Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

MSF-TrashDet: A Lightweight Multi-Scale Feature Fusion Network for Automatic Waste Object Detection Using the TrashNet Dataset

Tao Yanfen
Jiujiang Vocational University
Lei Lijuan
Jiujiang Vocational University
Gao Yousheng
2732377751@qq.com

Jiujiang Vocational University
Fan Dongdong
Jiujiang Vocational University
Mei Pengfei
Jiujiang Vocational University

Research Article

Keywords: waste detection, multi-scale feature fusion, lightweight convolutional neural network, model pruning, object detection, TrashNet

Posted Date: July 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10399139/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

MSF-TrashDet: A Lightweight Multi-Scale Feature Fusion Network for Automatic Waste Object Detection Using the TrashNet Dataset

Yanfen Tao[1], Lijuan Lei[1], Yousheng Gao[1*], Dongdong Fan[1], Pengfei Mei[1]

[1]Jiujiang Vocational University, Jiangxi, China.

*Corresponding author(s). E-mail(s): 273237751@qq.com;
Contributing authors: 30770943@qq.com; 8427110@qq.com;
214982942@qq.com; 1491853121@qq.com;

Abstract

With the acceleration of urbanization and industrialization, the yield of municipal solid waste has increased dramatically. Automatic waste sorting based on computer vision and deep learning has become an effective solution for environmental protection and resource recycling. However, mainstream object detection models face challenges such as insufficient multi-scale feature extraction, weak robustness to cluttered backgrounds, excessive model parameters, and high computational complexity when applied to real-world waste detection scenarios. To address these issues, this paper proposes MSF-TrashDet, a lightweight and efficient object detection network tailored for waste detection tasks. The proposed model introduces a novel Multi-Scale Feature Enhancement (MSFE) block, which integrates dynamic weight adaptation, dilated convolution, and channel-split residual connections to strengthen feature representation while maintaining computational efficiency. In addition, structured channel pruning is applied to further compress the model for practical deployment on edge devices. Extensive experiments are conducted on the public TrashNet dataset to evaluate the effectiveness of the proposed method. The experimental results demonstrate that MSF-TrashDet achieves 91.91% mAP50 with only 1.93 M parameters and 5.00 GFLOPs, which outperforms several state-of-the-art lightweight detectors including YOLOv5n, YOLOv8n, and YOLOv11n in terms of accuracy-efficiency trade-off. Ablation experiments verify the effectiveness of each key component in the proposed network. The MSF-TrashDet model provides a practical, reliable, and efficient solution for automatic waste detection and intelligent recycling systems.

Keywords: waste detection, multi-scale feature fusion, lightweight convolutional neural network, model pruning; object detection, TrashNet

1 INTRODUCTION

With the rapid expansion of urban population and industrial manufacturing activities, the output of municipal solid waste keeps rising year by year, which brings severe pressure to urban ecological governance and daily waste management worldwide [1]. Traditional manual waste classification relies heavily on human labor, which suffers from low sorting efficiency, high labor cost, and potential occupational health risks for workers. Therefore, developing automated, accurate and low-cost waste detection technology has become an urgent demand to promote resource recycling and build eco-friendly smart cities [2].

Benefiting from the rapid progress of deep learning and computer vision, Convolutional Neural Networks (CNN) [3] have been widely adopted in various visual tasks including image classification, semantic segmentation and generic object detection [4]. Existing detection frameworks can be roughly divided into two-stage and one-stage architectures. Two-stage detectors such as Faster R-CNN [5] achieve high detection precision yet carry heavy computation burden, while one-stage algorithms represented by the YOLO series strike a better balance between inference speed and accuracy and are more suitable for real industrial deployment.

Nevertheless, directly applying mainstream off-the-shelf detectors to garbage recognition still faces many practical obstacles in real scenarios[6]. Garbage targets exhibit drastic differences in scale, texture and placement angle, and the detection process is easily interfered by cluttered backgrounds, uneven illumination and target occlusion[7]. Most lightweight YOLO models simplify network layers to reduce computation, which inevitably weakens their ability to extract multi-scale fine features; high-precision networks with complete feature extraction modules often contain massive parameters and computing overhead, and cannot be deployed on embedded edge devices with limited computing power. To solve the above practical contradictions, this work designs a lightweight multi-scale feature fusion detection network named MSF-TrashDet oriented to automatic waste identification tasks.

The core innovation of the proposed method lies in a newly designed Multi-Scale Feature Enhancement block, which combines dilated convolution branches and dynamic weight assignment to capture multi-scale garbage features adaptively, and introduces channel-split residual structures to retain shallow detail information while cutting redundant calculation. On this basis, structured channel pruning technology is further adopted to compress the overall model volume for edge terminal deployment. A full set of comparative and ablation experiments based on the public TrashNet dataset verifies that the proposed MSF-TrashDet can achieve outstanding detection accuracy with fewer parameters and lower GFLOPs compared with mainstream lightweight detectors, providing a feasible technical scheme for intelligent waste sorting and recycling monitoring systems. The rest of this paper is organized as follows. Section 2 reviews relevant research on lightweight networks and waste detection algorithms. Section 3 elaborates the overall structure and each core module of MSF-TrashDet in detail. Section 4 presents experimental settings and multi-dimensional experimental results. Section 5 analyzes the performance of the proposed method. Finally, Section 6 summarizes this work and puts forward directions for follow-up optimization.

2 RELATED WORK

Intelligent waste detection utilizing deep learning and computer vision has emerged as a pivotal technology for advancing urban environmental management and resource recycling[8]. In recent years, the YOLO series of models have been extensively employed in waste detection tasks, owing to their high efficiency and end-to-end processing capabilities[9]. Numerous enhanced algorithms have been developed to address the complex characteristics inherent in waste targets.

Real-time intelligent waste monitoring systems built upon YOLOv5 and YOLOv8 have been deployed on embedded hardware platforms, which can realize recognition of filled trash cans, scattered waste, plastic bags and wet garbage. Comparative tests on public TrashNet dataset and self-built garbage data verify that YOLOv5 offers better adaptability on GPU-free embedded devices, while YOLOv8 achieves higher detection precision [10]. Restricted by insufficient scene diversity of training samples, such methods still suffer from limited generalization under uneven illumination and partial occlusion.

To improve detection robustness in cluttered urban environments, a lightweight YOLOv5s-based framework called EcoDetect-YOLO was proposed for exposed domestic waste detection [11]. By embedding the CBAM attention module, adding a P2 small-target detection head, and replacing the feature fusion structure with BiFPN, the model significantly strengthened multi-scale feature representation and small-object detection ability. Although it achieves strong anti-interference performance and

robustness against noise, occasional misjudgments still occur on similar-appearance waste categories.

Further pursuing efficiency and deployability, an improved lightweight ABCS-YOLO detector was constructed by introducing adaptive spatial feature fusion, bidirectional feature pyramid with attention mechanism, and ShuffleNetV2 lightweight modules [12]. This design effectively enhances multi-scale feature fusion while compressing model parameters and computation, making it more suitable for edge devices. Validated on real-world detection tasks, the method achieves a favorable balance between accuracy and speed, yet remains less effective on ultra-small and irregularly shaped waste targets.

