YOLO-CFDU: A Deep Learning Approach for Urban Road Garbage Detection

Peizhen Chen, Yingchao Zhou, Bo Li, Zhen Xiao and Bofeng Xue

Abstract—**With the acceleration of urbanization, the efficient and precise detection of road garbage is critical for urban environmental health management. To address the challenges of variable object morphology size and interference from complex backgrounds in urban road garbage detection, we propose an improved model—YOLO-CFDU based on YOLOv11. Firstly, the C3K2 module is replaced by C3K2-FE, which integrates the Fasterblock and EMA attention mechanism to enhance multi-scale target feature extraction while reducing computational complexity. Secondly, the Dynamic Spatial Weighted Fusion (DSWF) module is proposed in Neck network to strengthen target region features by generating spatial weight maps. Thirdly, the Unified-IoU (U-IoU) loss is introduced to optimize the accuracy of the bounding box regression. Lastly, LAMP technology is used to prune the model, reducing the number of parameters and calculations. Experimental results demonstrate that the proposed model achieves 0.735 for mAP@0.5 on a self-constructed road garbage dataset, outperforming the original YOLOv11n by 4.6%. In addition, the number of model parameters and computations is reduced by 44.0% and 29.7%, respectively, meeting the requirements for lightweight deployment on embedded devices. Real-vehicle tests further validate the model's superior real-time detection performance, providing an efficient and cost-effective solution for intelligent urban garbage management.**

Index Terms—**Deep learning, Garbage detection, YOLOv11, Lightweight, EMA**

I. Introduction

With the acceleration of the urbanization process, the surge of urban population density and the increase of consumption level have led to a continuous rise in the amount of domestic garbage generation [1]. The random disposal and accumulation of garbage not only affects the city appearance, but also may cause air pollution, traffic accidents and other problems, which adversely affects the urban ecological construction [2]. To promote urban environmental health management, various types of sweepers, sanitation robots and other sweeping equipment have been gradually applied to urban roads, squares, parks and other scenes. However, the existing equipment mainly relies on preset routes or timed sweeping modes and lacks real-time sensing of the distribution of roadway trash [3], resulting in insufficient sweeping in some areas, while relatively clean areas are still unnecessarily swept, which affects the overall sweeping efficiency. Therefore, how to realize the real-time perception of road garbage through accurate garbage detection and provide intelligent decision support for sweeping equipment has become an urgent problem in the field of urban sanitation [4].

In recent years, deep learning techniques have made significant breakthroughs in the field of object detection [5], especially convolutional neural network (CNN)-based detection frameworks such as Faster R-CNN [6], SSD [7] and YOLO [8] [9]. Among them, two-stage detection models such as Faster R-CNN rely on the regional proposal network to generate candidate frames, which, despite its high accuracy, has a high computational complexity and is difficult to meet the real-time requirements for deployment on mobile devices. In contrast, single-stage detection models such as YOLO directly regress the target categories and bounding boxes in an end-to-end manner, which can realize efficient inference while guaranteeing accuracy, and is more suitable for real-time detection application scenarios.

In response to the challenges in garbage detection tasks, scholars have proposed various optimization solutions based on the above detection approaches. Feng et al [10] introduced ResNet50 network to optimize the Faster R-CNN network framework for the target size difference challenge, but the inference speed is relatively slow. To improve the detection efficiency, single-stage detection models are gradually used to deal with the task of garbage detection [11]. He et al [12] proposed the EC-YOLOX model, which enhances the feature extraction of garbage of different morphologies by introducing the CA and ECA dual-attention mechanisms, but the increase in model computation leads to a slight decrease in the detection speed. In the garbage detection scenario, small target garbage can be difficult to accurately identify due to background interference. Kuang et al [13] introduced Transformer predictive detection head based on YOLOv5 and combined it with CBAM attention mechanism to enhance the detection effect of small targets. Li et al [14] designed noise suppression module ANSM to reduce the interference of background noise such as illumination, reflection, and so on; Xia et al [15] proposed a lightweight YOLO-MTG model, introduced MobileViTv3 to enhance the global feature expression, and optimized the Neck network by feature reuse technique, which solved the problem of balancing the

Manuscript received April 10, 2025; revised July 6, 2025.

This work was supported by the National Natural Science Foundation of China (52375105), the Excellent Young Talents Fund of Shandong Province (ZR2022YQ51), and the Major Science and Technology Innovation Project of Shandong Province (2019JZZY010911).

Peizhen Chen is a postgraduate student at the School of Transportation and Vehicle Engineering, Shandong University of Technology, Zibo 255000, China (e-mail: chen15063702271@163.com).

Yingchao Zhou is a lecturer at the School of Transportation and Vehicle Engineering, Shandong University of Technology, Zibo 255000, China (Corresponding author, e-mail: zhouyingchao@sdut.edu.cn).

Bo Li is a professor at the School of Transportation and Vehicle Engineering, Shandong University of Technology, Zibo 255000, China (e-mail: libo@sdut.edu.cn).

Zhen Xiao is a lecturer at the School of Transportation and Vehicle Engineering, Shandong University of Technology, Zibo 255000, China (e-mail: xiaozhen@sdut.edu.cn).

Bofeng Xue is a postgraduate student at the School of Transportation and Vehicle Engineering, Shandong University of Technology, Zibo 255000, China (e-mail: yudjffi@163.com).

efficiency and accuracy in multi-target spam detection. In order to reduce the number of model parameters and realize the real-time deployment requirements, Bai et al [16] used group-level pruning method to improve YOLOv7 in a lightweight way, and the experimental results show that the number of parameters is reduced by 6.05% compared with the original model; Jiang et al [17] used lightweight convolution Ghost Conv to replace the standard convolution in YOLOv8 and remove the C2f module, and introduced the GTR module and the SimAM attention mechanism guided pyramid network to optimize the feature extraction efficiency.

In the above context, although existing research has improved the feature extraction and detection effect of different morphological sizes of garbage to a certain extent by introducing multiple attention mechanism and feature fusion strategy, such improvement is often accompanied by a larger number of model parameters and computational burden, which is difficult to meet the real-time detection requirements of embedded devices. Moreover, the complex real road environment such as uneven road texture, light intensity changes and other background interference will also reduce the garbage detection precision. To address the above problems, we propose an improved model YOLO-CFDU based on YOLOv11n, with the following main contributions:

1) Design the C3K2-FE module to replace the C3K2 module, reducing redundant computation through PConv and enhancing multi-scale target feature extraction by combining the EMA attention mechanism;

2) Propose the DSWF module to highlight the feature expression of target region and suppress the background noise interference through weight assignment;

3) Introduce U-IoU loss, combining dynamic scaling strategy and Focal Loss mechanism to optimize the bounding box regression accuracy;

4) Adopt LAMP pruning technique to reduce the number of model parameters and calculations, realizing lightweight deployment of the model.

