_CSP module into the backbone network of YOLOv5, which combines the Swin Transformer window self-attention module with the CSP Bottleneck. This enhances the model's feature extraction capability and its ability to detect small-scale object such as holes, dents, and cracks in damaged images.
2. The incorporation of the BiFPN [4] in the neck enhances the multi-scale feature fusion capability in damaged images, reducing computational complexity and information loss. The use of the Focaler-IoU [5] function helps the model better balance feature extraction for different samples in the dataset.
3. Utilize K-Means clustering on the dataset to obtain 9 initial anchor boxes that are more suitable for the container damage dataset. During training, we introduce a multi-scale training method and label smoothing algorithm to enhance the model's generalizability.

2 Related Works

Container damage detection technologies typically include optical character recognition, laser scanning, and 3D imaging [6]. Since most container damages are manifested in appearance, using computer vision methods to detect damaged images can enhance detection accuracy and efficiency.

Nakazawa et al. [7] proposed a three-dimensional automatic detection device that sets different threshold values for segmenting holes and cracks using traditional image

segmentation algorithms, based on the differences in light reflection between damaged areas such as holes and cracks, and damages like paint, dirt, and dents. However, this method can only detect holes and cracks with weak light reflection, and has low accuracy. Son et al. [8] introduced the Capsize-Gaussian-Function to detect damage in containers. Based on this research, Son and Kim [9] utilized image preprocessing and Canny edge detection techniques to estimate the damage on the surface of containers and verified it on the Busan Port dataset. However, this algorithm is only suitable for locating damaged images with clear boundaries and cannot effectively detect other types of damage, such as blurring. To address the issue of unclear identification caused by the blurred surface of damaged containers, Kim et al. [10] proposed a container automatic recognition system based on the ART2 self-supervised learning algorithm. However, this method also has the drawback of only detecting single types of damage. Compared to traditional computer vision methods, deep learning models offer significant advantages in feature learning and recognition accuracy efficiency. Emil [6] investigated an automatic detection method for identifying damages to container corner castings using Faster R-CNN, MobileNet, and ResNet. Zixin et al [11] . proposed a model for detecting multiple types of container damage using transfer learning and MobileNetV2. The algorithm was capable of identifying various types of container damages such as dents, holes, rust, etc., but there is still significant room for improvement in detection accuracy. Zhiming et al. [12] presented a deep learning-based detection algorithm that can identify container numbers even under complex lighting conditions and background contamination. Furthermore, literature on road damage detection based on deep learning also provides some insights into damage detection for this paper. Wang et al. [13] used Faster R-CNN as the recognition framework to accomplish road damage detection. Guo and Zhang [14] proposed an improved road damage detection algorithm MN-YOLOv5 based on YOLOv5. They extracted the main features of the MobileNetV3 network to replace the backbone network of YOLOv5 and introduced a lightweight coordinate attention module. Compared to the original model, the mAP was improved by 2.5%, and the F1 score was increased by 2.6%. The above studies demonstrate the broad application prospects of damage detection algorithms based on deep learning.

In recent years, there has been a growing trend of applying Transformer [15] to computer vision tasks. Carion et al. proposed an end-to-end object detection model based on Transformer, named DETR [16]. It is the first object detection framework to successfully integrate Transformer as the central building blocks in the detection pipe-

line, and its performance is comparable to Faster R-CNN. The Visual Transformer model ViT [17], proposed by Dosovitskiy et al., fully adopts the standard structure of Transformer. Building upon the design principles of ViT, Liu et al. introduced Swin Transformer [18], which utilizes shifted windows for self-attention computation and integrates cross-window information of images. The ViT-FRCNN [19] model combines ViT with FRCNN for large-scale object detection tasks. There is also a mask-based Visual Transformer (MVT) [20], which improves the detection robustness of object detection models on complex background images.

In summary, computer vision-based container damage detection methods commonly suffer from low accuracy in detecting small-scale object damage and exhibit specificity and singularity. It notes that container damage can take various forms and that focusing on specific types may not meet inspection requirements for actual damage scenarios.