Currently, research in garbage detection predominantly concentrates on the design of lightweight networks suitable for edge deployment, multi-scale feature learning to accommodate targets with substantial variations in size and shape, and robust feature extraction under complex environmental conditions such as variable lighting, occlusion, and clutter. Concurrently, the integration of detection systems with IoT monitoring and intelligent recycling equipment has emerged as a significant direction for future development.

Despite significant advancements in current methodologies, the detection of waste materials continues to encounter numerous challenges in practical applications. The substantial variability in the scale, shape, and texture of waste items often results in the omission of small and fragmented debris. Furthermore, complex real-world conditions—such as uneven lighting, cluttered backgrounds, and occlusion of targets—adversely impact the robustness and reliability of detection outcomes. High-precision models frequently pose deployment difficulties on edge devices due to their considerable size, whereas lightweight models typically experience a reduction in accuracy. Additionally, limitations in public waste datasets, including inadequate scene diversity and imbalanced sample distributions, constrain the model's ability to generalize effectively across different environments. Most existing systems primarily support real-time monitoring but lack predictive capabilities regarding waste accumulation trends, thereby hindering the intelligent optimization of waste collection routes and management strategies.

3 METHOD

This section provides a detailed exposition of the proposed YOLO11n-MSFE detector, specifically designed for real-time waste classification on devices with limited computational resources. As illustrated in Fig. 1, the framework builds upon the compact architecture of YOLO11n and incorporates three principal enhancements: (1) optimization of the lightweight backbone through the integration of ShuffleNetV2 [13], (2) multi-scale feature augmentation enabled by the novel MSFE block, and (3) adaptive scale-consistent detection facilitated by the ASFF head. Collectively, these modifications enhance multi-scale feature representation, mitigate background noise, reduce occlusion effects, and decrease computational complexity, thereby establishing an efficient and robust system for waste detection.

3.1 Overall Architecture

The comprehensive architecture of the proposed MSF-TrashDet, illustrated in Fig. 2, is constructed upon the lightweight YOLO11n baseline, which is selected as the fundamental framework due to its native balance between inference latency and basic detection performance for edge-side vision tasks. To address the inherent drawbacks of vanilla YOLO11n when applied to complex waste detection scenarios, including insufficient multi-scale feature capture ability and relatively high computational cost, targeted optimizations are implemented across backbone, neck and detection head branches. The backbone network incorporates ShuffleNetV2 units in lieu of the original high-complexity convolution components; by introducing channel split, depthwise convolution and channel shuffle lightweight operations, the backbone largely eliminates redundant matrix multiplication during feature extraction, resulting in a substantial reduction in both parameter count and computational overhead while retaining stable feature extraction capacity for garbage targets with diverse textures and

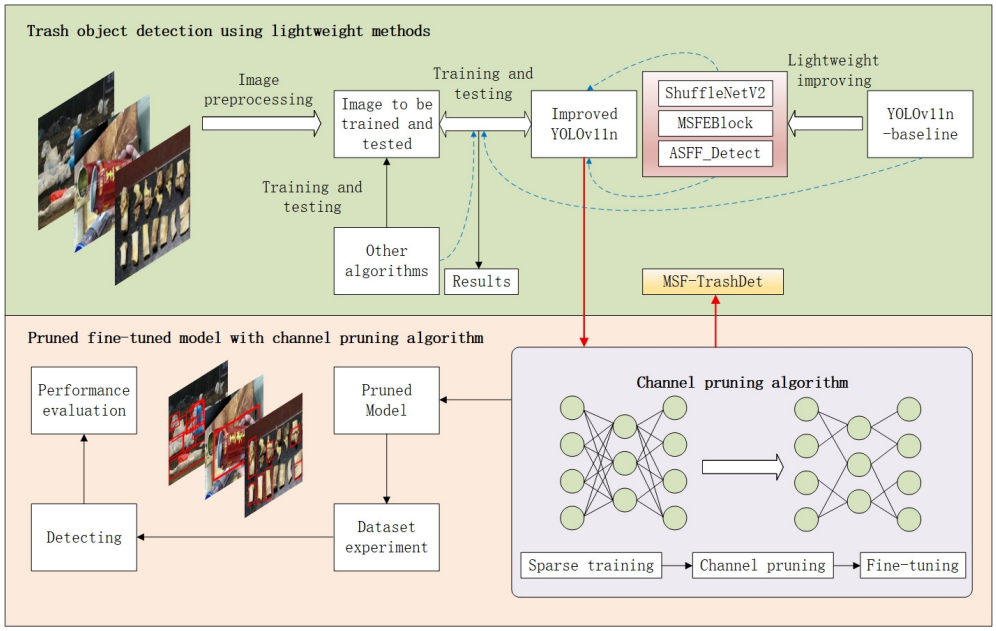


Fig. 1 Overall framework of the proposed MSF-TrashDet detection network.

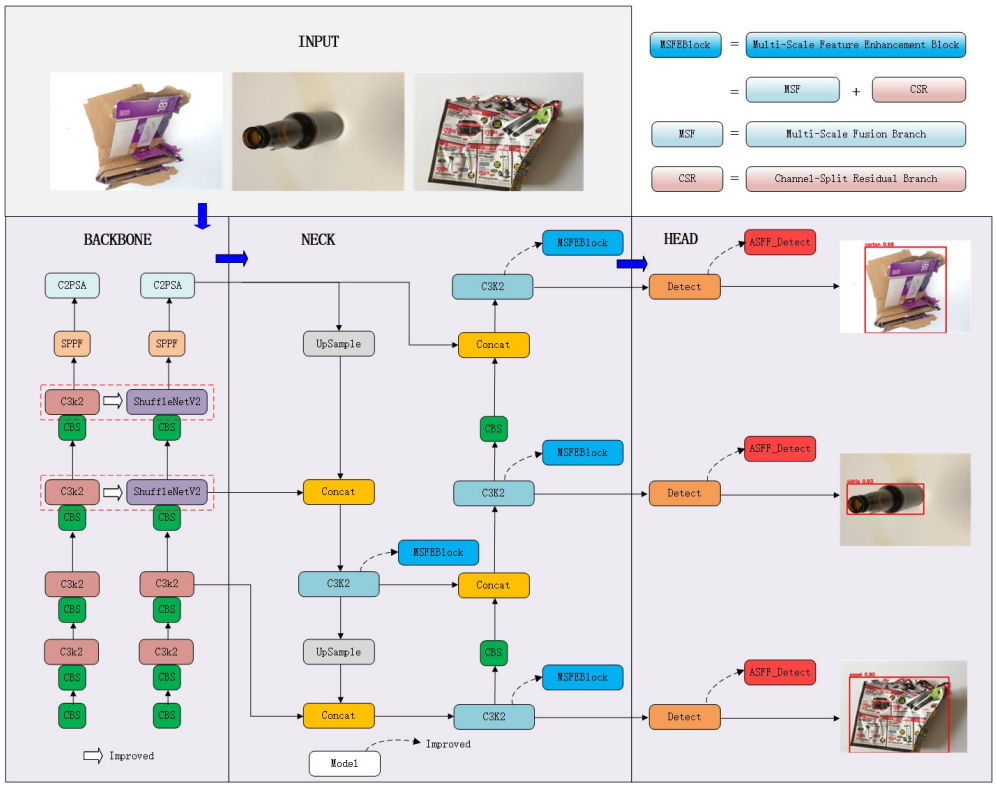


Fig. 2 Detailed network architecture of MSF-TrashDet model.

irregular shapes. Within the neck section, the customized multi-scale feature enhancement (MSFE) block is embedded as the core feature fusion module, which leverages multi-dilation convolution branches and dynamic weight assignment to strengthen multi-scale feature representation and effectively mitigate the interference brought by cluttered backgrounds, uneven illumination and frequent target occlusion in actual urban waste sorting scenes.

