Pattern Analysis and Applications (2024) 27:116
https://doi.org/10.1007/s10044-024-01338-0

ORIGINAL ARTICLE

YOLOv7-GCM: a detection algorithm for creek waste based on improved YOLOv7 model

Jianhua Qin[1,2] · Honglan Zhou[1,2] · Huaian Yi[1,2] · Luyao Ma[1,2] · Jianhan Nie[1,2] · Tingting Huang[1,2]

Received: 8 July 2024 / Accepted: 6 September 2024

Abstract
To enhance the cleanliness of creek environments, quadruped robots can be utilized to detect for creek waste. The continuous changes in the water environment significantly reduce the accuracy of image detection when using quadruped robots for image acquisition. In order to improve the accuracy of quadruped robots in waste detection, this article proposed a detection model called YOLOv7-GCM model for creek waste. The model integrated a global attention mechanism (GAM) into the YOLOv7 model, which achieved accurate waste detection in ever-changing backgrounds and underwater conditions. A content-aware reassembly of features (CARAFE) replaced a up-sampling of the YOLOv7 model to achieve more accurate and efficient feature reconstruction. A minimum point distance intersection over union (MPDIOU) loss function replaced the CIOU loss function of the YOLOv7 model to more accurately measure the similarity between target boxes and predictive boxes. After the aforementioned improvements, the YOLOv7-GCM model was obtained. A quadruped robot to patrol the creek and collect images of creek waste. Finally, the YOLOv7-GCM model was trained on the creek waste dataset. The outcomes of the experiment show that the precision rate of the YOLOv7-GCM model has increased by 4.2% and the mean average precision (mAP@0.5) has accumulated by 2.1%. The YOLOv7-GCM model provides a new method for identifying creek waste, which may help promote efficient waste management.

Keywords YOLOv7 · Waste detection · Quadruped robot · Attention mechanism

1 Introduction

With the advancement of global agriculture, the complexity and quantity of creek waste have correspondingly increased, which exacerbates the impact of agriculture on the environment. Despite the increasing amount of waste, effective methods for detecting waste are still insufficient. At present, the mainstream methods of waste disposal are dumping and incineration, which seriously threaten sustainable development for humanity. Actually, wrong waste disposal methods not only cause irreversible disasters to the ecological environment, but also pose huge risks to organisms. Waste detection not only solves the problem of waste pollution, but also helps promote economic development and environmental protection. Therefore, imperative to implement efficient creek waste detection [1].

As machine technology rapidly progresses, deep learning methods have attracted widespread attention [2]. Representatives of one-stage algorithms are SSD [3], YOLOv1 [4], YOLOv2 [5], YOLOv3 [6], YOLOv5 [7] and so on. Representatives of two-stage algorithms are CNN, R-CNN [8] and Faster R-CNN [9]. In order to more effectively detect and classify various types of waste, scholars have conducted extensive research and proposed various deep learning models. These models not only utilize the powerful capabilities of deep learning in feature extraction and pattern recognition, but also incorporate the specific requirements of waste detection tasks for design and optimization. Cheng et al. [10] optimized the CenterNet network pass through feature fusion to better abstract subtle features of waste. YOLO network and the original CenterNet network were used for waste detection. The VGG network and DenseNet network were used to optimize the backbone of the YOLO model. A waste

✉ Jianhua Qin
qinjh2@sina.com

[1] Education Department of Guangxi Zhuang Autonomous Region, Key Laboratory of Advanced Manufacturing and Automation Technology (Guilin University of Technology), Guilin 541006, China

[2] College of Mechanical and Control Engineering, Guilin University of Technology, Guilin 541006, China

Published online: 17 September 2024

detection model was designed. And a recyclable waste dataset was constructed to ascertain the algorithmic efficiency and effectiveness. Tian et al. [11] came up with a trash detection model that can recognize objects quickly and accurately through enhancing the YOLOv4 network. Especially, this method selects YOLOv4 model as the fundamental neural model frame for object detection. The enhanced YOLOv4 model has exceptional detection velocity and precision rate, according to the experimental findings. Hou et al. [12] came up with a complex detection model under water environment objectives based on improved YOLOv5s model. This algorithm adds a self-attention layer to increase the model's capability and enhance precision rate. Accordingly, although scholars have made elemental achievements in the field of waste detection, they still face some challenges. The existing waste image datasets mainly focus on indoor garbage detection, but lack datasets that can be used for outdoor creek waste detection and recognition. The continuous changes in the water flow environment may cause blurring and deformation of the images when taking images with quadruped robots. Most indoor waste datasets cannot provide useful information for quadruped robots.

So as to better address these problems, the article not only construct a new image dataset for creek waste, but also propose a detection algorithm for creek waste. The following is a summary of this article's primary contributions:

(1) We construct a high-quality dataset for creek waste, which provides 24 type of images.
(2) This article integrates the GAM and CARAFE modules into the YOLOv7 network, which is specifically optimized for the detection of waste.
(3) At last, The CIOU loss function of the YOLOv7 model is substituted by a MPDIOU loss function. The MPDIOU loss function value of the network can be reduced while improving the precision rate and recall rate of YOLOv7 network.

2 Dataset and methods

2.1 Dataset construction

We have carefully constructed a brand new garbage image dataset in this article, which aims to provide strong support for environmental protection and waste classification research. In order to obtain real images of various types of waste, we specifically used quadruped robots to walk and take photos in the natural environment of streams. These images cover 24 different types of waste, including plastic bottles, paper, metal cans, etc., all presented in high-resolution RGB image format. The shooting scene is shown in Fig. 1a, which shows a quadruped robot walking by a small stream. Its high-definition camera captures various scattered waste in the natural environment. Through this method, we ensure the diversity and authenticity of the dataset, providing rich materials for subsequent waste classification and recognition research.

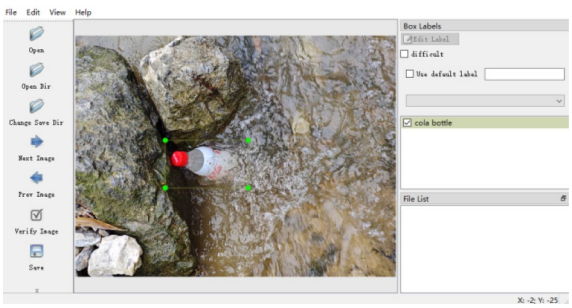
The entire dataset contains 730 images, which are not only representative but also cover various complex environments and lighting conditions. To improve the reliability of the dataset, we conducted detailed annotation work on each image before conducting the experiment. The annotation process is crucial for accurate annotation of images. In order to achieve precise labeling of each waste category, we used LabelImg [13], an open-source image labeling tool. We can easily create bounding boxes for targets and add corresponding labels to them through LabelImg. This process generates label files in txt format.