II. Methodology

A. YOLOv11

YOLOv11 [18] was released by Ultralytics in September 2024, and the algorithm further strengthens the support for multi-tasks such as target detection, instance segmentation, and image classification based on inheriting the real-time and efficient features of the YOLO series. Among them, YOLOv11 performs excellently in the task of target detection, can quickly and accurately locate and recognize targets, and is widely used in many fields such as security surveillance, autonomous driving, and industrial inspection. According to the network depth, YOLOv11 also provides five different model sizes: n, s, m, l, and x. The model detection accuracy increases with the number of model parameters and computation amount, which is suitable for different task requirements and computational resources. The network structure of YOLOv11 consists of three parts, the Backbone network, the Neck structure and the Head structure. Its network structure is shown in Fig. 1.

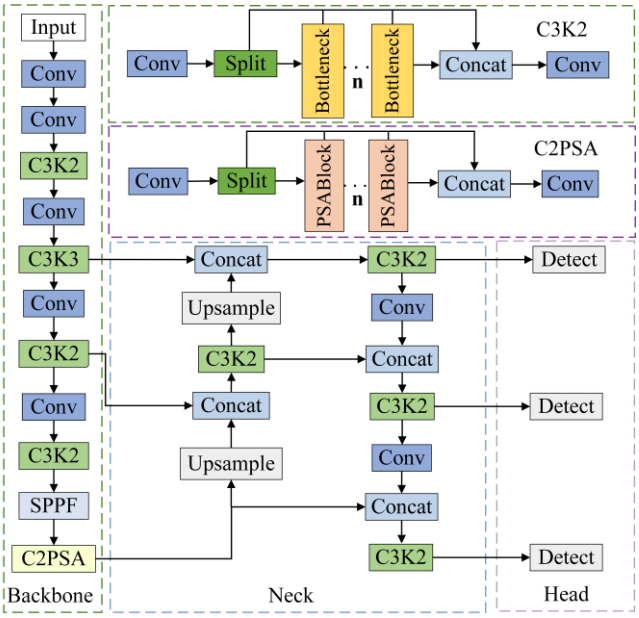


Fig. 1. YOLOv11 network structure

Compared with YOLOv8, YOLOv11 replaces C2f with C3K2 in Backbone network to improve the feature extraction efficiency and introduces the C2PSA module after the SPPF layer to improve the multiscale feature extraction. The Neck structure combines PAN and FPN as the feature fusion network, and makes features more fully fused by the cross-layer splicing design of top-down and bottom-up. The Head structure adopts decoupled design and DWConv operation, which significantly reduces the number of model parameters and computation while improving the detection flexibility.

B. YOLO-CFDU

Due to the different sizes of garbage morphology and the interference from the complex background of road surface and light intensity, the existing detection methods are still deficient in accuracy and efficiency. For this reason, we propose an object detection model YOLO-CFDU for urban road garbage based on YOLOv11, and its network structure is shown in Fig. 2.

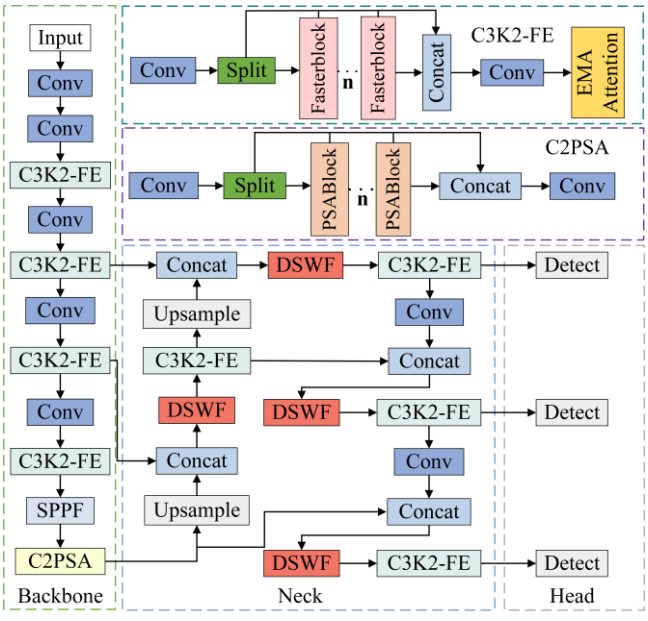


Fig. 2. YOLO-CFDU network structure

1) C3K2-FE

In the garbage detection task, garbage exhibits diverse morphological sizes, and the small-sized garbage is easy to be ignored, which puts high requirements on the feature extraction and expression ability of the model. In addition, the traditional convolutional operation has parameter redundancy in cross-channel feature fusion, leading to inefficient feature transfer and increased inference latency. To solve this problem, we adopt the C3K2-FE module to replace the C3K2 module of the original network, which realizes fast feature extraction by PConv convolution and improves the model's adaptability to targets of different sizes by adding the EMA attention mechanism. The structure of the C3K2-FE module is shown in Fig. 3.

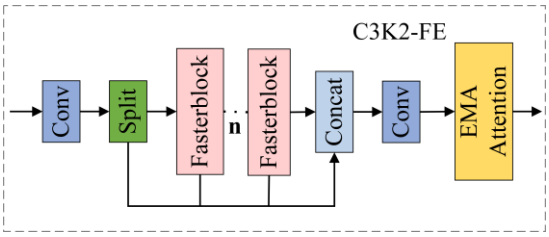


Fig. 3. C3K2-FE module structure

The core idea of Fasterblock [19], as a key component in the FasterNet structure, is to reduce redundant computations and memory accesses through PConv, which maintains or even improves the feature extraction and representation ability of the model while reducing the computational cost. Assuming that the size of the input feature map is $C\times H\times W$, PConv only performs $k\times k$ convolution operation on $C/4$ of these channels, and the rest of the channels are directly retained without any computation. The GFLOPs of PConv can be expressed as:

$$GFLOPs = H \times W \times k^2 \times C_p^2 \tag{1}$$

Where: C_p = 1/4C, the GFLOPs of PConv are only 1/16 of the normal Conv, significantly reducing the number of model parameters. The two subsequent 1×1 Conv are responsible for cross-channel information interaction, ensuring that the network captures both local details and global information. The structure of the Fasterblock is shown in Fig. 4.