Moreover, the original naive detection head relying on simple feature concatenation is fully substituted with the Adaptive Spatial Feature Fusion (ASFF)[14] module

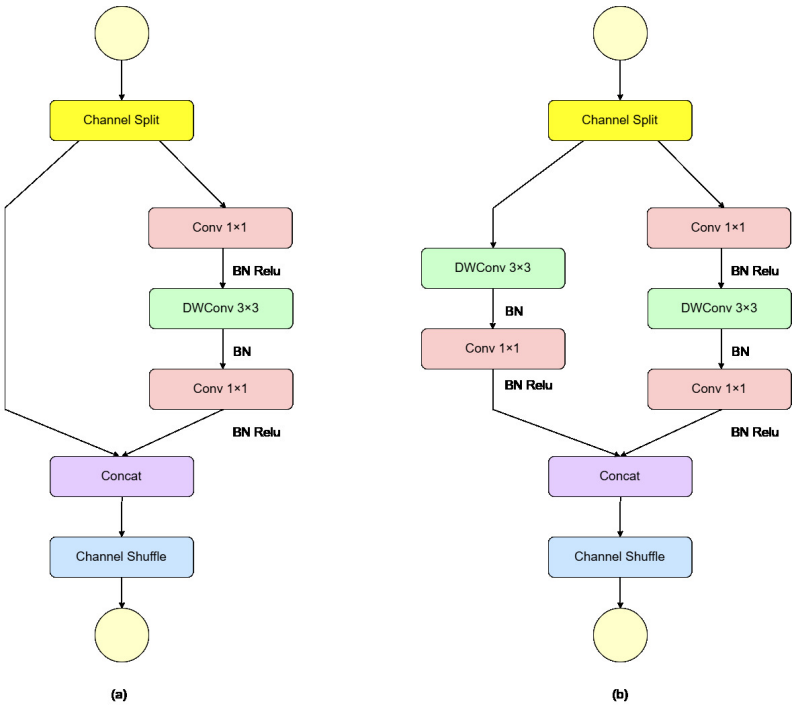


Fig. 3 Structural diagram of the ShuffleNetV2 block.

to facilitate adaptive cross-scale feature integration. Different from conventional feature pyramid networks that directly splice multi-scale feature maps and ignore spatial scale conflicts, ASFF dynamically learns independent spatial weighting coefficients for each feature layer at every pixel position, which reasonably suppresses invalid cross-scale feature noise and alleviates the scale mismatch issue thoroughly. Benefiting from this adaptive fusion mechanism, the model significantly improves the detection accuracy of tiny fragmental waste, partially occluded trash and densely stacked garbage targets that are prone to missed detection in original lightweight detectors. The ultimate detection layer generates object bounding box and category predictions at three distinct output scales, corresponding to feature maps downsampled by factors of 8×, 16×, and 32×. The three-scale prediction branches work collaboratively to capture small, medium and large waste objects respectively, enabling comprehensive and robust recognition of garbage items spanning a wide range of size variations in real-world monitoring environments.

3.2 Lightweight Backbone Based on ShuffleNetV2

To facilitate efficient deployment on resource-limited edge devices, ShuffleNetV2 is employed to optimize the backbone architecture of YOLO11n. ShuffleNetV2, as shown in Fig. 3, is a well-established lightweight network design that leverages channel splitting, depthwise convolution, and channel shuffle operations to enable effective feature extraction with minimal computational cost[15].

In the proposed framework, the original computationally intensive modules located in the intermediate and higher layers of the backbone are substituted with ShuffleNetV2 units. This alteration preserves the network's feature extraction capabilities while reducing the overall model size. The adoption of this lightweight backbone ensures the model's real-time performance in practical garbage detection scenarios, thereby providing a robust foundation for subsequent multi-scale feature enhancement and adaptive detection processes.

3.3 Multi-Scale Feature Enhancement Block (MSFE block)

To address the weak feature extraction capacity of lightweight detectors for multi-scale, irregular, and occluded waste targets under complex background interference,

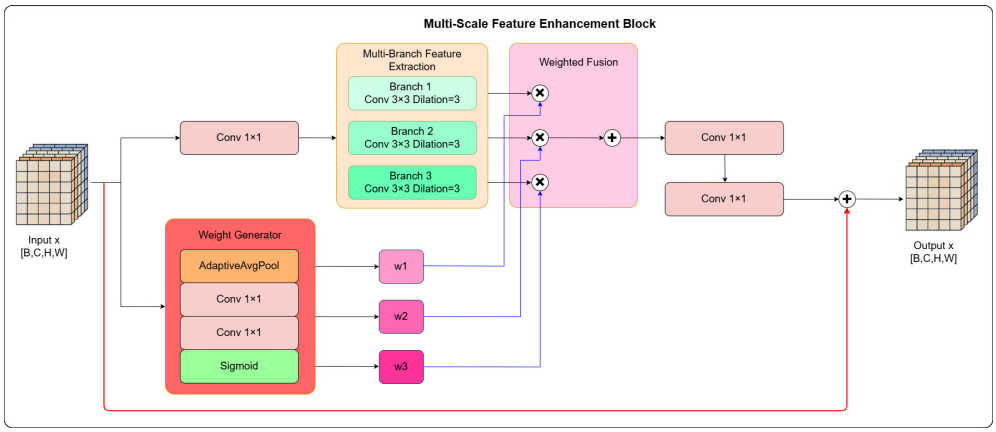


Fig. 4 Structural diagram of the Multi-Scale Feature enhancement block.

this paper designs a dedicated Multi-Scale Feature Enhancement (MSFE) block tailored for garbage detection, whose complete structure is visualized in Fig. 4. The core innovations of the proposed module lie in three complementary designs: multi-dilation parallel branches for multi-scale context capture, a global context-driven dynamic weight generator for adaptive feature fusion, and a channel-split residual branch to retain fine-grained shallow texture information. The whole module maintains low computational overhead and exhibits good compatibility with subsequent structured channel pruning.