As shown in Fig. 1b, we demonstrate an example of using the LabelImg tool for image annotation. In the figure, we can see a labeled plastic bottle target with its bounding box tightly fitting the contour of the target, and the label file also records the category information of the target. We ensure that the annotations for each target in the dataset are accurate and reliable through this method. Finally, to verify and evaluate the performance of the waste detection algorithm, we divided the dataset into training, testing, and validation sets in an 8:1:1 ratio. Through this partitioning method, we can ensure the generalization ability of the algorithm on unknown data, providing strong support for subsequent waste detection research.

The images used were categorized into 24 categories: nut jar, white brush, transparent plastic bag, laundry detergent bottle, jewelry box, banana skin, hanger, coffee cup, pink and white plastic bag, tissue bag, white board, mobile phone shell, plastic seal, ziplock bag, cola bottle, green toy, white paper shell, white lid, white shoes, tape, brown carton, bread bag, clear plastic bag, blue paper. The sample images in the dataset is shown in Fig. 1c.

Data augmentation has become an indispensable link to build a more efficient dataset. Data augmentation not only aims to expand the size of the dataset, but also enables machine learning models to learn more generalized and robust feature representations. Specifically, through operations such as mirroring, blurring, rotation, cropping and scaling, the model is able to learn the invariant features of images in different poses and perspectives. These operations not only enrich the diversity of the dataset, but also help the model make more accurate and stable predictions when facing new samples. We obtained 9033 images to enhance the persuasiveness of the experiment through data augmentation techniques. Figure 2 shows those waste images after data augmentation.

(a) Filming scene



(b)The schematic diagram of dataset annotation.

(c) Samples in the image

Fig. 1 Image capture and annotation

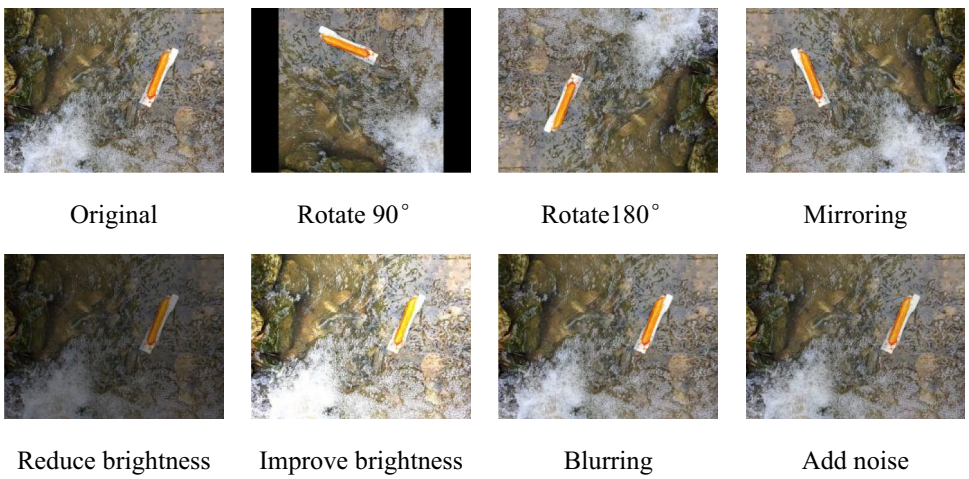


Fig. 2 Waste images after data augmentation

2.2 The YOLOv7-GCM model

Figure 3 depicts the YOLOv7 [14] model's network architecture. The YOLOv7-GCM network has made three important improvements on the basis of the YOLOv7 network. Firstly, to address the issue of information loss during feature extraction and up-sampling, we have introduced the CARAFE module, replacing the original up-sampling module. The CARAFE recombines features through content awareness, which can significantly reduce the loss of feature information in input images, especially in multi-scale feature fusion. The CARAFE ensures the effective transmission of important information. This improvement enables the network to more accurately capture the features of the target during the detection process.