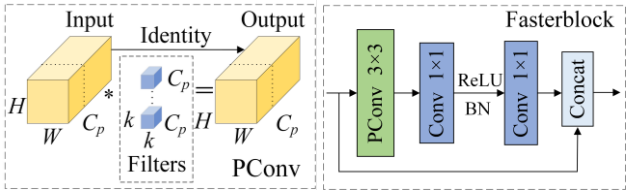


Fig. 4. Fasterblock structure

The EMA module [20] is a new multi-scale learning attention mechanism, which aims to enhance the context-awareness of the model through efficient multi-scale feature fusion. The structure of the EMA module is shown in Fig. 5, and its main computational flow is as follows: first, the input feature map will be divided into G groups, and the dimensionality of each group becomes $C/G\times H\times W$; based on the grouped features, average pooling is performed in the X and Y directions, respectively, and then the biaxial features are spliced and output the two sets of feature vectors by 1×1 Conv, after which they are activated by the Sigmoid function; meanwhile, the third branch captures the local detail features by 3×3 depth-separable Conv; finally, multi-scale fusion of channel attention weights with local features is performed by Cross-spatial learning network, and enhancement of the original input features is accomplished by Sigmoid function and multiplication operation, and the final output is the enhanced feature map that maintains the original resolution.

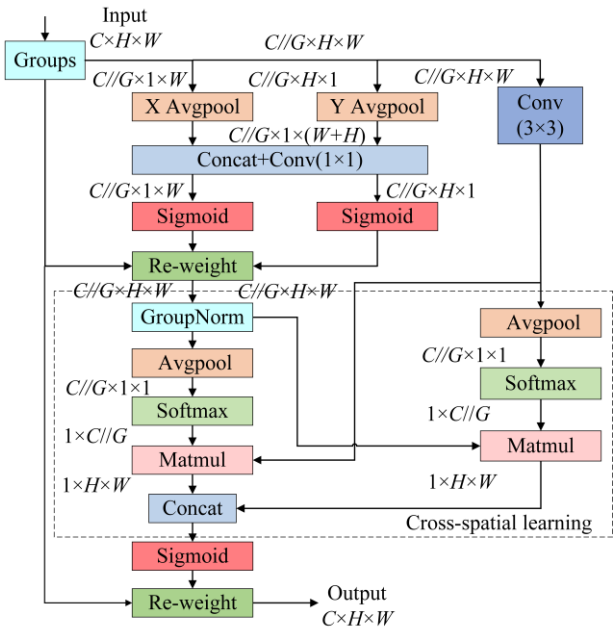


Fig. 5. EMA module structure

2) DSWF

Although the Neck network of YOLOv11 realizes cross-layer multi-scale feature splicing through PANET, in the case of low contrast between target and background or noise interference, it adopts a uniform weighting strategy, which ignores the significance difference of different layer features in spatial dimensions, resulting in that the detailed information of the shallow features is easy to be masked by the background noise [21], which makes it difficult to accurately highlight the target region, affecting the detection accuracy. For this reason, we propose DSWF to spatially selectively enhance multiscale fusion features by dynamically assigning spatial weights after feature splicing and suppressing background noise interference. The structure of DSWF module is shown in Fig. 6.

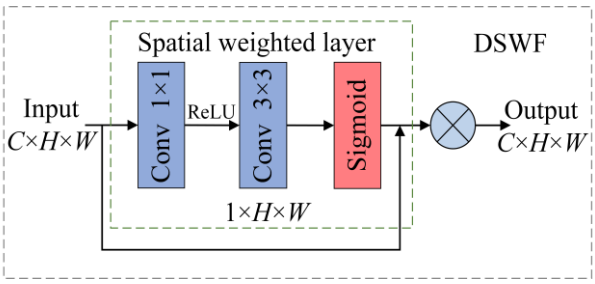


Fig. 6. DSWF module structure

The DSWF module consists of three parts: channel compression, spatial weighting and weighted fusion. In the channel compression stage, 1×1 Conv is used to perform channel dimensionality reduction on the input feature map $F_{in} \in \mathrm{R}^{H\times W\times C}$:

$$F_{comp} = \left(\text{Conv}_{1\times1}(F_{in}) + b\right) \in \text{R}^{C'\times H\times W} \quad (2)$$

Where: C' = $[C/r]$ is the number of channels after compression, r is the compression rate, and b is the bias variable.

In the spatial weighting stage, a single-channel weight feature map is generated based on W_{map}:

$$W_{\text{map}} = f_{\text{sig}}\left(\text{Conv}_{3\times3}^{d=2}\left(f_{\text{ReLu}}(F_{\text{comp}})\right)\right) \in \text{R}^{1\times H\times W} \quad (3)$$

Among them, f_{ReLu} implements feature sparsification to suppress background noise and expands the sensory field using $\text{Conv}_{3\times3}^{d=2}$ with a null rate of 2, f_{sig} constrains the output weights to the [0,1] interval to quantify the spatial feature importance.

Finally, the target area feature expression is highlighted by weighted fusion:

$$F_{out}^{(C,i,j)} = F_{in}^{(C,i,j)} \cdot \mathcal{B}\left(W_{map}^{(i,j)}\right) \in \text{R}^{C\times H\times W} \quad (4)$$

Where: $F_{out} \in \text{R}^{H\times W\times C}$ is the output feature map and $\mathcal{B}$ is the broadcast function extending single-channel weights to C channels.

3) U-IoU loss

Although the regression loss CIOU takes into account the distance between the centroid of the predicted box and the real box as well as the aspect ratio factor on the basis of IOU, similar to DIOU, EIOU and other such loss improvements are still confined to static refinement of geometrical differences, and are unable to dynamically control the gradient contribution of high and low quality samples, leading to slow convergence in the pre-training stage, and difficulty in improving the detection accuracy in the late stage under high IoU thresholds. For this reason, we adopt a new regression loss function U-IoU [22]. The core idea of this method is to dynamically assign weights to different quality prediction box during the training process, so that the model pays more attention to high-quality prediction box, and at the same time strikes a balance between training speed and detection accuracy. Specifically, the scaling ratio between predicted and real box is adjusted by introducing the hyperparameter ratio, as shown in Fig. 7.