Let the input feature tensor be $\boldsymbol{X} \in \mathbb{R}^{B\times C_{in}\times H\times W}$, where B, C_{in}, H, W separately denote batch size, input channel count, feature height and feature width. The module first executes a channel-split strategy to partition the output channel budget C_{out} into two independent parts for multi-scale processing and residual information preservation:

$$C_{mid} = \frac{C_{out}}{2}, \quad C_{res} = C_{out} - C_{mid}. \tag{1}$$

Here, C_{mid} serves as the channel dimension for multi-branch dilated convolution, while C_{res} reserves original shallow feature channels as a residual branch to avoid the loss of tiny fragmented waste details caused by repeated convolution downsampling. A 1×1 pointwise convolution compresses the input feature dimension to reduce computation of subsequent multi-scale branches:

$$\boldsymbol{X}_{red} = \mathrm{Conv}_{1\times 1}(\boldsymbol{X}). \tag{2}$$

Three parallel convolution branches with distinct dilation rates $d = \{1, 2, 3\}$ are constructed on top of $\boldsymbol{X}_{red}$ to capture local texture, medium-range contour and global large-scale context of garbage targets respectively. Each branch stacks two standard convolution layers to refine extracted multi-scale features:

$$\begin{aligned} \boldsymbol{F}_1 &= \mathrm{ConvStack}_{d=1}(\boldsymbol{X}_{red}), \\ \boldsymbol{F}_2 &= \mathrm{ConvStack}_{d=2}(\boldsymbol{X}_{red}), \\ \boldsymbol{F}_3 &= \mathrm{ConvStack}_{d=3}(\boldsymbol{X}_{red}). \end{aligned} \tag{3}$$

Instead of fixed equal fusion weights adopted in conventional multi-branch modules, this work builds a lightweight weight generator to produce data-adaptive fusion coefficients from global feature statistics. The pipeline integrates global average pooling, two stacked 1×1 convolutions and Sigmoid activation to output a three-channel weight vector $\boldsymbol{W} = [w_1, w_2, w_3] \in \mathbb{R}^{B\times 3\times 1\times 1}$:

$$\boldsymbol{W} = \sigma\left(\mathrm{Conv}_{1\times 1}\left(\mathrm{Conv}_{1\times 1}\left(\mathrm{AvgPool}(\boldsymbol{X})\right)\right)\right), \tag{4}$$

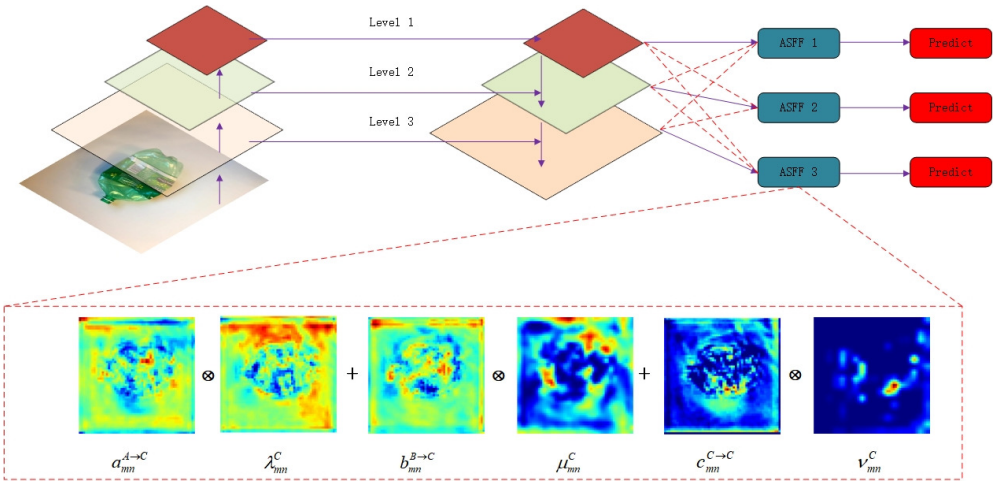


Fig. 5 Structural diagram of the ASFF block, including multi-level feature delivery pipeline and pixel-level weighted fusion calculation.

where $\sigma(\cdot)$ represents the Sigmoid function to constrain each weight $w_i \in (0,1)$. The three multi-scale feature maps are fused via element-wise weighted summation:

$$\boldsymbol{F}_{fusion} = w_1 \odot \boldsymbol{F}_1 + w_2 \odot \boldsymbol{F}_2 + w_3 \odot \boldsymbol{F}_3, \tag{5}$$

with $\odot$ denoting element-wise Hadamard multiplication. The dynamic weighting mechanism enables the model to automatically assign larger weights to branches matching the actual scale of garbage targets and suppress irrelevant scale features disturbed by clutter or occlusion.

For the channel-split residual branch, the module slices the first C_{res} channels of original input $\boldsymbol{X}$ as residual feature $\boldsymbol{X}_{res}$. Bilinear interpolation is adopted to align spatial resolution between $\boldsymbol{X}_{res}$ and $\boldsymbol{F}_{fusion}$ when mismatched. The residual branch and fused multi-scale features are concatenated along the channel dimension, followed by two successive 1×1 convolution layers to aggregate cross-information and project to the final output dimension C_{out}:

$$\boldsymbol{Y} = \mathrm{Conv}_{1\times1}\left(\mathrm{Conv}_{1\times1}\left(\mathrm{Concat}(\boldsymbol{F}_{fusion}, \boldsymbol{X}_{res})\right)\right). \tag{6}$$

If the input and output channel dimensions are identical, an additional shortcut residual connection is appended between $\boldsymbol{X}$ and $\boldsymbol{Y}$ to stabilize network training.

Benefiting from the joint effect of dilated multi-branch extraction, context-aware dynamic weight fusion and channel-split residual retention, the MSFE block significantly strengthens the model's discriminative perception of tiny, large and irregular waste objects. Meanwhile, the unified channel design of the module supports sparse training and channel pruning, which facilitates the subsequent lightweight compression of the whole detection network for edge device deployment.

3.4 ASFF-Based Detection Head

To enhance the integration of cross-scale features and mitigate the information loss and feature conflict brought by simple direct concatenation in traditional feature pyramids, the ASFF module is introduced and deployed on all three output branches of MSF-TrashDet. As illustrated in Fig. 5, the network extracts three multi-scale feature maps denoted as Level 1, Level 2 and Level 3 from the feature pyramid. For each target prediction branch, ASFF first aligns feature maps from all three levels to an identical spatial resolution, then learns pixel-wise adaptive weights to perform weighted feature fusion at every spatial coordinate (m, n), which suppresses invalid cross-scale feature interference according to the contribution of each scale feature[16].