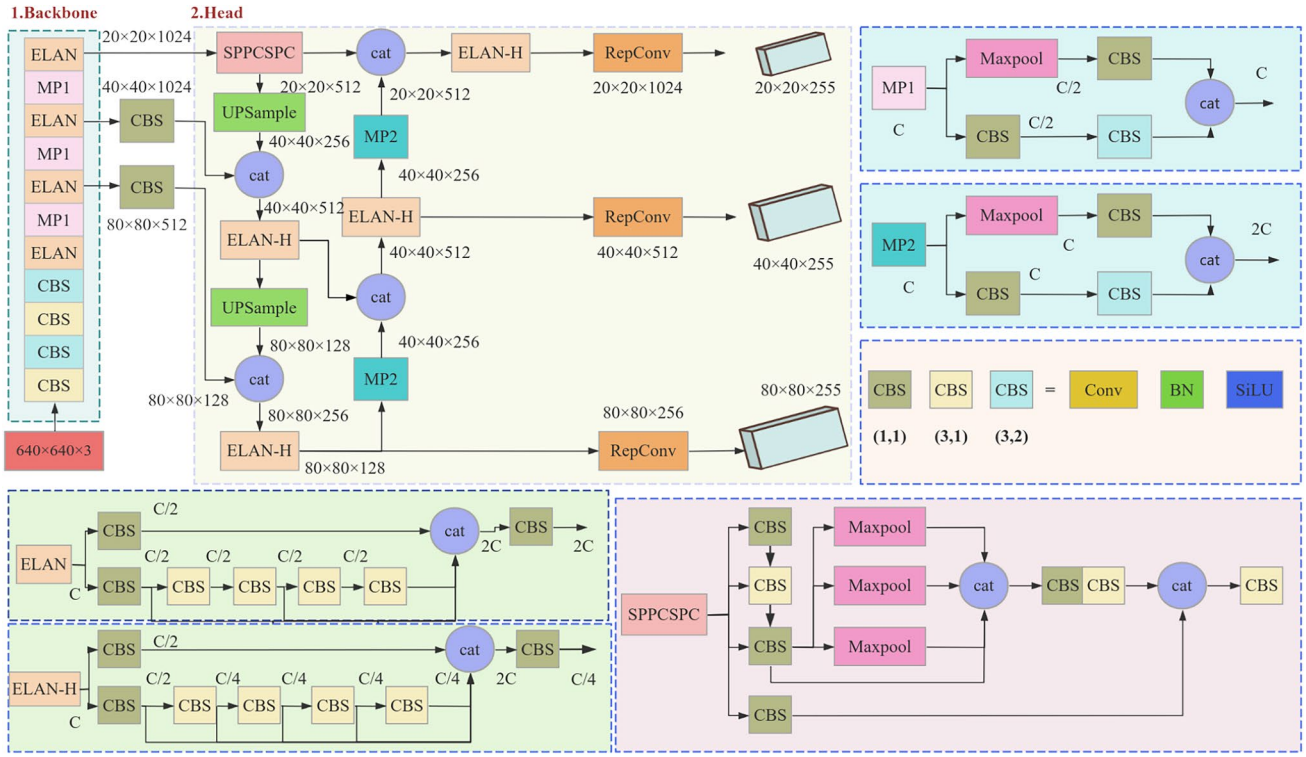


Fig. 3 The architecture of YOLOv7 network

We have decided to replace the original MP2 module with the GAM in the process of in-depth research and optimization of the YOLOv7 network. The purpose of this decision is to give the model higher attention to small target areas during prediction. The GAM captures features through a global perspective, allowing the YOLOv7-GCM model to more accurately adapt to the diversity and distribution characteristics of small target objects in images. This innovation not only improves the detection accuracy of the model for small targets, but also further enhances the object detection ability of YOLOv7-GCM in complex scenes.

Finally, to address the issue of unstable convergence of the YOLOv7 model in scenarios such as waste detection, we introduced the MPDIOU loss function, replacing the original CIOU loss function. The distance between the anticipated box corner and the actual box corner is taken into account by the MPDIOU loss function. This improvement enables the model to more accurately evaluate the differences between predicted results and actual labels during the training process. Meanwhile, the MPDIOU loss function's more precise consideration of the corner position of the predicted box.

We have improved the YOLOv7 network and successfully obtained the YOLOv7-GCM network through these three points. This network further improves its detection performance on the basis of inheriting the advantages of YOLOv7, especially its ability to detect waste. This makes YOLOv7-GCM more widely applicable.

2.2.1 The YOLOv7 model

The YOLOv7 model stands out in the field of single-stage detection algorithms due to its excellent performance and efficient detection speed, becoming one of the typical representatives. Its precise target recognition ability and fast response speed provide strong technical support for various application scenarios. The network architecture of YOLOv7 model is shown in Fig. 3. According to the structure diagram of YOLOv7 model, the YOLOv7 model is comprised of an input module, a backbone structure and a head structure. The YOLOv7 network preprocesses images by resizing them to $640 \times 640 \times 3$, which inputs the adjusted images into the backbone structure. Backbone structure is consist of some cross-stage-partial-connections (CBS) modules, ELAN [15] modules and MP1 [16] modules. The CBS module consists of convolution, batch normalization and SiLU activation function. The MP1 module mainly includes Maxpool [17] module and CBS module. And the ELAN module is a structure consist of multiple CBS modules, which enable the extraction of features of various sizes from a same feature map. The ELAN-H structure is similar to the composition

structure of ELAN structure, but the number of CATs between the two varies [18].

2.2.2 Global attention mechanism (GAM)

The GAM plays a crucial role in deep learning models especially in visual tasks, which aims to enhance the model's recognition ability for key regions or feature channels in images. As shown in Fig. 4, the structure of the GAM clearly demonstrates how these two sub-modules work together. Firstly, the channel attention sub-module utilizes a three-dimensional arrangement to protect the three-dimensional information in the input data. This arrangement ensures the information integrity of the data in the channel dimension, providing rich information for subsequent processing. The channel attention sub-module adopts a multi-layer perceptron (MLP) to further enhance the spatial dependence of cross dimensional channels. Because of its multi-layered structure, MLP is able to identify the intricate connections between several channels and assign a priority to each one. By using this method, the model can increase its processing accuracy and efficiency by concentrating more on the channels that are important to the outcome. However, the spatial attention sub-module concentrates on the image's spatial content. Two convolutional layers are used to combine spatial data. The model can use convolution operations to extract characteristics from local portions of the image and use that information to determine which regions are more relevant for the task at hand. With the enhancement of spatial attention, the performance of the model has been significantly improved. It can focus more accurately on the key information in the image, thereby improving the understanding and processing ability of images.

Where F_1 represents an input feature map, M_C represents the channel attention, M_S represents the spatial attention, F_2 denotes a feature map during processing, F_3 denotes a output feature map. This study introduced the GAM into the head structure of the YOLOv7 model. The GAM reduces the dispersion of information and enhances the interactive features of the global dimension to boost the model's performance. The GAM corrects the original feature map through two independent attention sub-modules based on the input feature map F_1. Firstly, the channel attention sub-module corrects the original feature map to obtain the feature map during processing F_2. Then, the spatial attention sub-module corrects the feature map during processing F_2 to obtain the feature map F_3. The feature map during processing F_2 and the final output F_3 are shown in Eqs. (1) and (2).

$$F_2 = M_C \otimes F_1 \tag{1}$$

$$F_3 = M_S \otimes F_2 \tag{2}$$

where $\otimes$ denotes the calculation of product for elemental methods.

2.2.3 Content-aware reassembly of features (CARAFE)

Feature up-sampling is a common operation in neural network and machine learning. It is usually used to enlarge the size of feature maps which indirectly increases the spatial resolution of the original image after model processing. Due to the chaotic background in water, it is often difficult to extract clear semantic information from waste images. The CARAFE up-sampling operator is based on input image and has a wider perceptual field, which enables more efficient utilization and