The improved model enhances the average precision of container damage detection and reduces the missed detection rate, which is of great significance for promoting the intelligent development of port transportation. The specific detection process of the model is illustrated in Figure 1.

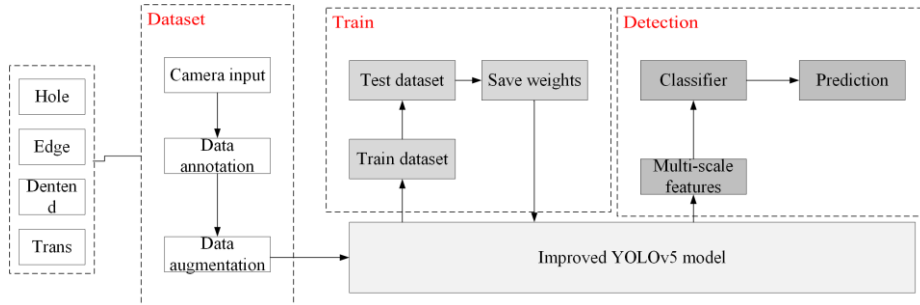


Fig. 1. Container damage detection process

3 Improved YOLOv5

The YOLO series is an advanced object detection framework widely used in computer vision for real-time object detection tasks due to its outstanding speed and accuracy. YOLOv5 is a widely used and classic framework. The latest YOLOv8 model is essentially an improvement of YOLOv5, with both using the same backbone network. YOLOv8 has improved training speed and detection accuracy compared to YOLOv5,

but the Frames Per Second (FPS) has decreased. Moreover, the Anchor-Base method in YOLOv5 has a higher recall rate in small-scale object detection tasks compared to the Anchor-Free method in YOLOv8. For container damage detection tasks such as detecting holes and cracks, the Anchor-Base method performs better. Therefore, this paper proposes improvements based on the YOLOv5 framework. Experimental results show that the performance of the improved model is not inferior to YOLOv8. The structure of the improved model is shown in Figure 2.

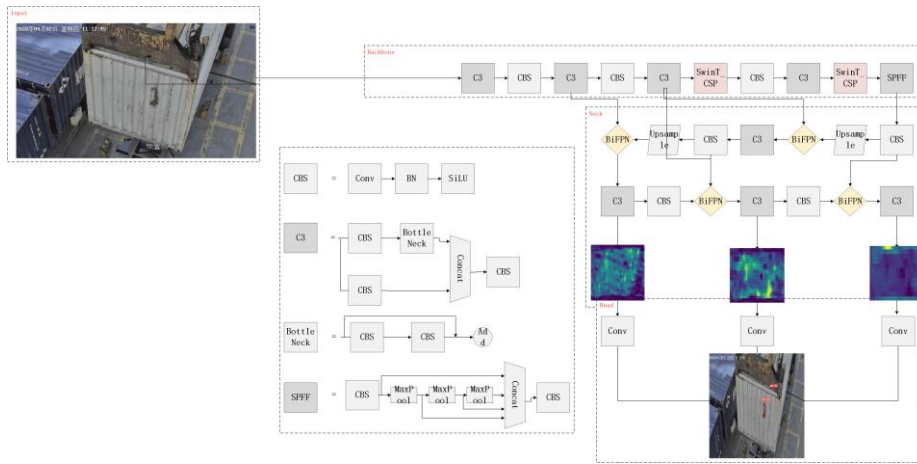


Fig. 2. Improved YOLOv5 network structure diagram

3.1 Transformer Integration in Backbone

In container damage detection, there are instances of small-scale object damages like holes, dents, which occupy a small proportion in the image and are densely distributed. Therefore, this paper adds the Transformer self-attention mechanism to the convolution module of the YOLOv5. This helps the model better learn the feature representation of damaged images and improves its ability to detect small targets.

In computer vision, one application of Transformer [错误!未找到引用源。] is to partition the image into fixed-size blocks (patches), obtain embedding representations for each block through linear transformations (patch embeddings), and then feed these embeddings into the Transformer for feature extraction and classification. However, relying solely on attention mechanisms and disregarding convolutions causes the model to lose the natural advantages of convolutions, including translation equivariance and locality. Therefore, when combining Transformer for container damage

(SW-MSA) module is utilized, which is essentially an offset version of the W-MSA, as illustrated in Layer $l + 1$ in Figure 4.