Taking the fusion process for the Level 3 output branch as an example, let λ_{mn}^{C}, μ_{mn}^{C}, ν_{mn}^{C} represent the aligned feature values from Level 1, Level 2 and Level 3 at position (m, n) respectively. The learnable weight coefficients corresponding to three

feature sources are defined as $a_{mn}^{A\to C}$, $b_{mn}^{B\to C}$, $c_{mn}^{C\to C}$. These weight coefficients are normalized to satisfy the sum constraint:

$$a_{mn}^{A\to C} + b_{mn}^{B\to C} + c_{mn}^{C\to C} = 1 \tag{7}$$

The final fused feature $\mathcal{F}_{mn}^{C}$ at spatial coordinate (m, n) is obtained by element-wise multiplication and summation as follows:

$$\mathcal{F}_{mn}^{C} = a_{mn}^{A\to C} \otimes \lambda_{mn}^{C} + b_{mn}^{B\to C} \otimes \mu_{mn}^{C} + c_{mn}^{C\to C} \otimes \nu_{mn}^{C} \tag{8}$$

where $\otimes$ denotes the element-wise Hadamard product operation shown in the heatmap subgraph of Fig. 5. All weight terms $a_{mn}^{A\to C}$, $b_{mn}^{B\to C}$ and $c_{mn}^{C\to C}$ are generated by stacked convolution layers with Sigmoid activation, which enables the network to automatically assign larger weights to feature layers matching the real target scale at each pixel.

This pixel-adaptive weighted fusion mechanism fundamentally resolves scale mismatch and spatial inconsistency defects existing in conventional fixed-concat feature pyramid networks. By embedding the ASFF fusion unit into all three detection layers corresponding to Level 1, Level 2 and Level 3, the model can dynamically filter out irrelevant feature responses from mismatched scales. The design achieves prominent accuracy improvement for tiny fragmented waste, partially occluded trash and densely stacked garbage samples, and greatly strengthens the model's anti-interference robustness under complex real waste sorting scenes with cluttered backgrounds and uneven illumination.

3.5 Structured Channel Pruning for Model Compression

After completing the network construction of MSF-TrashDet, structured channel pruning is introduced as a post-processing compression strategy to further reduce model overhead for edge deployment. The pruning pipeline includes sparse training, channel pruning and fine-tuning. During sparse training, the Batch Normalization scaling factors of each convolutional channel are constrained to approach zero, so channels with lower scaling factors are identified as redundant ones that contribute little to feature extraction.

In the pruning stage, redundant channels with small BN scaling factors are directly removed to realize structured network compression without irregular weight sparsity, which guarantees hardware acceleration compatibility. After pruning, fine-tuning is conducted to recover slight detection performance degradation caused by channel deletion. Multiple pruning ratios are tested to balance accuracy and model volume.

4 EXPERIMENTS

4.1 Experimental Setup

All experiments were performed on a desktop workstation configured with an Intel Core i7-12700 CPU, an NVIDIA GeForce RTX 3090 GPU featuring 24 GB of memory, and 64 GB of DDR4 RAM operating at 3200 MHz. The system ran Windows 11 as the operating system. The deep learning framework utilized was PyTorch version 1.13.1, integrated with CUDA 11.7 and cuDNN 8.5.

The TrashNet[17] dataset served as the basis for all experimental evaluations. This dataset was partitioned randomly into training, validation, and test subsets following a 70:20:10 split ratio. All input images were uniformly resized to 640×640 pixels. To enhance model generalization, data augmentation techniques—including random horizontal flipping, random scaling, and color jittering—were applied during the training phase.

Model optimization was conducted using the AdamW optimizer, initialized with a learning rate of 0.001, a weight decay parameter of 0.0005, and a batch size of 16. The training regimen spanned 100 epochs.

Performance was assessed using multiple evaluation metrics, comprising Precision[18], Recall[19], F1-score[20], mean Average Precision at 50% IoU (mAP50)[21], the number of model parameters (in millions), model size (in

Table 1 Ablation study of key components on the TrashNet dataset.

Module			Precision	mAP50	Recall	F1	Size (MB)	GFLOPs
MSFE	ShuffleNetV2	ASFF						
-	-	-	83.09	89.27	82.05	82.56	11.4	6.50
✓	-	-	86.77	89.92	81.89	84.26	10.8	6.21
✓	✓	-	87.86	89.59	**82.42**	85.05	**10.2**	**6.07**
✓	✓	✓	**91.80**	**90.79**	80.72	**85.90**	12.1	7.30

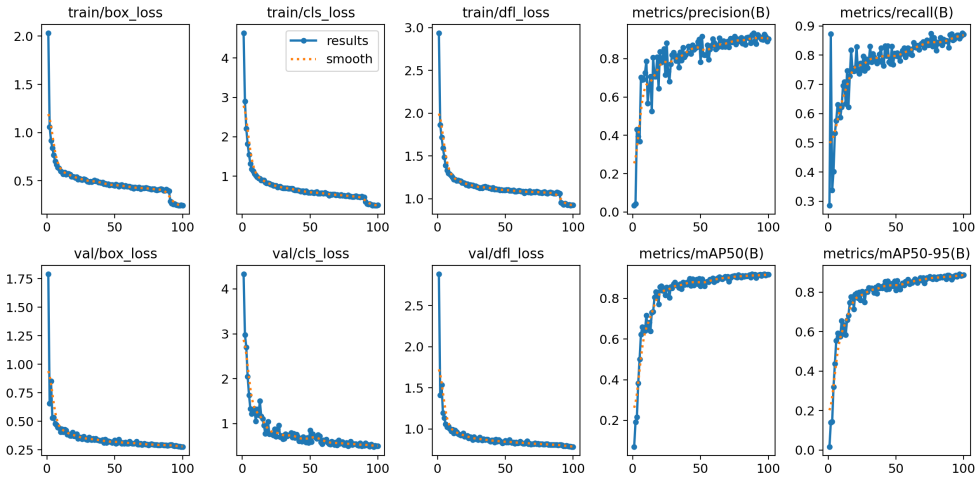


Fig. 6 Loss and performance metric curves during model training.

megabytes), and computational complexity measured in giga floating-point operations (GFLOPs)[22].

4.2 Ablation Study Results

To verify the effectiveness of each key component in MSF-TrashDet, ablation experiments are performed on the TrashNet dataset. The baseline is YOLOv11n. Three core components are evaluated: the MSFE block, ShuffleNetV2 backbone, and ASFF module. In the table, "√" denotes that the corresponding component is enabled, while "-" denotes disabled. The ablation results are listed in Table 1.

The baseline model attains a precision of 83.09% and an mAP50 of 89.27%, with a model size of 11.4 MB. Incorporating the MSFE block results in a notable enhancement in detection accuracy, accompanied by a marginal reduction in both model size and computational complexity. The adoption of the ShuffleNetV2 backbone further streamlines the model architecture while preserving stable detection performance. When all three modules are integrated concurrently, the network achieves the highest overall detection accuracy, albeit with a moderate increase in model size and GFLOPs. This observation suggests that the proposed modules effectively improve feature representation, and the slight augmentation in model complexity is justified by the significant gains in accuracy. Figure 6 presents the training curves of loss values and key performance metrics over epochs, demonstrating the convergence behavior and performance improvements of MSF-TrashDet.

4.3 Model Compression Results

To further optimize the lightweight performance and facilitate edge device deployment, structured channel pruning is adopted. Multiple pruning rates from 0% to 90% are set for comparative analysis. The detailed results including model size are presented in Table 2.

As the pruning rate increases, both the number of parameters and the model size exhibit a consistent decline. At a pruning rate of 60%, the model attains an

Table 2 Model compression results under different pruning rates.