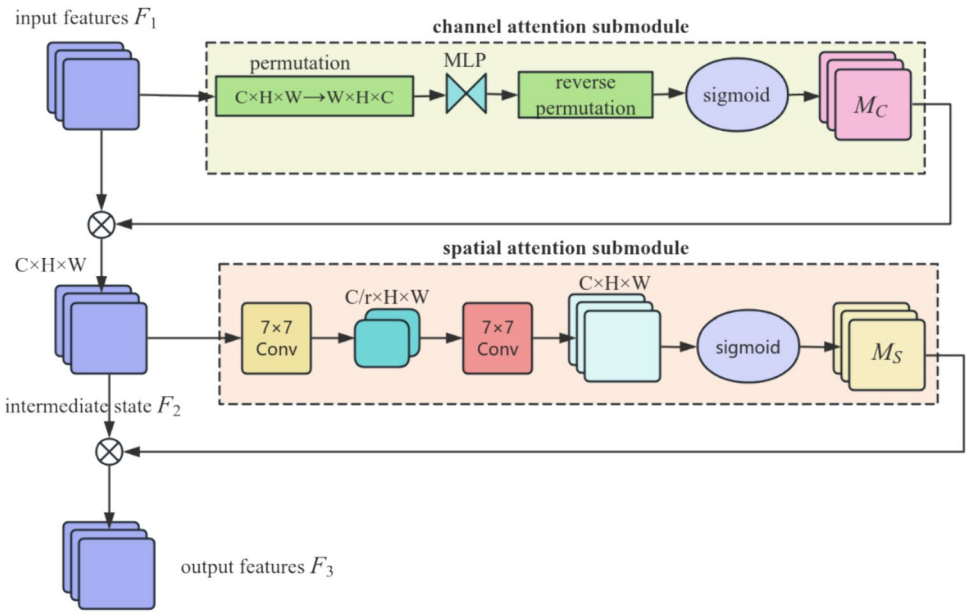


Fig. 4 The structure diagram of the global attention mechanism

integration of surrounding information. These information are then combined with the deep information of the feature map. The CARAFE replaced the up-sampling [19] in the head structure of YOLOv7 model to enhance the capability of extracting waste feature information. The structure of the CARAFE is shown in Fig. 5.

In the sampling process, given a feature map X of size $C \times H \times W$ and an up-sampling rate σ (assuming σ is an integer), the CARAFE will generate a new feature map X^* of size $C \times \sigma H \times \sigma W$. For any target location $L^* = (i^*, j^*)$ of the new map X^*, there is a corresponding source location $L = (i, j)$ on the map X, where $i = \left\lfloor i^*/\sigma \right\rfloor, j = \left\lfloor j^*/\sigma \right\rfloor$. Where $N(X_L, K)$ denotes the new map X generated for a region of size $K \times K$, centered on the location L on the feature map [20]. The kernel prediction module g predicts the location kernel W_{L^*} for each location L^* based on the domain of X_L. The formula is shown in Eq. (3):

$$W_{L^*} = g\left(N\left(X_L,\ X_{encoder}\right)\right) \tag{3}$$

After that the features are reorganized by the content-aware reorganization module ψ, which recombines the domains of X_L with the kernel W_{L^*}. The formula is shown in Eq. (4):

$$X^*_{L^*} = \psi(N(X_L \,.\, K_u), W_{L^*}) \tag{4}$$

2.2.4 Minimum point distance intersection over union (MPDIOU)

The YOLOv7 network adopts the CIOU loss function [21]. This function can more accurately measure the overlap and shape difference between the predicted box and the target box, thereby significantly improving the accuracy and robustness of detection. This function evolved from the IOU loss function [22], which intuitively reflects the similarity between the predicted bounding box and the actual target bounding box by calculating the ratio of intersection and union. However, in order to more accurately guide the regression of the target framework, the CIOU loss function is introduced in the YOLOv7 model. By comprehensively measuring these error amplitudes, the regression process of the target framework is more stable and reliable. The mathematical expression for the CIOU function is shown in Eq. (5), (6),(7):

$$Loss_{CIOU} = 1 - Loss_{IOU} + \frac{\rho^2\left(b_1, b_2\right)}{Z^2} + \alpha\nu \tag{5}$$

$$\nu = \frac{4}{\pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h}\right)^2 \tag{6}$$

$$\alpha = \frac{\nu}{\left(1 - \mathrm{Loss}_{\mathrm{IOU}}\right) + \nu} \tag{7}$$

where b_1 denotes the center point of the predicted box, b_2 denotes the center point of the real box, ρ is to calculate the euclidean distance between the two center points, Z is the diagonal distance of the smallest closure, w^{gt} and h^{gt} denote the width and height of the real box, w and h denote the width and height of the predicted box, α is a parameter to make the trade-off, and ν is the parameter used to measure the consistency of aspect ratios. The CIOU loss function

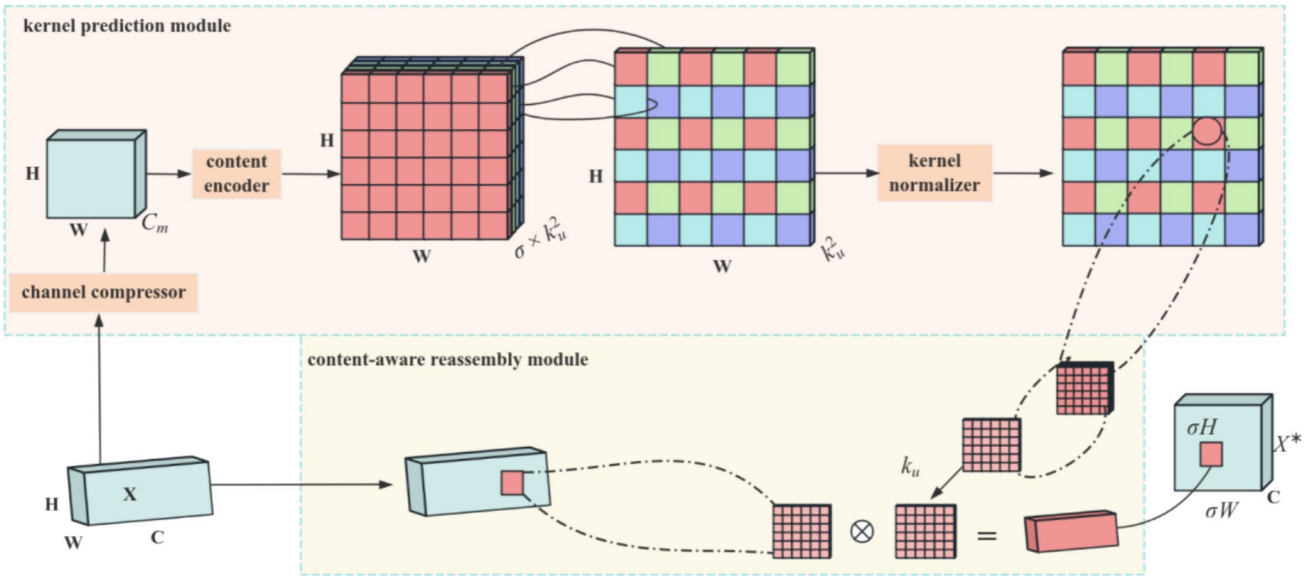


Fig. 5 The structure of the CARAFE module

shows excessive sensitivity to the size of bounding boxes when dealing with object detection tasks. This sensitivity makes the model more susceptible to changes in the size of bounding boxes, resulting in the network focusing too much on size adjustment and ignoring other key features of the target, such as shape, texture, and contextual information. This bias will limit the model's comprehensive understanding and accurate detection of the target, thereby affecting the overall accuracy of target detection. Therefore, balancing the focus on bounding box size with other important features is key to improving object detection performance [23].