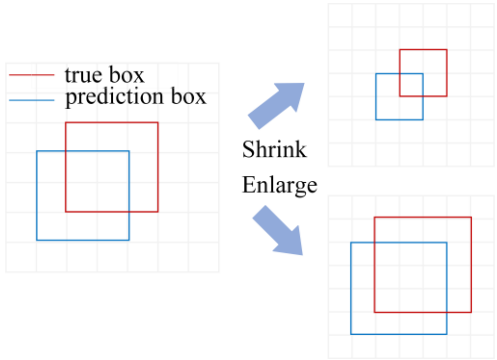


Fig. 7. Illustration of bounding box

At the initial stage of training, the ratio is set to a higher value to enlarge the bounding box and focus on low-quality samples to accelerate convergence, and as the number of training rounds is increasing, the score-decreasing strategy is used to gradually reduce the ratio and shrink the box size to enhance the loss weight of high-quality samples, so as to improve the detection accuracy under the high IoU threshold. In addition, drawing on the idea of Focal Loss, a dual-attention mechanism is introduced to further optimize the weight allocation of bounding box regression by adjusting the weights of different quality prediction boxes through the Focal box and forcing the model to focus on samples that are more difficult to locate according to the difference in confidence.

The ratio is related to the number of training rounds as follows:

$$ratio = \frac{150}{epoch + 100} \quad (5)$$

The Focal Loss expression is:

$$Focal = -\alpha_t\left(1 - p_t\right)^{\gamma}\log(p_t) \quad (6)$$

Where: α_t is a balancing factor, p_t is the predicted probability of target detection incorporating real labels, and γ is a focus factor to control the rate of weight decay for easy and difficult samples.

4) LAMP pruning

To further meet the lightweight deployment requirements of mobile and embedded devices, we adopt the LAMP technique for pruning optimization of YOLO-CFDU. LAMP is a layer adaptive sparse pruning method based on weight magnitude [23], and its core idea is to maximize the compression of the number of model parameters and computation under the premise of guaranteeing the accuracy of the model by dynamically assigning pruning ratios to each layer. LAMP evaluates the importance of the weights of each layer of the network and dynamically allocates pruning ratios for different layers, with less pruning for layers with more critical information and more pruning for layers with more redundant parameters. At the same time, the pruning threshold is determined based on the weight magnitude, and the weights below the threshold are set to zero. Finally, the model is iteratively fine-tuned to recover the accuracy to achieve a good balance between model lightweighting and detection performance. The specific score formula is as follows:

$$Socre(u;W_t) = \frac{(W_t[u])^2}{\sum_{v\geq u}(W_t[v])^2} \quad (7)$$

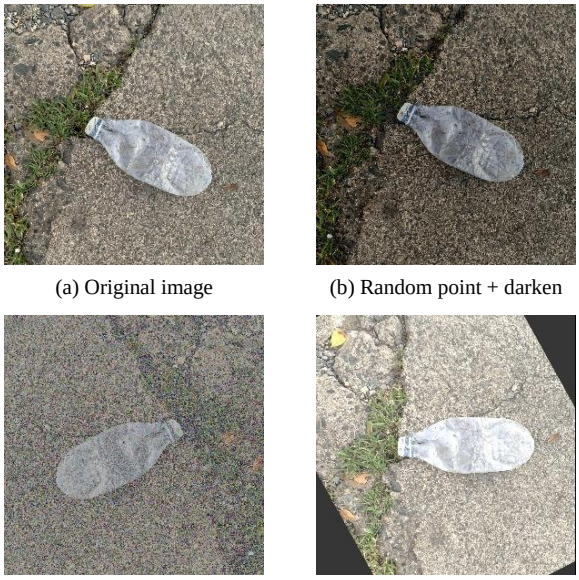
Where: u and v denote the index mappings corresponding to different weights, $W_t[u]$ and $W_t[v]$ denote the weights of index u and v mappings respectively.

III. Experimentation and Analysis

A. Dataset

Currently, the existing publicly available target detection datasets have not yet existed a labeled dataset specifically for urban road garbage scenarios. To meet the research needs, we independently constructed a diverse urban road garbage dataset through field photography and public resource screening. The dataset includes five categories of common road litter such as cigarette butts, plastic bags, plastic bottles, and waste paper. To improve the

generalization ability of the model and enhance the complex environment of the working road, the collected garbage images are enhanced by brightness adjustment, noise and random point addition, panning, flipping, etc., and the dataset is finally expanded to 7043. The effect of garbage image enhancement is shown in Fig. 8.



(a) Original image (b) Random point + darken

(c) Noise + flip + darken (d) Rotate + random point + lighten

Fig. 8. Image enhancement

In the data labeling stage, LabelImg tool was used to create labels for the garbage images, construct the urban road garbage dataset, and divide the dataset into training, validation, and testing sets according to the ratio of 8:1:1. The distribution of the number of instances in each category in the dataset is shown in Fig. 9.

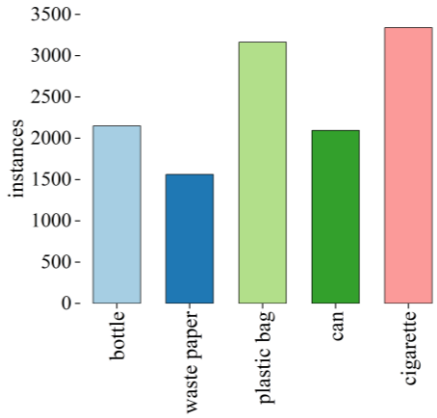


Fig. 9. Distribution of the number of instances of each type of garbage

B. Evaluation Index

In order to comprehensively evaluate the performance of the improved YOLO-CFDU model, we select the evaluation indexes such as Precision, Recall, mAP, Params and FPS. The related calculation formula is shown below:

$$Precison = \frac{TP}{TP + FP} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{10}$$

Where: Precision denotes the proportion of positive cases among the samples predicted by the model; Recall denotes the proportion of correctly predicted positive cases among the samples predicted by the model; TP denotes the number of correctly predicted positive cases; FP denotes the number of incorrectly predicted positive cases; FN denotes the number of samples that are missed; mAP is used to measure the performance of the model under different IoU thresholds, and the higher the value of mAP, the better is the performance of the model detection. Params and FPS reflect the parameter size and computing speed of the model, respectively, which are important indicators of whether the model can run lightly.

C. Experimental Environment

The experiments were completed under the ubuntu18.04 operating system, using the GPU model NVIDIA GeForce RTX 4090D, based on the Python3.8 compilation environment, using the Pytorch2.3.1 Deep learning framework, and the CUDA version of 11.8.The model training parameter settings are detailed in TABLE I.