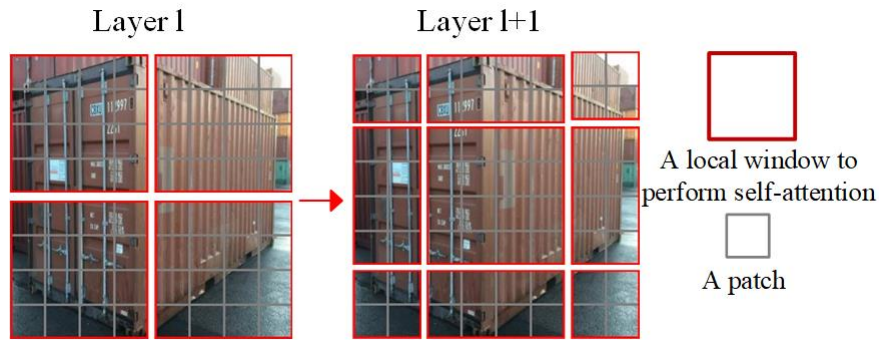


Fig. 4. Shift window method for calculating self-attention in Swin Transformer architecture

Figure 5 shows the combination of the Swin Transformer block with the CSP Bottleneck. This approach reduces computational costs, allowing the model to learn more feature representations from container damage images.

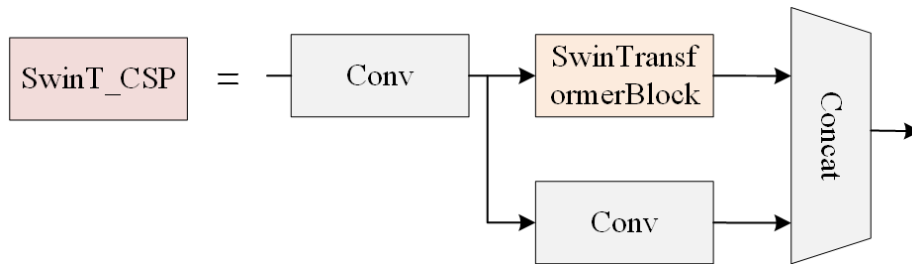


Fig. 5. The combination of Swin Transformer block and CSP Bottleneck (SwinT_CSP)

3.2 BiFPN feature fusion

The container image contains many small-scale object damage scenarios, such as holes, depressions, as well as larger affected areas like rust and deformation. Therefore, a mechanism is required to handle damage targets of different scales and sizes. In YOLOv5, the PANet feature fusion network has only one top-down and one bottom-up path, which limits its ability to fully utilize different scale feature information.

To address this issue, the BiFPN feature pyramid network has been introduced in the neck part of YOLO.

BiFPN integrates bidirectional cross-scale connections and fast normalized fusion. The feature fusion process is described by Equation 3.

$$\begin{aligned} P_l^t &= Conv\left(\frac{w_1 \cdot P_l^{in} + w_2 \cdot Resize(P_{l+1}^{in})}{w_1 + w_2 + \epsilon}\right) \\ P_l^{out} &= Conv\left(\frac{w'_1 \cdot P_l^{in} + w'_2 \cdot p_l^t + w'_3 \cdot Resize(P_{l-1}^{out})}{w'_1 + w'_2 + w'_3 + \epsilon}\right) \end{aligned} \quad (3)$$

Where P_l^t is the intermediate feature of the l -th layer of the top-down path, P_l^{out} is the output feature of the l -th layer of the bottom-up path, and $Resize$ is a downsampling or upsampling operation.

The incorporation of the BiFPN feature pyramid network into the YOLOv5 model's neck structure improves the model's ability to fuse multi-scale features. This results in better detection performance for low-resolution damaged objects by capturing more information about the affected areas in images.