To address this problem, the MPDIOU function is used instead of the CIOU loss function of the YOLOv7 model. The MPDIOU function is a loss function for bounding box regression. The MPDIOU function is designed to solve the problem of that the existing loss functions cannot be optimized efficiently when the predicted bounding box is completely different from the real bounding box. The process can be formulated as follows:

$$d_1^2 = \left(x_1^B - x_1^A\right)^2 + \left(y_1^B - y_1^A\right)^2 \tag{8}$$

$$d_2^2 = \left(x_2^B - x_2^A\right)^2 + \left(y_2^B - y_2^A\right)^2 \tag{9}$$

$$Loss_{MPDIOU} = Loss_{IOU} - \frac{d_1^2}{w^2 + h^2} - \frac{d_2^2}{w^2 + h^2} \tag{10}$$

where the input of two arbitrary rectangles as $A, B \subseteq S \in R^n$, the output is $Loss_{MPDIOU}$, rectangular boxes A and B, $\left(x_1^A, y_1^A\right)\left(x_2^A, y_1^A\right)$ denotes the coordinates of the upper-left and lower-right points of rectangular box A and $\left(x_1^B, y_1^B\right)\left(x_2^B, y_1^B\right)$ denotes the coordinates of the upper-left and lower-right points of rectangular box B, respectively.

3 Experimental design

3.1 Experimental environment

(1) Hardware configuration for the experiment: Intel(R) Core(TM) i7-7700 CPU @ 3.60 GHz, AMD Radeon R7 430 and Intel(R) HD Graphics 630.
(2) The software environment for the experiment: Windows 10, Python 3.8, Pytorch 1.9.0.
(3) Parameter settings: period learning rate is 0.1, input image size is 640×640×3, batch size is 8 and epoch is 100.

3.2 Model training

When training the YOLOv7-GCM network using a dataset, it is necessary to place the images and their corresponding labels in specified file paths. The image files and label files need to be added to the images and labels sub-directories within the data folder. After adding the data correctly, it needs to be processed to generate train.txt, val.txt and test.txt files for training. The data.yaml file was modified in the directory. And the names of the 24 types of waste was entered the data.yaml file. The period learning rate and the batch size were adjusted in the train.py file. The weight files generated during training are saved in the runs folder, which is used to predict new images.

3.3 Evaluation metrics

To evaluate the effect of detection for YOLOv7-GCM model, this paper experiments with three evaluation indexes: precision rate [24], recall rate [25], mean of average precision(mAP@0.5) [26]. Compared to the actual category, the predicted results able to divide into four categories: false positive (*FP*), true positive (*TP*), true negative (*TN*) and false negative (*FN*). We often use TP and TN to describe the correct prediction of the model for positive and negative categories in the fields of data analysis and machine learning. TP means that the model correctly identified the actual positive samples, while TN means that the model correctly identified the actual negative samples. Corresponding to true positives and true negatives are FP and FN, which respectively represent the situation where the model makes incorrect predictions. FP refer to the model mistakenly predicting negative samples as positive, while FN refer to the model mistakenly predicting positive samples as negative. The precision rate is a commonly used indicator for evaluating model performance. Another important evaluation metric is the recall rate. The recall rate focuses on how many samples that are actually positive are correctly predicted by the model as positive. The calculation formula for recall rate is TP divided by (TP+FN). A high recall rate usually means that the model can capture more positive samples, but it may also be accompanied by a higher false positive rate [27]. Their calculations are as follows:

$$\text{Pr}\,ecision = \frac{TP}{TP + FP} \times 100\% \tag{11}$$

$$\text{Re}call = \frac{TP}{TP + FN} \times 100\% \tag{12}$$

$$AP = \sum_{i=1}^{n-1} \left(r_{i+1} - r_i\right) P_{\text{inter}}\left(r_i + 1\right) \tag{13}$$

$$mAP = \frac{\sum_{i=1}^{N} AP_i}{N} \tag{14}$$

where $r_1, r_2, ..., r_{N-1}, r_N$ is the value of the first interpolated value of the precision interpolated segment according to the ascending order to interpolate the corresponding recall [28]. N denotes the number of category number, which is taken as 24 in this experiment.

4 Experiment

4.1 Experimental results

Confusion matrix is a commonly used evaluation tool in the field of machine learning, especially in classification problems, which can clearly reveal the differences between model predictions and actual labels. When discussing the performance of the YOLOv7 model, the confusion matrix provides an intuitive and quantitative perspective. Object detection typically involves multiple categories in the YOLOv7 model, each of which has a corresponding prediction probability. These predicted probabilities are based on the model's analysis of the features of each region in the image. The confusion matrix helps us understand the comparison between these predicted results and the actual labels [29]. The confusion matrix of the model is shown in Fig. 6a below, which includes 24 types of waste labels. The horizontal axis labels represent the true target samples in the confusion matrix, while the vertical axis labels represent the predicted target samples. A model would present a diagonal matrix in an ideal state that, which indicates the model has excellent classification capabilities. The relationship between the true values and predicted values of the labels in this experiment basically conforms to the characteristics of a diagonal matrix. Among them, some labels have fewer training samples, which may lead to lower detection accuracy. Secondly, water flow may cause waste to deform during capturing images of waste for a quadruped robot, which results in lower detection accuracy for 'pink-white plastic bags' and 'white boards' compared to other types of waste. The model's higher performance is demonstrated by the excellent prediction and classification effect for labels.