TABLE I
PARAMETER SETTINGS FOR MODEL TRAINING

Parameters	Parameter value
Epoch	200
Batch	16
Optimizer	SGD
Initial learning rate	0.01
Maximum learning rate	0.01
Weight Decay	0.0005
Image Resolution	640×640

D. Analysis of Experimental Results

1) Comparative Experiment

To verify the performance of the improved YOLO-CFDU model, we selected the mainstream models of YOLO series, YOLOv5n, YOLOv6, YOLOv8n, YOLOv9t, YOLOv10n, and YOLOv11n, to conduct comparative experiments under the same experimental conditions, and the experimental results are shown in TABLE II.

TABLE II
COMPARISON OF EXPERIMENTAL RESULTS

Model	Precision	Recall	mAP@0.5	mAP@0.5-0.95	Params/M	GFLOPs	FPS	Size/MB
YOLOv5n	0.816	0.620	0.683	0.544	2.5	7.2	588.2	5.3
YOLOv6	0.850	0.529	0.602	0.488	4.2	11.9	571.4	8.2
YOLOv8n	0.874	0.645	0.719	0.578	3.0	8.2	606.1	6.2
YOLOv9t	0.855	0.625	0.700	0.569	2.0	7.9	568.7	4.6
YOLOv10n	0.866	0.631	0.714	0.579	2.7	8.4	645.1	5.7
YOLOv11n	0.873	0.630	0.703	0.541	2.5	6.4	602.4	5.5
YOLO-CFDU	0.913	0.666	0.735	0.582	1.4	4.5	617.2	3.3

(a)YOLOv11n (b)YOLO-CFDU

Fig. 10. Confusion matrix

As can be seen from the data in the TABLE II, the improved YOLO-CFDU model significantly outperforms the above comparison models in terms of detection accuracy, computational efficiency, and lightweight performance, with Precision and Recall reaching 0.913 and 0.666, and the mAP@0.5 and mAP@0.5-0.95 are improved by 4.6% and 7.6%, compared to the YOLOv11n baseline model, reflecting the higher detection accuracy at different IoU thresholds. In terms of lightweighting, thanks to LAMP pruning technology, the number of model parameters and computational volume are compressed to 1.4M and 4.5GFLOPs, the inference speed is increased to 617.2FPS, and the model volume is reduced to 3.3MB, which is 28.3% smaller compared to the smallest model in the table, YOLOv9t, and 40.0% smaller compared to the original model, which can meet the requirements of real-time detection and mobile deployment.

The Confusion Matrix can intuitively reflect the prediction accuracy and misclassification of the model for different target categories, and Fig. 10 demonstrates the change status of the confusion matrix of models YOLOv11n and YOLO-CFDU before and after improvement. Compared with the pre-improvement period, YOLO-CFDU significantly improves the recognition accuracy of each garbage category. Among them, the recognition accuracy of cigarette butt is improved by 37%, and the leakage problem of small-size garbage is effectively alleviated. In addition, the misclassification rate of background is significantly reduced, indicating that the model's anti-interference ability against complex background is improved. Overall, the improved model YOLO-CFDU performs better in recognition accuracy, and its generalization ability and robustness are significantly enhanced.

To evaluate the optimization effect of different attention mechanisms on model performance, we further compared the mAP@0.5 variations over 200 training epochs when incorporating ECA, CBAM, SE, and EMA attention modules. The experimental results are shown in Fig. 11. As shown in Fig. 11, the mAP@0.5 of all attention mechanisms exhibits a similar growth trend during the early stages of training. However, as training progresses, the EMA attention mechanism demonstrates a clear advantage, ultimately achieving the highest mAP@0.5 of 0.73, outperforming the other attention modules overall.

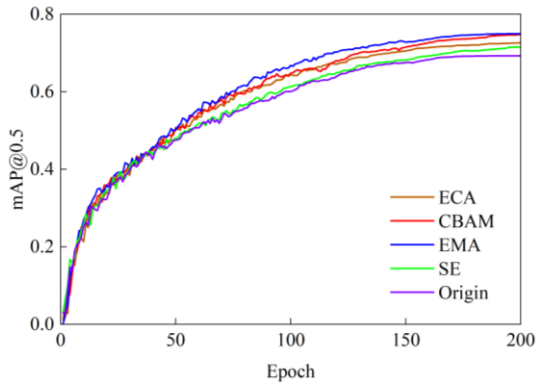


Fig. 11. Variation curves of mAP@0.5 for different attention mechanisms

2) Ablation Experiment

To further explore the contribution of each improved module to the performance of the YOLO-CFDU model, we design ablation experiments by gradually introducing each module and comparing the performance changes when different modules are combined, and the comparison results of the ablation experiments are shown in TABLE III.

TABLE III
COMPARISON RESULTS OF ABLATION EXPERIMENT

C3K2-FE	DSWF	U-IoU	LAMP	Precision	Recall	mAP@0.5	mAP@0.5-0.95	Params/M	GFLOPs	FPS	Size/MB
×	×	×	×	0.873	0.630	0.703	0.541	2.5	6.4	602.4	5.5
√	×	×	×	0.902	0.633	0.731	0.582	2.4	6.7	510.2	5.2
√	√	×	×	0.911	0.656	0.738	0.607	2.5	7.7	492.6	5.4
√	√	√	×	0.918	0.685	0.755	0.617	2.5	7.7	492.6	5.4
√	√	√	√	0.913	0.666	0.735	0.582	1.4	4.5	617.2	3.3