3.3 Focaler-IoU loss function

During training, YOLOv5 mainly includes three types of losses: bounding box loss ($loss_{rect}$), confidence loss ($loss_{obj}$), and classification loss ($loss_{cls}$). Therefore, the loss function of the YOLOv5 network is defined as Equation 4:

$$Loss = a \times loss_{obj} + b \times loss_{rect} + c \times loss_{cls} \quad (4)$$

YOLOv5 utilizes the CIoU loss to calculate the bounding box loss. In the task of container damage detection, from the analysis of the scale of detection objects, rust, deformation, etc., can be considered as simple samples, while extremely small-scale object such as hole damages can be regarded as difficult samples due to the challenge of precise localization. The CIoU loss function considers the overlap area of bounding box regression, center point distance, and aspect ratio. However it does not address the issue of balancing simple and difficult samples, limiting its effectiveness in container damage detection tasks. Therefore, the Focaler-IoU loss function is chosen to replace CIoU, allowing the model to focus on samples of different difficulty levels.

The Focaler-IoU loss function reconstructs the IoU loss using a linear interval mapping method, allowing it to focus on different regression samples in various detection tasks. This is represented by Equation 5:

$$IoU^{focaler} = \begin{cases} 0, & IoU < d \\ \frac{IoU - d}{u - d}, & d \ll IoU \ll u \\ 1, & IoU > u \end{cases} \quad (5)$$

Where $IoU^{focaler}$ is the reconstructed Focaler-IoU, $[d, u] \in [0, 1]$. By adjusting the values of d and u , $IoU^{focaler}$ can focus on different regression samples. Its loss function is defined by Equation 6:

$$L_{Focaler-IoU} = 1 - IoU^{focaler} \quad (6)$$

4 Experiments

4.1 Dataset

A dataset of container images was collected based on the official dataset provided by Tianjin Port and photos of containers captured. To expand the data set, data enhancement techniques such as horizontal mirroring and the addition of Gaussian noise were applied. The dataset was divided into training, validation, and testing sets in an 8:1:1 ratio to ensure sufficient samples for training, effective validation, and objective accuracy in testing evaluations. Figure 6 illustrates examples of data augmentation.



Fig. 6. Container image data enhancement legend

The dataset consists of four main categories of container damage: Hole, Edge, Dented, and Trans, as shown in Figure 7.



Fig. 7. 4 illustrations of different types of damaged containers

4.2 Experimental environment

The experimental environment and parameter settings adopted in this study are presented in Table 1, ensuring the reproducibility and reliability of the experimental results. Key information includes the operating system, CPU, GPU, programming language, memory, and number of training epochs, among others.

Table 1. Experimental environment configuration and parameter settings

Configuration items	Configuration parameters
Operating system	Ubuntu 20.04.5 LTS
CPU	intel(R)Core(TM)i5-13600K
GPU	NVIDIA GeForce RTX3060
Compiled language	Python 3.8.12
Running memory	32GB
Epochs	300

4.3 Training strategy

1. KMeans clustering

In container damage detection tasks, the model not only needs to detect the types of damage but also to learn the position and size of the damage. The prior anchor box mechanism divides the space where objects of different scales and aspect ratios are located into multiple subspaces. The initial anchor boxes for the YOLOv5 model only applicable to the COCO dataset. KMeans clustering is applied to the container damage dataset to recluster and obtain new anchor boxes suitable for the dataset. A comparison of the two sets of anchor boxes is illustrated in Figure 8.