Figure 6b illustrates the results of precision rate for YOLOv7 model and YOLOv7-GCM model. As illustrated in Fig. 6b, the comparison curves of precision rate can be seen that both models have not stabilized before the first 60 rounds of training. The precision rate of both models improved rapidly in first 60 rounds of training. The recognition performance of YOLOv7-GCM model gradually stabilized at around 0.958 when the number of training epochs reaches 80. YOLOv7 model was not as stable as YOLOv7-GCM model for precision rate. The precision rate of YOLOv7 model remained stable at around 0.916. Figure 6c illustrates the results of recall rate for YOLOv7 model and YOLOv7-GCM model. As shown in Fig. 6c, the comparison curves can be seen that the recall rate of the YOLOv7 model is superior to that of the YOLOv7-GCM model. As depicted in Fig. 6d, it illustrates the mAP@0.5 of two models. The mAP@0.5 of YOLOv7-GCM network and YOLOv7 network tend to stabilize after training for about 60 epochs but the mAP@0.5 of the YOLOv7-GCM network is superior to the YOLOv7 network.

The loss curves of the YOLOv7-GCM model is shown in Fig. 6e as follows. Box denotes bounding box loss which is used to measure the difference between the predicted bounding box of the model and the actual bounding box [30]. The initial value of the box loss curve was 0.0748 which eventually stabilized at 0.0185. The model has become increasingly accurate in predicting the positions of target bounding boxes. The objectness loss is used to measure the model's performance in determining whether a bounding box contains an object or not. The initial value of the objectness loss curve was 0.0208 which gradually decreased and eventually converged to 0.0075. This indicates that the model's performance in determining whether a bounding box contains an object has gradually improved. And the model has achieved a good detection capability. The classification loss is used to measure the model's performance on the classification task, which specifically the difference between the predicted target categories and the actual categories. The initial value of the classification loss curve was 0.046, which eventually converged to 0. The model's performance on the object detection task has improved. The val box, val classification and val objectness losses denote the box loss, classification loss and objectness loss on the validation set, respectively. Data augmentation can be utilized to boost the diversity of training data, which can help reduce the loss on the validation set.

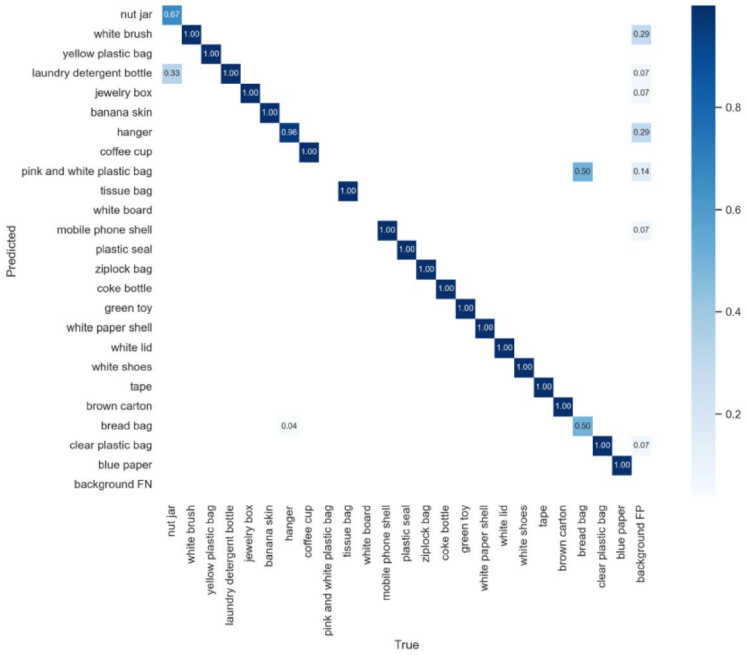
Some of the waste detection results of the YOLOv7-GCM model is shown in Fig. 7. The YOLOv7-GCM model is better detection performance for different types of waste. This model has few false positives and omissions which can accurately detect incomplete targets. However, there are also some individual types of waste that don't have a high confidence level. Because the exposed area of waste is small in complex environments. And the color of waste is similar to the background color.

4.2 Comparative experiments

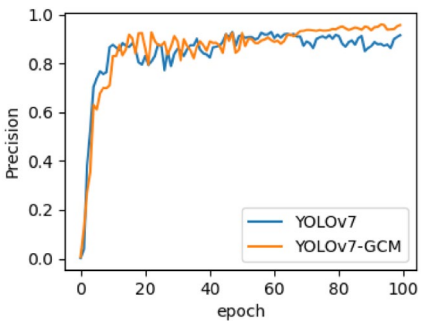
In order to verify the efficacy of the YOLOv7-GCM model, it is compared with detection methods using identical settings and datasets. These detection models include YOLOv7, YOLOv7-GCM, YOLOv5s, Faster RCNN and SSD models. The performance of each model is shown in Table 1.

Table 1 compares the precision rates, recall rates, mAP@0.5, FLOPs and parameters of those models on the same datasets. The mAP@0.5 of the improved

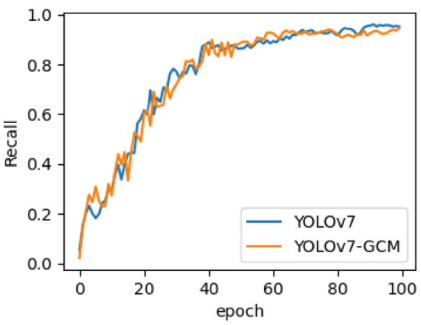
Fig. 6 The curve chart of evaluation indicator



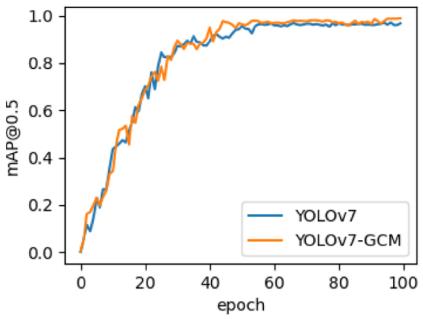
(a) confusion-matrix



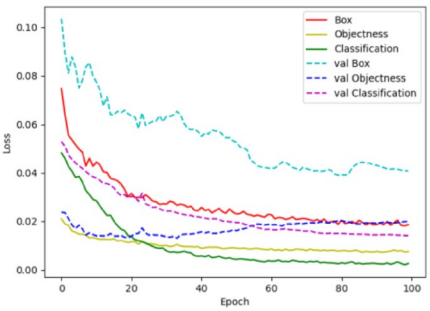
(b) precision rate curve



(c) recall rate curve



(d) mAP@0.5 curve



(e) loss curve

YOLOv7-GCM model in this comparative experiment is as high as 98.8%. The precision rate is also the highest among these models to reach 95.8%. The recall rate of the improved YOLOv7-GCM model is as high as 94.4%. The precision rate of the YOLOv7-GCM model has been improved by 4.2% compared to the YOLOv7 model, which has been improved the performance of the YOLOv7-GCM model. Combined with the above analysis, the YOLOv7-GCM model has obvious advantages in precision rate and arithmetic power compared to other detection algorithms for