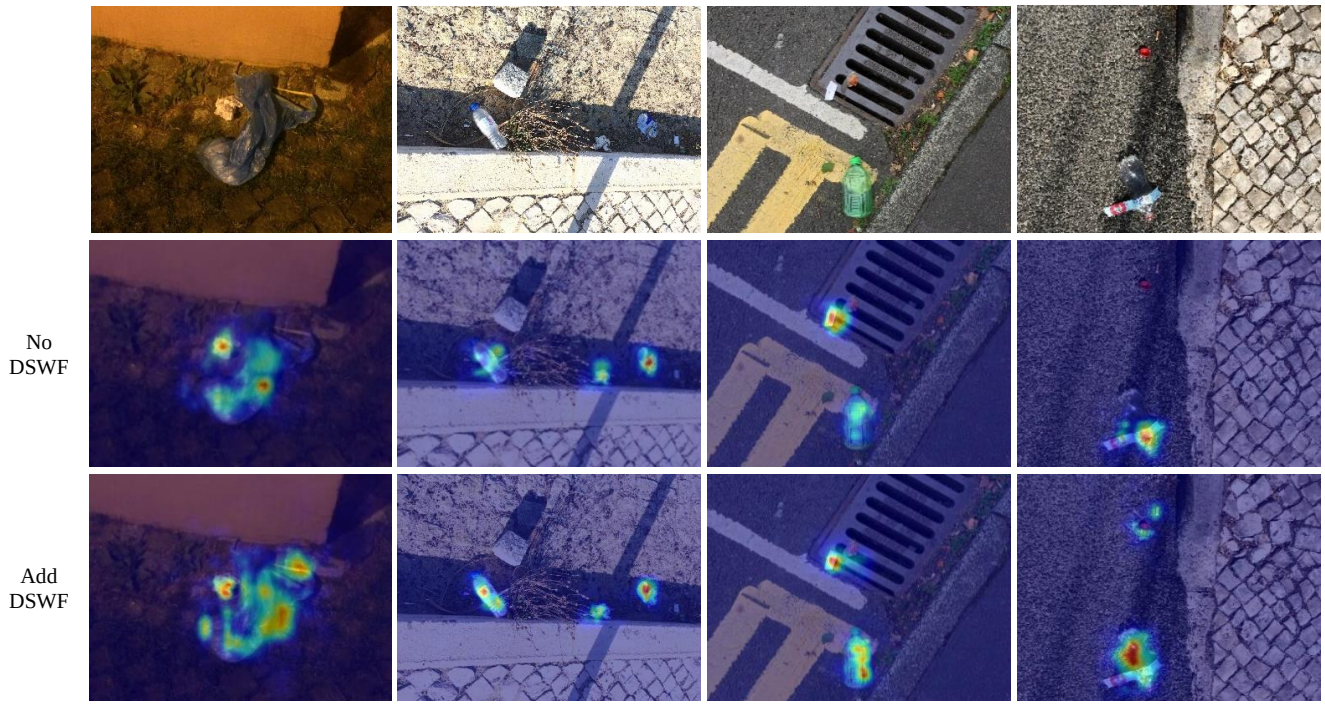


Fig. 12. Feature visualization

The data in the TABLE III shows that after the introduction of the C3K2-FE module, Precision is improved to 0.902, mAP@0.5 is improved by 4.0%, but FPS decreases by 15.3% due to the increase of the computational complexity; after further integration of the DSWF module, Recall is improved by 3.6%, and mAP@0.5 reaches 0.738, which indicates that the module can effectively enhance the target area feature expression; after the introduction of the U-IoU loss function, mAP@0.5 is improved by 2.3% compared with the previous stage, which verifies its optimization effect on the regression accuracy of the bounding box; finally, applying the LAMP pruning technique, the Params is compressed by 44.0%, and the model volume is reduced to 3.3MB. Although mAP@0.5 is slightly decreased by 2.6% compared with that before the pruning, it significantly balances the detection accuracy and the lightweight performance to meet the needs of mobile deployment.

In the ablation experiments, the DSWF module enhances the model's ability to extract target features in complex backgrounds through the dynamic spatial weight allocation strategy. To visually verify its mechanism of action, Fig. 12 demonstrates the feature enhancement effect of the DSWF module. When the DSWF module is not added, the model does not focus on the target region prominently enough in the face of limited illumination, shadow occlusion, complex road texture and other environmental interferences, resulting in an affected detection effect. After adding the DSWF module, the high response region of the heat map is more focused on the target object, indicating that the module can effectively overcome the environmental interference and improve the recognition of target features.

3) U-IoU Loss Analysis

To evaluate the advantage of U-IoU over the original CIoU for bounding box regression, we recorded the curves of the two loss functions with epochs under the same training settings, as shown in Fig. 13.

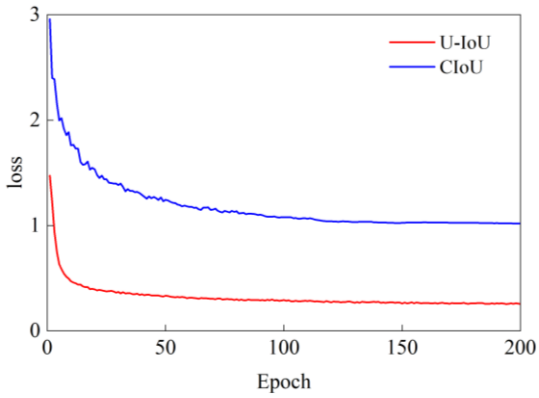


Fig. 13. Loss function variation curve

From the change in the loss curve in Fig. 13, U-IoU shows a faster decline in loss at the beginning of the training period, showing stronger convergence ability. Throughout the entire training process, its loss remains consistently below that of CIoU and ultimately converges to a lower level. In addition, the U-IoU loss curve remains smooth and stable throughout the whole process without significant fluctuations, while the CIoU has slight oscillations during the descent process. These results fully demonstrate the superiority of U-IoU over traditional CIoU in the bounding box regression task, both in terms of convergence speed, final results and training stability.

4) Pruning Experiment

LAMP pruning, as a layer-adaptive structured pruning algorithm, has the advantage that automated model compression can be achieved without intervention in hyperparameter tuning. It is worth noting that due to the heterogeneous nature of topological connections between neural network layers, different pruning strengths impair inter-layer dependencies significantly, which is often accompanied by unavoidable accuracy degradation.

TABLE IV
COMPARISON RESULTS OF DIFFERENT PRUNING RATES

Speed_up	Precision	Recall	mAP@0.5	mAP@0.5-0.95	Params/M	GFLOPs	FPS	Size/MB
1.0	0.918	0.685	0.755	0.617	2.5	7.7	492.6	5.4
1.5	0.915	0.669	0.745	0.594	1.7	5.1	598.8	3.6
1.7	0.913	0.666	0.735	0.582	1.4	4.5	617.2	3.3
2.0	0.906	0.639	0.728	0.578	1.3	3.8	636.9	2.9
2.2	0.899	0.619	0.703	0.552	1.1	3.5	645.1	2.5
2.4	0.869	0.602	0.681	0.534	1.0	3.2	657.8	2.3

Although subsequent accuracy fine-tuning can reconfigure the model representation capability, there is a significant difference in the magnitude of accuracy regain under different Speed_up pruning rates. In order to seek the optimal balance between model compression and accuracy maintenance, we design multiple sets of Speed_up parameter comparison experiments, and the experimental results are shown in TABLE IV.