Prune Rate (%)	Precision	mAP50	Recall	F1	Params (M)	Size (MB)	GFLOPs
0	91.80	90.79	80.72	85.90	2.67	5.44	7.30
10	92.73	92.37	85.79	89.12	2.52	5.14	6.80
20	**93.30**	**92.66**	86.09	**89.55**	2.39	4.89	6.50
30	92.51	91.98	84.85	88.52	2.25	4.63	6.00
40	88.81	91.19	87.02	87.91	2.14	4.42	5.70
50	92.73	91.62	85.52	88.98	2.04	4.22	5.40
60	92.78	91.91	85.65	89.07	1.93	4.02	5.00
70	86.89	90.61	87.01	86.80	1.85	3.86	4.80
80	31.67	43.40	81.96	45.69	1.76	3.69	4.50
90	14.03	18.39	**98.86**	24.48	**1.71**	**3.59**	**4.30**

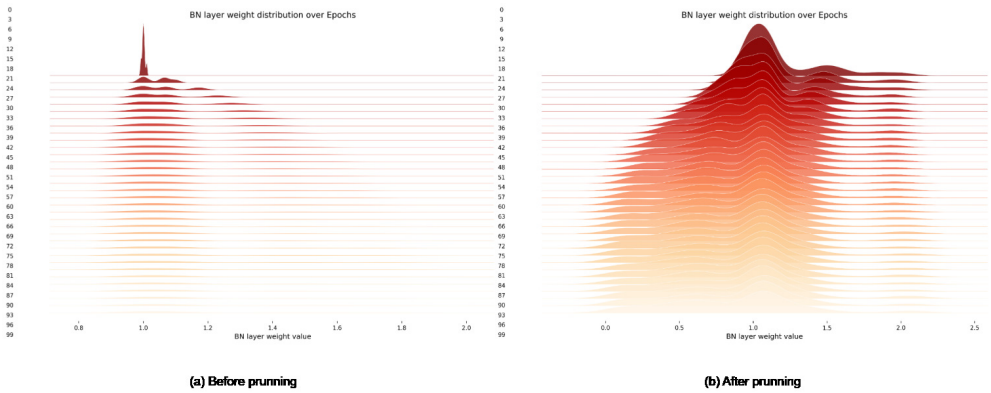


Fig. 7 Comparison of model weight distribution before and after structured pruning.

optimal balance among accuracy, parameter count, model size, and computational complexity, containing only 1.93 million parameters and occupying 8.7 MB of storage. Beyond this threshold, although the model size undergoes further reduction, there is a marked decrease in detection precision and mAP50, indicating that excessive pruning compromises essential feature channels and results in significant performance deterioration. Consequently, a pruning rate of 60% is identified as the most effective ratio for practical implementation.

Figure 7 visualizes the BatchNorm layer weight distribution of the MSF-TrashDet model before and after structured channel pruning. Before pruning, the weight values are widely dispersed, which indicates abundant redundant convolutional channels inside the network. After applying a 60% structured pruning rate, redundant feature channels are eliminated according to the scaling factors of BatchNorm layers, making the weight distribution much more concentrated. After pruning, the model parameter volume is compressed from 2.67 M to only 1.93 M, and computational complexity (GFLOPs) decreases from 7.30 to 5.00. Meanwhile, the model retains competitive detection performance, achieving 92.78% Precision and 91.91% mAP50 on the TrashNet dataset. The above results fully validate that the structured channel pruning strategy proposed in this paper can well balance model lightweight degree and detection robustness, which is highly suitable for deployment on resource-limited edge devices.

4.4 Comparative Experiment Results

To comprehensively validate the superiority of the proposed MSF-TrashDet, comparative experiments with a series of state-of-the-art detection models are carried out under identical experimental conditions. All evaluation indicators including model size are summarized in Table 3.

As demonstrated in the comparative analysis table, the proposed MSF-TrashDet model attains the highest precision, mAP50, and F1-score among all lightweight

Table 3 Performance comparison with state-of-the-art detectors on TrashNet.

Model	Precision	mAP50	F1	Recall	Params (M)	Size (MB)	GFLOPs
YOLOv5n[23]	87.45	89.29	84.86	82.41	2.64	5.30	7.70
YOLOv8n[24]	89.38	89.86	86.25	83.33	3.15	6.23	8.70
YOLOv11n[25]	83.09	89.27	82.56	82.05	2.61	5.28	6.50
YOLOv12n[26]	83.78	86.01	80.64	77.72	2.59	5.32	6.50
YOLOv26n[27]	84.45	86.32	81.31	82.85	2.37	5.13	5.20
SSD 300[28]	82.68	88.97	85.68	89.69	24.15	190.67	30.58
RetinaNet[29]	71.95	82.56	67.33	63.33	36.39	289.05	18.76
Swin-Transformer[30]	90.92	89.77	**90.47**	**90.05**	44.76	537.82	23.31
Faster-RCNN[31]	56.43	70.01	53.24	50.74	28.20	108.6	108.29
RT-DETR[31]	80.61	74.04	77.68	74.95	31.99	63.09	103.40
ABCS-YOLO[12]	87.50	89.75	84.62	81.91	2.27	4.62	6.10
YOLO11-improved	91.80	90.79	85.90	80.72	2.67	5.44	7.30
MSF-TrashDet (Ours)	**92.78**	**91.91**	89.07	85.65	**1.93**	**4.02**	**5.00**

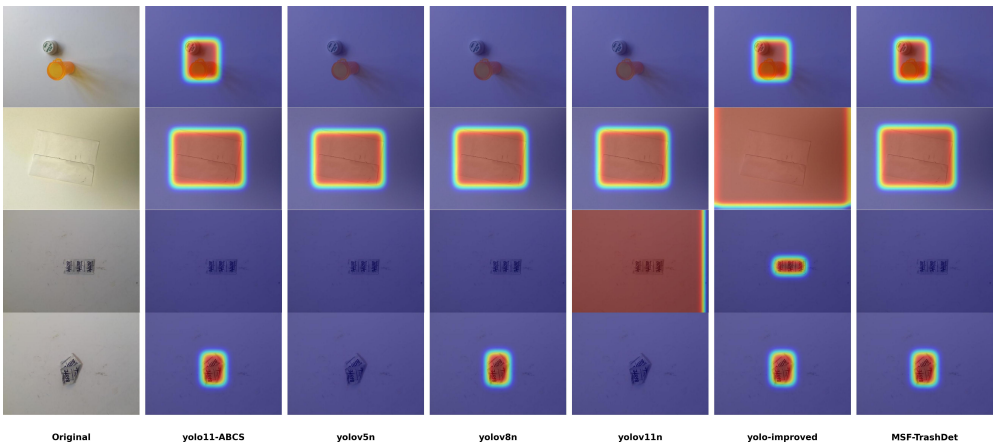


Fig. 8 Feature activation heatmap visualization of waste detection results.

models, while simultaneously exhibiting the lowest parameter count, model size, and GFLOPs. In comparison to conventional two-stage detectors and large-scale transformer-based architectures, our approach achieves comparable or superior detection accuracy with substantially reduced model size and computational demands. This exemplary balance between detection performance and model efficiency substantiates the suitability of MSF-TrashDet for intelligent waste detection applications and deployment on edge devices. Furthermore, Figure 8 illustrates the feature activation heatmaps, providing an intuitive visualization of the model's attention regions and its capacity for discriminative feature learning in waste object detection.