Fig. 7 The YOLOv7-GCM model detection effect

Table 1 The outcomes of these models' experiments

Model	Precision (%)	Recall (%)	mAP@0.5 (%)	FLOPs(G)	Parameters (M)
YOLOv7	91.6	95.2	96.7	103.6	36.6
YOLOv7-GCM	95.8	94.4	98.8	102.0	35.3
Faster RCNN	86.0	67.7	79.1	238.6	136.8
YOLOv5s	86.2	88.9	93.4	16.1	7.5
YOLOv3	76.5	78.9	73.4	27.9	61.5
SSD	79.3	57.1	74.4	62.3	92.16

creek waste. The YOLOv7-GCM model has the best detection performance.

4.3 Ablation experiments

This article integrates the GAM into the head structure of the YOLOv7 model, which can achieve accurate waste identification in complex backgrounds and underwater conditions by minimizing information dispersion. The up-sampling in the YOLOv7 network was modified to the CARAFE. And the CIOU function was modified the MPDIOU function. Therefore,There are three components to the ablation experiment that was utilized in this article to confirm the effectiveness of the improvement points. Three parts are respectively used to verify the GAM, the CARAFE and the ablation experiment of modifying the CIOU function of the YOLOv7 network to the MPDIOU loss function. To verify the effectiveness of the improvement points in the YOLOv7 model in this experiment, the improved model is called the YOLOv7-GCM model. The results of each improved part is shown in Table 2.

Table 2 presents the results of the ablation experiment, which shows that adding the GAM, CARAFE module and MPDIOU function to enhance the performance of the model. It can be seen that adding the GAM to the YOLOv7 network model. The experiment has smaller floating-point calculations and parameter but the recall rate, precision rate and mAP@0.5 were decrease. We can see that using the CARAFE of YOLOv7 model has improved its precision rate and mAP@0.5. But the recall rate of the experiment has decreased by 0.4% and the FLOPs has increased by 0.2G. The CARAFE used different up-sampling kernels for different feature layers which focuses more on global information of features than traditional up-sampling. The YOLOv7 model has added the MPDIOU loss function, which includes the recall rate, precision rate and mAP@0.5 of the model has been improved by 0.8%, 0.4% and 0.8%, respectively. To sum up, the YOLOv7-GCM model in this study showed

Table 2 Comparison table of ablation experiments

Model	GAM	CARAFE	MPDIOU	Precision (%)	Recall (%)	mAP@0.5 (%)	FLOPs(G)	Parameters (M)
YOLOv7	–	–	–	91.6	95.2	96.7	103.6	36.6
	√	–	–	90.6	94.3	95.0	101.1	34.7
	–	√	–	93.7	94.8	96.9	103.8	35.3
	–	–	√	92.4	95.6	97.5	103.6	36.6
	√	√	√	95.8	94.4	98.8	102.0	35.3

a slight decrease in recall rate compared to the YOLOv7 network. But the YOLOv7-GCM model increased the precision rate and mAP@0.5 by 4.2% and 2.1%, respectively. Additionally, the FLOPs and parameters were also reduced which has indicated the effectiveness of the YOLOv7-GCM model.

4.4 Practical detection

We will connect the UP board on the quadruped robot to a GPU to improve its computing power and enable the quadruped robot to successfully run the YOLOv7-GCM model. Firstly, we turn off the UP board and disconnect the power supply of the quadruped robot, and connect the corresponding interface of the GPU to the UP board. Then restart and start the UP board, configure and install the GPU driver, and finally verify whether the GPU is successfully installed. We loaded the YOLOv7-GCM model into a quadruped robot and ran the robot in a stream for object detection. The test results are shown in Fig. 8. The YOLOv7-GCM model initially did not detect the target in the garbage detection process. But after 1 s, it detected the target appearing in the camera, and then continued to detect the target until it disappeared. Figure 8 shows the real-time detection image of a quadruped robot running the YOLOv7-GCM model.

5 Conclusions

The article explores the optimization and improvement of a detection algorithm for creek waste based on the YOLOv7 model. We particularly focused on optimizing the head structure of the model by cleverly integrating the GAM into the head structure of YOLOv7, enabling the model to more effectively capture global contextual information in images. In addition, we have also innovated the up-sampling module of the model. Traditional up-sampling methods may lead to the loss of feature information, which is particularly evident when processing high-resolution images. To overcome this challenge, we adopted CARAFE, which not only achieves efficient up-sampling but also ensures the stability and accuracy of detection performance. We have also made improvements in the loss function. We used the MPDIOU loss function to replace the original CIOU function. The MPDIOU loss function can better adapt to objects of different scales and shapes, especially in scenarios such as creek waste detection. The accuracy and robustness of detection can be greatly enhanced by employing the MPDIOU loss function because waste comes in a variety of shapes and sizes. We have developed a new algorithm called YOLOv7-GCM through this series of improvements. The outcomes of the experiment demonstrate that the YOLOv7-GCM model not only has faster convergence speed, but also significantly improves detection performance. Specifically, its mAP@0.5 reached 98.8%, which is 2.1% higher than the original YOLOv7 network. Although the recall rate decreased by 0.8%, the precision rate significantly improved by 4.2%. This indicates that the model can more accurately identify waste in creek while maintaining a high recall rate. This research achievement not only provides an efficient and accurate solution for the field of creek waste detection, but also provides new ideas and methods for target detection tasks in similar complex scenes in the future.

Fig. 8 Real-time detection images

Acknowledgements This project is sponsored by the Natural Science Foundation of Guangxi Zhuang Autonomous Region (2021GXNSFAA220091) and the Wuzhou Central Leading Local Science and Technology Development Fund Project Grant No. 202201001.

Author contribution Jianhua Qin: conceptualization Honglan Zhou: writing Huaian Yi: supervision Luyao Ma: data curation Jianhan Nie: resources Tingting Huang: visualization.

Data availability The code and data supporting this research are stored in the Science Data Bank of generalist data repository, and the access link is https://www.scidb.cn/en/s/6FrUJf.

Declarations

conflict of Interests The authors declare no competing interests.