The data in the TABLE IV shows that when Speed_up is 1.7, the model Params and Size shrink by 44.0% and 38.9%, while mAP@0.5 decreases by only 2.6%. When Speed_up is 1.5, mAP@0.5 drops by just 1.3%, but the computational cost decreases relatively small. When Speed_up exceeds 1.7, the model Size is further reduced, but the loss of accuracy is aggravated, and the compression ratio tends to slow down. Therefore, after comprehensive consideration, the pruning model with a Speed_up of 1.7 is selected as the final optimized model, which significantly reduces the computational cost while maintaining the model characterization ability to the greatest extent. The comparison of the number of channels before and after pruning is shown in Fig. 14.

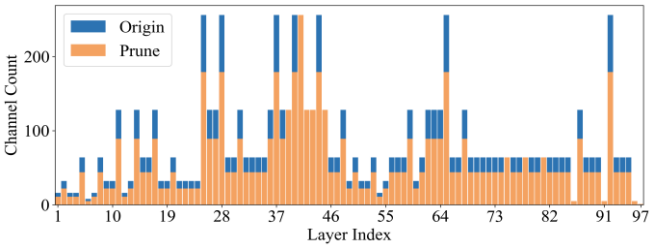


Fig. 14. Comparison of channels before and after pruning

5) Visualization Results

To intuitively demonstrate the garbage detection performance of the improved YOLO-CFDU model, four representative urban road garbage scenarios from the test set are selected for visual comparison. The results are shown in Figs. 15 to 18, with each figure displaying: (a) original image, (b) YOLOv11n output, and (c) YOLO-CFDU output.

Fig. 15 presents a scenario characterized by pavement cracks and weed interference. As shown in Fig. 15(b), the baseline model fails to detect a plastic bag target obscured by the complex background. In contrast, the improved model successfully identifies the target, demonstrating enhanced detection accuracy under challenging visual conditions. For the densely piled garbage scene in Fig. 16, the improved feature fusion strategy effectively enhances the target area feature expression. As a result, multiple overlapping or partially occluded targets, which are difficult to distinguish in Fig. 16(b), are clearly and accurately identified in Fig. 16(c). In Fig. 17, which contains small and inconspicuous targets such as cigarette butts on the road surface, the integration of the EMA attention mechanism significantly boosts the model's sensitivity to fine details, enabling precise localization of the small targets. As shown in Fig. 18, under the interference of strong illumination and shadows, the improved model further detects fuzzy targets that were previously missed due to light interference, while ensuring the accurate identification of the original targets.

These results collectively demonstrate that the improved YOLO-CFDU model achieves higher accuracy and stronger robustness across a wide range of complex urban garbage detection scenarios.

(a) original image (b) YOLOv11n output (c) YOLO-CFDU output

Fig. 15. Comparison of detection visualization results of the two models in the first scenario.

(a) original image (b) YOLOv11n output (c) YOLO-CFDU output

Fig. 16. Comparison of detection visualization results of the two models in the second scenario.

(a) original image (b) YOLOv11n output (c) YOLO-CFDU output

Fig. 17. Comparison of detection visualization results of the two models in the third scenario.

(a) original image (b) YOLOv11n output (c) YOLO-CFDU output

Fig. 18. Comparison of detection visualization results of the two models in the fourth scenario.

6) Real-vehicle Experiment

To verify the detection performance of the improved YOLO-CFDU model in real road scenarios, we conduct real-vehicle testing experiments based on the AKM ROS smart car, which is equipped with the Astra camera, embedded computing unit (Jetson Xavier NX), board computer and other equipment required for this experiment, The experimental vehicle is shown in Fig. 19.

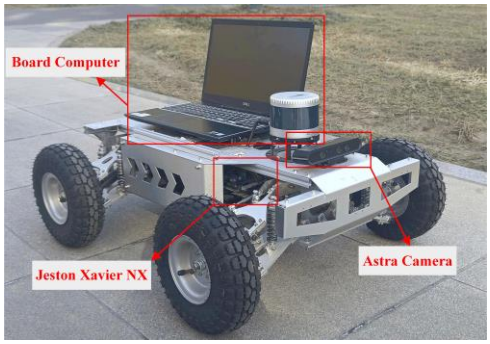


Fig. 19. AKM ROS smart car

The experiment was conducted on a 200-meter secondary road and sidewalk in a city, with randomly distributed cigarette butts, plastic bags, plastic bottles, and other litter. The ROS smart car traveled the route at a constant speed of 1 m/s, capturing road images with an Astra camera. These images were processed by an embedded unit for garbage detection, with results transmitted to an onboard computer for storage and real-time display. Experimental results demonstrate that YOLO-CFDU delivers excellent detection performance in real road environments, achieving a stable inference speed of 43.6 FPS, which meets real-time requirements. As shown in Fig. 20, the model effectively detects targets of varying sizes in mixed scenes and maintains high accuracy even under complex conditions such as dense garbage piles, uneven road textures, and lighting interference, with no missed or false detections observed.