5 DISCUSSION

5.1 Effectiveness of Multi-Scale Feature Enhancement Module

Variable target scales, occlusion and cluttered backgrounds degrade the detection performance of lightweight YOLO models. This paper proposes MSFE block to strengthen multi-scale garbage feature extraction, and ablation results in Table 1 validate its effectiveness.

With only MSFE block embedded into YOLO11n baseline, precision rises from 83.09% to 86.77% and mAP50 increases to 89.92%, while model size and GFLOPs slightly decline. The module adopts three dilated convolution branches with different dilation rates to capture multi-scale garbage features, and dynamically assigns fusion weights via global pooling and Sigmoid activation. Channel-split residual branches retain fine details of tiny fragmented waste and avoid feature loss.

Feature heatmaps in Figure 8 intuitively demonstrate that MSF-TrashDet generates more concentrated, accurate foreground response regions than YOLOv5n, YOLOv8n and original YOLO11n, effectively suppressing background noise for occluded and small trash objects.

5.2 Lightweight Optimization of Backbone and Detection Head

Two lightweight improvements including ShuffleNetV2 backbone replacement and ASFF detection head are adopted to balance accuracy and deployment cost.

Table 1 shows that replacing original backbone modules with ShuffleNetV2 further cuts model size and GFLOPs without obvious mAP50 drop. Channel split, depthwise convolution and channel shuffle operations reduce backbone computation overhead while preserving feature extraction capability.

ASFF solves scale mismatch defects of simple feature concatenation by learning spatial adaptive fusion weights for cross-scale feature maps, greatly boosting detection performance on dense and small waste. After integrating MSFE block, ShuffleNetV2 and ASFF, the model achieves 91.80% precision and 90.79% mAP50 with mild computation growth.

As shown in Table 3, compared with mainstream lightweight YOLOs, two-stage detectors and Transformer models, the pruned MSF-TrashDet achieves the highest mAP50 (91.91%) with only 1.93 M parameters and 5.00 GFLOPs, possessing obvious edge deployment advantages over large heavyweight models.

5.3 Analysis and Limitations of Structured Channel Pruning

Structured channel pruning is utilized for further model compression, and Table 2 records performance under different pruning ratios.

Metrics remain stable or slightly improved when pruning rate increases from 0% to 60%. At the optimal 60% pruning ratio, the model is compressed to 1.93 M parameters with 92.78% precision and 91.91% mAP50. Figure 7 shows BN weights become more concentrated after pruning, confirming redundant low-contribution channels are removed without damaging core feature extraction channels.

Excessive pruning leads to severe performance degradation: mAP50 sharply falls when pruning ratio exceeds 70%, since critical feature channels for irregular and tiny waste are discarded.

This work has two limitations: experiments only rely on the single TrashNet dataset lacking complex real-world scenes, limiting generalization; the model only implements static object detection without temporal analysis for waste accumulation prediction. Future work will expand multi-scene garbage datasets and introduce time-series modules to realize intelligent collaborative monitoring with recycling IoT devices.

6 CONCLUSION

A lightweight multi-scale feature fusion detection network named MSF-TrashDet is proposed for automatic waste classification in this paper, aiming to address the defects of insufficient multi-scale feature extraction, excessive computational cost and poor anti-interference ability existing in traditional lightweight detectors under complex garbage detection scenarios. On the basis of YOLO11n, three targeted optimizations are designed: ShuffleNetV2 lightweight units replace the original backbone modules to reduce baseline computation; a novel Multi-Scale Feature Enhancement (MSFE) block is constructed to capture multi-scale garbage features via dilated convolution branches and dynamic weight fusion; the ASFF adaptive spatial feature fusion head is introduced to mitigate cross-scale feature mismatch for small and occluded trash targets. Furthermore, structured channel pruning with an optimal 60% pruning rate is adopted to further compress the model for edge-device deployment.

A series of ablation, compression and comparative experiments on the public TrashNet dataset verify the superiority of the proposed method. The integrated MSF-TrashDet before pruning achieves 91.80% precision and 90.79% mAP50. After 60% channel pruning, the final model only contains 1.93 M parameters with 5.00 GFLOPs, and its mAP50 reaches 91.91%. Compared with mainstream lightweight YOLO series

(YOLOv5n, YOLOv8n, YOLOv11n), two-stage detection models and Transformer-based architectures, MSF-TrashDet achieves the optimal trade-off between detection accuracy and model complexity. Visualization results of BN layer weight distribution and feature activation heatmaps further prove that the proposed modules effectively filter redundant network parameters and strengthen the model's perception of multi-scale waste targets under cluttered backgrounds.

There remains room for further improvement in this research. The current detection framework only completes static image recognition of garbage objects without supporting time-series analysis for waste accumulation trend prediction. Subsequent research will introduce temporal feature learning modules to couple the detection network with IoT intelligent recycling monitoring systems, so as to provide more comprehensive technical support for intelligent urban waste collection and management scheduling.

Statements and Declarations

Funding

The authors declare that no funds, grants, or other financial support were received for conducting this research and preparing the manuscript.

Competing interests

The authors have no relevant financial or non-financial interests to disclose.

Data availability

The TrashNet public dataset used in this study is openly available at the official Stanford project repository. All model training codes and experimental results will be released upon paper acceptance for reproducibility.

Authors' contributions

Conceptualization: Yousheng Gao; Methodology: Yanfen Tao, Lijuan Lei; Formal analysis and investigation: Dongdong Fan, Pengfei Mei; Writing – original draft preparation: Yanfen Tao; Writing – review and editing: Yousheng Gao; Supervision: Yousheng Gao. All authors read and approved the final manuscript.