References

1. Plötz T, Guan Y (2018) Deep learning for human activity recognition in mobile computing. Computer 51(5):50–59. https://doi.org/10.1109/MC.2018.2381112
2. Mittal P, Singh R, Sharma A (2020) Deep learning-based object detection in low-altitude UAV datasets: a survey. Image Vis Comput 104(3):104046. https://doi.org/10.1016/j.imavis.2020.104046
3. Chen X, Li J (2019) Research on an Efficient Single-Stage Multi-object Detection Algorithm. In: 2019 International Conference on Smart Grid and Electrical Automation (ICSGEA). https://doi.org/10.1109/ICSGEA.2019.00110
4. Kumar A, Kalia A, Verma K, Sharma A, Kaushal M (2021) Scaling up face masks detection with YOLO on a novel dataset. Optik 239:166744. https://doi.org/10.1016/j.ijleo.2021.166744
5. ALEXEY B, WANG C, LIAO H. YOLOv4: Optimal speed and accuracy of object detection, https://arxiv.org/abs/2004.10934
6. Gai R, Chen N, Yuan H (2023) A detection algorithm for cherry fruits based on the improved YOLO-v4 model. Neural Comput Applic 35:13895–13906. https://doi.org/10.1007/s00521-021-06029-z
7. Al Muksit A, Hasan F, Emon MF, Haque MR, Anwary AR, Shatabda S (2022) YOLO-Fish: A robust fish detection model to detect fish in realistic underwater environment. Ecol Informatics 72:101847. https://doi.org/10.1016/j.ecoinf.2022.101847
8. Girshick R, Donahue J, Darrell T et al (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. IEEE Comput Soc. https://doi.org/10.1109/CVPR.2014.81
9. Ren S, He K, Girshick R et al (2017) Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Analy Machine Intell 39(6):1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031
10. Cheng X, Fei H, Song L, Zhu J, Ming Z, Wang C, Yang L, Ruan Y (2023) A novel recyclable garbage detection system for waste-to-energy based on optimized CenterNet with feature fusion. J Signal Process Syst 95(1):67–76. https://doi.org/10.1007/s11265-022-01811-1
11. Tian M, Xiali LI, Kong S et al (2022) A modified YOLOv4 detection method for a vision-based underwater garbage cleaning robot. J Front Inf Electron Eng 23:12
12. Hou C, Guan Z, Guo Z et al (2023) An improved YOLOv5s based scheme for target detection in a complex underwater environment. J Marine Sci Eng 11:1041
13. Zhang Q, Chang X, Meng Z et al (2021) Equipment detection and recognition in electric power room based on faster R-CNN. J Procedia Comput Sci 183:324–330. https://doi.org/10.1016/J.PROCS.2021.02.066
14. Abdulghani AM, Abdulghani MM, Walters WL et al (2023) Multiple data augmentation strategy for enhancing the performance of YOLOv7 object detection algorithm. J Tech Sci Press. https://doi.org/10.32604/JAI.2023.041341
15. Shamsuzzaman JM (2022) YOLObin: non-decomposable garbage identification and classification based on YOLOv7. J Comput Commun 10:104–121
16. Stancilas S, Pathinarupothi RK, Gopalakrishnan U (2013) Detection of Pathological Markers in Colonoscopy Images using YOLOv7. In: 2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS).0 [2023–12–29], https://doi.org/10.1109/ICICCS56967.2023.10142724
17. Wang S, Wu D, Zheng X (2023) TBC-YOLOv7: a refined YOLOv7-based algorithm for tea bud grading detection. Front Plant Sci 14:1223410. https://doi.org/10.3389/fpls.2023.1223410
18. Sun Y, Zhang S, Shi Y, Tang F, Chen J, Xiong Y, Dai Y, Li L (2024) "YOLOv7-DCN-SORT: An algorithm for detecting and counting targets on Acetes fishing vessel operation. Fisheries Res 274:106983
19. Wang J, Chen K, Rui X, Liu Z, Loy CC, Lin D (2021) CARAFE++: unified content-aware ReAssembly of FEatures. IEEE Trans Pattern Analy Machine Intell. https://doi.org/10.1109/TPAMI.2021.3074370
20. An K, Duanmu H, Zhiyang W, Liu Y, Qiao J, Shangguan Q, Song Y, Xiaonong X (2024) Enhancing small object detection in aerial images: a novel approach with PCSG model. Aerospace 11:392
21. Kim TK, Kim JS, Cho HC (2023) Deep-learning-based gestational sac detection in ultrasound images using modified YOLOv7-E6E model. J Animal Sci Technol 65:627–637. https://doi.org/10.5187/JAST.2023.E43
22. Zheng Z, Wang P, Ren D et al (2021) Enhancing geometric factors in model learning and inference for object detection and instance segmentation. IEEE. https://doi.org/10.1109/TCYB.2021.3095305
23. Duan K, Xie L, Qi H et al. (2021) Location-Sensitive Visual Recognition with Cross-IOU Loss. https://doi.org/10.48550/arXiv.2104.04899
24. Liu X, Gan H, Yan Y (2021) Study on improvement of YOLOv3 algorithm. J. Phys Conf Series 1884:012031. https://doi.org/10.1088/1742-6596/1884/1/012031
25. Jin Q, Han Q, Su N et al (2023) A deep learning and morphological method for concrete cracks detection. J Circuits Syst Comput. https://doi.org/10.1142/S0218126623502717
26. Konala TR, Nammi A, Tella DS (2023) Analysis of Live Video Object Detection using YOLOv5 and YOLOv7. In: 4th International Conference for Emerging Technology (INCET).0 [2023–12–29]. https://doi.org/10.1109/INCET57972.2023.10169926
27. Modha DS, Akopyan F et al (2023) Neural inference at the frontier of energy, space, and time. Science 382:329–335. https://doi.org/10.1126/science.adh1174
28. Gang X, Yue Q, Liu X (2023) Realtime monitoring of concrete crack based on deep learning algorithms and image processing techniques. Adv Eng Inf 58:102214
29. Song Z, Huang X, Ji C, Zhang Y (2023) Intelligent identification method of hydrophobic grade of composite insulator based on efficient - former network. IEEJ Trans Electr Electron Eng 18:1160
30. Li Z, Zhu Y, Sui S, Zhao Y, Liu P, Li X (2024) Real-time detection and counting of wheat ears based on improved YOLOv7. Comput Electron Agricul 218:108670