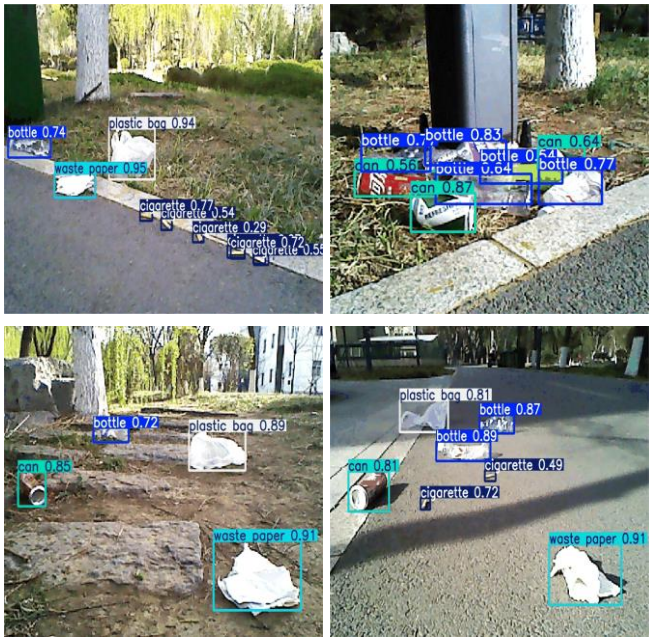


Fig. 20. Vehicle testing effect

IV. Conclusion

To address the challenges of variable object morphology size and complex background interference in urban road garbage detection, this paper proposes an improved model YOLO-CFDU based on YOLOv11. The C3K2-FE module is designed to replace the C3K2 module, incorporating Fasterblock to reduce redundant computations while integrating the EMA attention mechanism to enhance multi-scale feature extraction. Additionally, the DSWF module is introduced to dynamically allocate spatial weights and reinforce target feature representation. The U-IoU loss function is adopted to optimize bounding box regression accuracy and balance gradient contributions from high- and low-quality samples. To meet lightweight deployment requirements, the LAMP pruning technique is used to compress model parameters and computational overhead. Experimental results on the self-constructed urban road garbage dataset demonstrate that the improved model achieves mAP@0.5 of 0.735, outperforming the baseline by 4.6%. Meanwhile, model parameters and computations are reduced by 44.0% and 29.7%, respectively, significantly lowering computational overhead while maintaining detection accuracy. This study provides an efficient and cost-effective garbage detection solution for urban sanitation management, offering technical support for advancing urban environmental governance.

References

[1] W. Ma, H. Chen, W. Zhang, H. Huang, J. Wu, X. Peng, and Q. Sun, "DSYOLO-trash: An attention mechanism-integrated and object tracking algorithm for solid waste detection," Waste Management, vol. 178, pp. 46-56, 2024.

[2] Y. Chen, A. Luo, M. Cheng, Y. Wu, J. Zhu, Y. Meng, and W. Tan, "Classification and recycling of recyclable garbage based on deep learning," Journal of Cleaner Production, vol. 414, p 137558, 2023.

[3] C. Ou, C. Yang, Q. Zeng, and G. Tan, "Research on garbage recognition of intelligent sweeper vehicle based on improved PSPNet algorithm," SAE Technical Paper, no. 2022-01-0220, 2022.

[4] S. Jin, Z. Yang, G. Królczykg, X. Liu, P. Gardoni, and Z. Li, "Garbage detection and classification using a new deep learning-based machine vision system as a tool for sustainable waste recycling," Waste Management, vol. 162, pp. 123-130, 2023.

[5] T. Yuan, J. Lin, K. Hu, W. Chen, and Y. Hu, "Domain adaptive urban garbage detection based on attention and confidence fusion," Information, vol. 15, no. 11, p. 699, 2024.

[6] M. Arulmozhi, N. G. Iyer, S. J. Sophia, P. Sivakumar, C. Amutha, and D. Sivamani, "Comparison of YOLO and Faster R-CNN on garbage detection," Optimization Techniques in Engineering: Advances and Applications, vol. pp. 37-49, 2023.

[7] M. Karthikeyan, T. S. Subashini, and R. Jebakumar, "SSD based waste separation in smart garbage using augmented clustering NMS," Automated Software Engineering, vol. 28, no. 2, p. 17, 2021.

[8] X. Cai, F. Shuang, X. Sun, Y. Duan, and G. Cheng, "Towards lightweight neural networks for garbage object detection," Sensors, vol. 22, no. 19, p. 7455, 2022.

[9] S. Liu, R. Chen, M. Ye, J. Luo, D. Yang, and M. Dai, "EcoDetect-YOLO: A lightweight, high-generalization methodology for real-time detection of domestic waste exposure in intricate environmental landscapes," Sensors, vol. 24, no. 14, p. 4666, 2024.

[10] W. Feng, W. Yan, and J. Xie, "Garbage object detection method based on improved Faster R-CNN," Proceedings of the 5th International Conference on Computer Science and Software Engineering, pp. 415–419, 2022.

[11] L. Tan, H. Wu, Z. Xu, and J Xia, "Multi-object garbage image detection algorithm based on SP-SSD," Expert Systems with Applications, vol. 263, p. 125773, 2025.

[12] J. He, Y. Cheng, W. Wang, Y. Gu, Y. Wang, W. Zhang, A. Shankar, S. Selvarajan, and S. A. P. Kumar, "EC-YOLOX: A deep-learning algorithm for floating objects detection in ground images of complex water environments," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 17, pp. 7359–7370, 2024.

[13] E. Z. Kuang, K. R. Bhandari, and J. Gao, "Optimizing waste management with advanced object detection for garbage classification," arXiv:2410.09975, 2024.

[14] N. Li, M. Wang, G. Yang, B. Li, B. Yuan, and S. Xu, "DENS-YOLOv6: A small object detection model for garbage detection on water surface," Multimedia Tools Appl., vol. 83, no. 18, pp. 55751-55771, 2024.

[15] Z. Xia, H. Zhou, H. Yu, H. Hu, G. Zhang, J. Hu, and T. He, "YOLO-MTG: A lightweight YOLO model for multi-target garbage detection," Signal Image Video Process., vol. 18, no. 6, pp. 5121-5136, 2024.

[16] X. Bai, H. Shi, Z. Wang, H. Hu, and C. Zhang, "Study of YOLOv7 model lightweighting based on group-level pruning," IEEE Access, vol. 12, pp. 12345-12356, 2024.

[17] L. Jiang, F. Liu, J. Lv, B. Liu, and C. Wang, "GST-YOLO: A lightweight visual detection algorithm for underwater garbage detection," Journal of Real-Time Image Process., vol. 21, no. 4, p. 114, 2024.

[18] R. Khanam, M. Hussain, "Yolov11: An overview of the key architectural enhancements," arXiv:2410.17725, 2024.

[19] J. Chen, S. Kao, H. He, W. Zhou, and S. Wen, "Run, don't walk: chasing higher FLOPS for faster neural networks," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12021–12031, 2023.

[20] D. Ouyang, S. He, G. Zhang, M. Luo, H. Guo, J. Zhan, and Z. Huang, "Efficient multi-scale attention module with cross-spatial learning,"

Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5, 2023.
[21] J. Qin, C. Wang, X. Ran, S. Yang, and B. Chen, "A robust framework combined saliency detection and image recognition for garbage classification," Waste Management, vol. 140, pp. 193-203, 2022.
[22] X. Luo, Z. Cai, B. Shao, and Y. Wang, "Unified-IoU: For high-quality object detection," arXiv:2408.06636, 2024.
[23] J. Lee, S. Park, S. Mo, S. Ahn, and J. Shin, "Layer-adaptive sparsity for the magnitude-based pruning," arXiv:2010.07611, 2020.