References

[1] Awino, F.B., Apitz, S.E.: Solid waste management in the context of the waste hierarchy and circular economy frameworks: An international critical review. Integrated Environmental Assessment and Management **20**(1), 9–35 (2024) https://doi.org/10.1002/ieam.4774

[2] Addas, A., Khan, M.N., Naseer, F.: Waste management 2.0 leveraging internet of things for an efficient and eco-friendly smart city solution. PLOS one **19**(7), 0307608 (2024) https://doi.org/10.1371/journal.pone.0307608

[3] Zhao, X., Wang, L., Zhang, Y., Han, X., Deveci, M., Parmar, M.: A review of convolutional neural networks in computer vision. Artificial Intelligence Review **57**(4), 99 (2024) https://doi.org/10.1007/s10462-024-10721-6

[4] Chen, F., Li, S., Han, J., Ren, F., Yang, Z.: Review of lightweight deep convolutional neural networks: F. chen et al. Archives of Computational Methods in Engineering **31**(4), 1915–1937 (2024) https://doi.org/10.1007/s11831-023-10032-z

[5] Zhang, H., Shao, F., Chu, W., Dai, J., Li, X., Zhang, X., Gong, C.: Faster r-cnn based on frame difference and spatiotemporal context for vehicle detection. Signal, Image and Video Processing **18**(10), 7013–7027 (2024) https://doi.org/10.1007/s11760-024-03789-8

[6] Zhou, H., Chen, C., Xia, Z., Ding, Q., Liao, Q., Wang, Q., Yu, H., Hu, H., Zhang, G., Hu, J., *et al.*: Hgcs-det: a deep learning-based solution for localizing and recognizing household garbage in complex scenarios. Sensors **25**(12), 3726 (2025) https://doi.org/10.3390/s25123726

[7] Yan, M., Sun, Z., Xu, Y., Gong, C., Kang, C.: A review of lightweight object detection technologies for densely occluded scenarios in agricultural fields. Agronomy **16**(11), 1059 (2026) https://doi.org/10.3390/agronomy16111059

[8] Shahab, S., Anjum, M.: Solid waste management scenario in india and illegal dump detection using deep learning: an ai approach towards the sustainable waste management. Sustainability **14**(23), 15896 (2022) https://doi.org/10.3390/su142315896

[9] Rajeev, P.A., Dharewa, V., Lakshmi, D., Vishnuvarthanan, G., Giri, J., Sathish, T., Alrashoud, M.: Advancing e-waste classification with customizable yolo based deep learning models. Scientific Reports **15**(1), 18151 (2025) https://doi.org/10.1038/s41598-025-94772-x

[10] Abo-Zahhad, M.M., Abo-Zahhad, M.: Real time intelligent garbage monitoring and efficient collection using yolov8 and yolov5 deep learning models for environmental sustainability. Scientific Reports **15**(1), 16024 (2025) https://doi.org/10.1038/s41598-025-99885-x

[11] Liu, S., Chen, R., Ye, M., Luo, J., Yang, D., Dai, M.: Ecodetect-yolo: A lightweight, high-generalization methodology for real-time detection of domestic waste exposure in intricate environmental landscapes. Sensors **24**(14), 4666 (2024) https://doi.org/10.3390/s24144666

[12] Ji, Y., Zhang, D., He, Y., Zhao, J., Duan, X., Zhang, T.: Improved yolo11 algorithm for insulator defect detection in power distribution lines. Electronics **14**(6), 1201 (2025) https://doi.org/10.3390/electronics14061201

[13] Zhou, H., Chen, J., Niu, X., Dai, Z., Qin, L., Ma, L., Li, J., Su, Y., Wu, Q.: Identification of leaf diseases in field crops based on improved shufflenetv2. Frontiers in Plant Science **15**, 1342123 (2024) https://doi.org/10.3389/fpls.2024.1342123

[14] Zhang, W., Shao, X., Mei, C., Pan, X., Lu, X.: Multiscale adaptively spatial feature fusion network for spacecraft component recognition. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **18**, 3501–3513 (2025) https://doi.org/10.1109/JSTARS.2024.3523273

[15] Udeh, C.P., Chen, L., Du, S., Liu, Y., Li, M., Wu, M.: Improved shufflenet v2 network with attention for speech emotion recognition. Information Sciences **689**, 121488 (2025) https://doi.org/10.1016/j.ins.2024.121488

[16] Li, S., Wei, J., Su, S., Zhao, Z., Gao, W., Wang, Z., Li, Y., Ou, T.: An innovative multi-scale feature fusion network for change detection of remote sensing images. Electronics **14**(23), 4628 (2025) https://doi.org/10.3390/electronics14234628

[17] Thung, G., Yang, M.: Classification of trash for recyclability status. Cs229 project report, Stanford University (2016). https://github.com/garythung/trashnet

[18] Zhang, X., Jiang, H., Chen, N., Duan, S., Fan, Z.: Rdl-yolo: an efficient algorithm for detecting road surface defects. Scientific Reports (2026) https://doi.org/10.1038/s41598-026-50902-7

[19] Shen, Y., Yang, Z., Khan, Z., Liu, H., Chen, W., Duan, S.: Optimization of improved yolov8 for precision tomato leaf disease detection in sustainable agriculture. Sensors **25**(5), 1398 (2025) https://doi.org/10.3390/s25051398

[20] Zhu, J., Hua, L.S., Chen, B., Kou, W., Rong, J., Zhou, Q., Sun, Y.: Yolo-dcs: A high-precision edge detection method for latex bowl state and latex trail state. Industrial Crops and Products **237**, 122310 (2025) https://doi.org/10.1016/j.indcrop.2025.122310

[21] Dan, H.-C., Yan, P., Tan, J., Zhou, Y., Lu, B.: Multiple distresses detection for asphalt pavement using improved you only look once algorithm based on convolutional neural network. International Journal of Pavement Engineering **25**(1), 2308169 (2024) https://doi.org/10.1080/10298436.2024.2308169

[22] Song, D., Liu, P., Zhu, Y., Li, T., Zhang, K.: Fgs-yolov8s-seg: A lightweight and efficient instance segmentation model for detecting tomato maturity levels in greenhouse environments. Agronomy **15**(7), 1687 (2025) https://doi.org/10.3390/agronomy15071687

[23] Zhang, Y., Guo, Z., Wu, J., Tian, Y., Tang, H., Guo, X.: Real-time vehicle detection based on improved yolo v5. Sustainability **14**(19), 12274 (2022) https://doi.org/10.3390/su141912274

[24] Wang, G., Chen, Y., An, P., Hong, H., Hu, J., Huang, T.: Uav-yolov8: A small-object-detection model based on improved yolov8 for uav aerial photography scenarios. Sensors **23**(16), 7190 (2023) https://doi.org/10.3390/s23167190

[25] He, L.-h., Zhou, Y.-z., Liu, L., Cao, W., Ma, J.-h.: Research on object detection and recognition in remote sensing images based on yolov11. Scientific Reports **15**(1), 14032 (2025) https://doi.org/10.1038/s41598-025-96314-x

[26] Jia, Y., Li, B., Du, Y., Zheng, W., Li, Y., Wu, S.: Research on coal gangue image recognition method based on emam-yolo. ACS Omega (2026) https://doi.org/10.1021/acsomega.5c12454

[27] Çarklı Yavuz, B.: Scale-dependent performance analysis of yolo26 and yolov11 for ppe detection. Electronics **15**(6), 1146 (2026) https://doi.org/10.3390/electronics15061146

[28] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A.C.: Ssd: Single shot multibox detector. In: European Conference on Computer Vision, pp. 21–37. Springer, ??? (2016). https://doi.org/10.1007/978-3-319-46448-0_2

[29] Mohammed, A., Ibrahim, H.M., Omar, N.M.: Optimizing retinanet anchors using differential evolution for improved object detection. Scientific Reports **15**(1), 20101 (2025) https://doi.org/10.1038/s41598-025-02888-x

[30] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)

[31] Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448 (2015). https://doi.org/10.1109/ICCV.2015.169