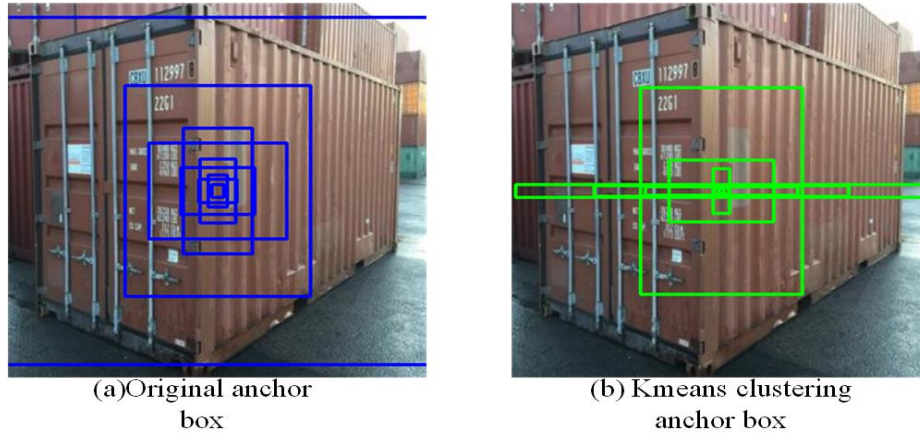


Fig. 8. Visual comparison of two anchor box sizes

2. Multi-scale training

In container damage detection tasks, such as detecting small-scale object like holes, the backbone network generates feature maps several times smaller than the original image during the feature extraction stage. This often results in difficulty for the detection network to capture detailed feature descriptions. Therefore, During training, every few iterations, a random scale was chosen for training. By inputting larger and varied sizes of images during training, the detection model's robustness to different sizes of container damage targets was somewhat improved.

3. Label smoothing

In container damage detection, the label smoothing regularization strategy is employed to prevent overfitting due to the presence of noise in the training data and the inability to ensure that all samples in the dataset are correctly labeled. This strategy perturbs the target variables to impose constraints on the model, preventing overfitting and improving the model's generalization ability and robustness.

4.4 Experimental results

To evaluate the effectiveness of the improved YOLOv5 model objectively, employed several metrics for model evaluation, including precision, recall, and accuracy (mAP@0.5). mAP is one of the most important evaluation metrics in the field of ob-

ject detection. It is the average of the area under the PR curve calculated for all categories at different Intersection over Union (IoU) thresholds. mAP@0.5 refers to the mAP value when the IoU threshold is set to 0.5.

Trained the improved YOLOv5 model on the Tianjin Port container dataset and obtained the following training results:



Fig. 9. Training results of the Tianjin Port container damage dataset

Figure 9 shows that the YOLOv5 model reaches a converged state during training and performs well on the Tianjin Port container damage dataset, achieving prediction accuracy and mAP values both exceeding 90%.

To further evaluate the improved YOLOv5 model, it was compared with commonly used object detection algorithms, including YOLOv3, YOLOv4, YOLOv5 and YOLOv8. The comparison metrics include precision, recall, and mAP values. The specific results are presented in Table 2.

Table 2. Comparison of training results of Tianjin Port container damage data set

Algorithm	Backbone	Precision	Recall	mAP ₅₀
Faster-RCNN	ResNet-101	78.9%	79.4%	76.1%
SSD	VGG-16	69.6%	71.3%	70.8%
YOLOv3	Darknet53	70.2%	78.4%	72.3%
YOLOv4	CSPDarknet53	83.7%	81.1%	82.4%
YOLOv5	CSPDarknet53	85.3%	82.5%	88.6%
YOLOv5	CSPDarknet53+CBAM	94.3%	86.2%	93.7%
YOLOv8	CSPDarknet53	96.1%	93.6%	94.8%
Improved YOLOv5	CSPDarknet53+ SwinT_CSP	95.8%	96.3%	95.4%

Based on the comparison of experimental results, it is evident that the improved YOLOv5 model achieves the highest mAP value, reaching 95.4%. Indicating the model's excellent capability in detecting damaged targets.

Figure 10 illustrates a visual comparison of the improved YOLOv5 model with other advanced detection models. It can be observed that the improved YOLOv5 model and YOLOv8 exhibit the highest detection accuracy.

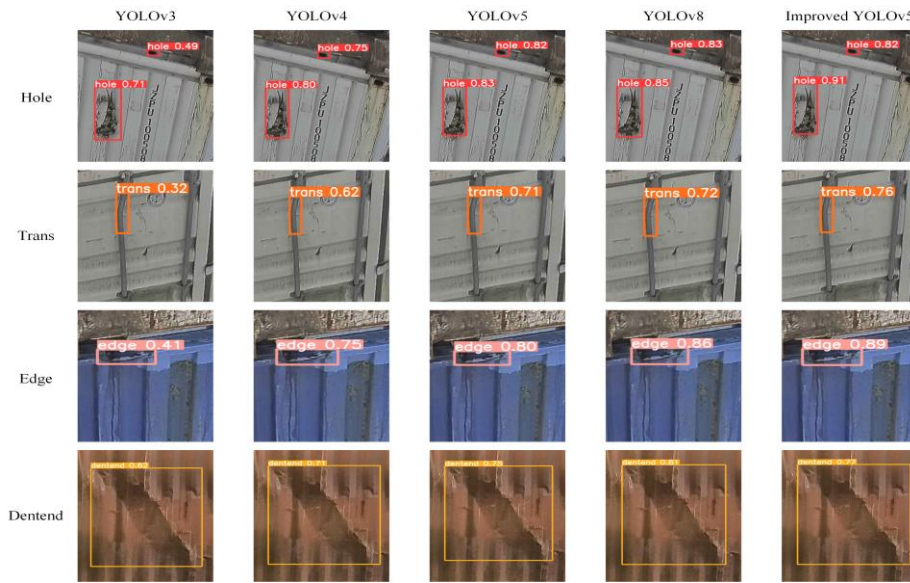


Fig. 10. Comparison between the improved YOLOv5 model and other advanced models

In the improved YOLOv5 model, it mainly combines three modules: SwinT_CSP, BiFPN, and Focaler-IoU. In order to further analyze the impact of each module on the container damage detection results, an ablation experiment was designed. Using the YOLOv5 model as the baseline, each module was added separately for experiments. The results are shown in Table 3.

Table 3. Single Variable Ablation Experiment Results Comparison

Algorithm	Precision	Recall	mAP50
YOLOv5(base)	85.3%	82.5%	88.6%
YOLOv5 + SwinT_CSP	89.5%	87.2%	90.2%
YOLOv5 + BiFPN	87.2%	84.8%	89.5%
YOLOv5 + Focaler-IoU	86.8%	83.9%	89.0%
Improved YOLOv5	95.8%	96.3%	95.4%

To evaluate the generalization ability of the improved YOLOv5 model, testing was performed on the MS COCO official dataset. This dataset includes more than 330,000 images and annotations for 80 categories, making it one of the most authoritative benchmark datasets in the field of object detection. The specific results are presented in Table 4, which demonstrates the model's excellent performance.

Table 4. Comparison of training results on MS COCO dataset

Algorithm	Backbone	mAP ₅₀
Faster-RCNN	ResNet-101	55.7%
SSD	VGG-16	50.4%
YOLOv3	Darknet53	57.9%
YOLOv4	CSPDarknet53	61.7%
YOLOv5	CSPDarknet53	63.6%
Improved YOLOv5	CSPDarknet53+ SwinT_CSP	64.1%

5 Conclusion

This paper proposes an improved YOLOv5 container damage detection algorithm based on the Transformer mechanism to address the issues of the singularity of current container damage detection methods and the difficulty in detecting small damaged areas in containers. By integrating the window self-attention calculation of Swin Transformer with convolution operations into the Backbone, the model's feature extraction capability is enhanced, thereby improving the detection ability of complex small-scale object damage. In the Neck, the BiFPN is used instead of the PANet structure to improve the multi-scale fusion capability of the detection head. The Focaler-IoU loss function is employed to focus on damaged samples of different scales, enhancing the model's robustness. Ablation experiments were designed to demonstrate the positive impact of each module on container damage detection tasks. To validate the effectiveness and generalization ability of the model, experiments are conducted on both the Tianjin Port official damaged container dataset and MS COCO. The improved model achieves an mAP of 95.4% on the container dataset, meeting the requirements of port container damage detection tasks.

The study still faces several issues. The dataset mainly comprises damaged container images from Tianjin Port, which may result in sample selection bias and limit the universality of the research. Although the accuracy of container damage detection has improved, assessing the severity of multiple damaged areas in containers still

requires enhancement. In the next stage, the main tasks include incorporating more damaged container images from other ports to enhance the diversity of the dataset. Efforts will also be made to improve the model's object detection capability by employing multi-modal data fusion techniques to enhance its performance in detecting the severity of damages.

Acknowledgments. This research is support[2]ed by The second batch of Tianjin's manufacturing high-quality development special construction for intelligent and digital application scenarios in 2023 and the project "Research on Energy-saving Technology for District Heating Based on Big Data and Deep Learning(Z20220192)".

References

1. G. Bee, L. Hontz, Detection and prevention of post-processing container handling damage, *Journal of Food Protection* 43(6) (1980) 458-461.
2. S.-H. Cha, C.-K. Noh, A Case Study of Automation Management System of Damaged Container in the Port Gate, *Journal of Korean navigation and port research* 41 (2017) 119-126.
3. O.J. Ho, H.S. Woo, C.G. Jong, K.M. Ho, A.D. Sung, Development of the container damage inspection system, *Journal of the Korean Society for Precision Engineering* 22(1) (2005) 82-88.
4. M. Tan, R. Pang, Q.V. Le, Efficientdet: Scalable and efficient object detection, *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10781-10790.
5. H. Zhang, S. Zhang, Focaler-IoU: More Focused Intersection over Union Loss, *arXiv preprint arXiv:2401.10525* (2024).
6. E. Kattainen, Object detection for container corner detection, (2019).
7. K. Nakazawa, I. Iwasaki, I. Yamashita, Development of damage detection system for container, *Proceedings of IECON'95-21st Annual Conference on IEEE Industrial Electronics*, IEEE, 1995, pp. 1160-1163.
8. T.N.H. Son, Y.-S. Ha, H.-S. Kim, AN APPLICATION OF DIGITAL IMAGE PROCESSING TECHNIQUES IN DETECTING DAMAGE OR DEFORMATION SHAPE ON EXTERNAL SURFACE OF CONTAINER.
9. S.T.N. Hoang, Estimating directly damage on external surface of container from parameters of capsized-Gaussian-function, *Proceedings of the Korean Institute of*

- Navigation and Port Research Conference, Korean Institute of Navigation and Port Research, 2005, pp. 297-302.
10. K.-B. Kim, S. Kim, Y.-J. Kim, Container image recognition using ART2-Based self-organizing supervised learning algorithm, *Advances in Natural Computation: Second International Conference, ICNC 2006, Xi'an, China, September 24-28, 2006. Proceedings, Part I 2*, Springer, 2006, pp. 385-394.
 11. Z. Wang, J. Gao, Q. Zeng, Y. Sun, Multitype damage detection of container using CNN based on transfer learning, *Mathematical Problems in Engineering 2021 (2021)* 1-12.
 12. W. Zhiming, W. Wuxi, X. Yuxiang, Automatic container code recognition via faster-RCNN, *2019 5th International Conference on Control, Automation and Robotics (ICCAR), IEEE, 2019*, pp. 870-874.
 13. W. Wang, B. Wu, S. Yang, Z. Wang, Road damage detection and classification with faster R-CNN, *2018 IEEE international conference on big data (Big data), IEEE, 2018*, pp. 5220-5223.
 14. G. Guo, Z. Zhang, Road damage detection algorithm for improved YOLOv5, *Scientific reports 12(1) (2022)* 15523.
 15. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, *Advances in neural information processing systems 30 (2017)*.
 16. N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, *European conference on computer vision, Springer, 2020*, pp. 213-229.
 17. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, An image is worth 16x16 words: Transformers for image recognition at scale, *arXiv preprint arXiv:2010.11929 (2020)*.
 18. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, *Proceedings of the IEEE/CVF international conference on computer vision, 2021*, pp. 10012-10022.
 19. J. Beal, E. Kim, E. Tzeng, D.H. Park, A. Zhai, D. Kislyuk, Toward transformer-based object detection, *arXiv preprint arXiv:2012.09958 (2020)*.
 20. H. Li, M. Sui, F. Zhao, Z. Zha, F. Wu, MVT: mask vision transformer for facial expression recognition in the wild, *arXiv preprint arXiv:2106.04520 (2021)*.