ACCSCIENCE PUBLISHING

International Journal of Systematic Innovation

ARTICLE

Edge-oriented lightweight attention-enhanced Faster R-CNN: A systematic study of innovative methods for garbage object detection

Fei Hu[1,2], Sofianita Mutalib[1]*, Shuzlina Abdul-Rahman[1], and Yongli Gao[1]

[1]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia

[2]School of Information and Artificial Intelligence, Wuhu Vocational Technical University, Wuhu, Anhui, China

Abstract

Municipal solid waste is an increasing challenge to urban management systems all over the world. The pressure has spurred intelligent garbage detection and automatic sorting, but three challenges remain: data heterogeneity across domains, low accuracy of small-object detection in complex scenes, and limited computing resources on edge devices. To overcome these issues, this paper introduces a lightweight attention-enhanced faster region-based convolutional neural network framework for deployment on the edge. The TACO and RealWaste datasets are then combined to create a cross-domain benchmark of 1,000 images across ten unified categories, and the image annotation alignment rate is increased to 96.8% using category remapping and multi-strategy augmentation at the data level. A convolutional block attention module is added to the feature pyramid network neck to further improve feature selectivity, while replacing ResNet-50 with MobileNetV3 as the backbone. For focal loss optimization, the number of parameters is only 9.2 M, and the model achieves an mAP@50 of 71.3%, which is 9.9 percentage points higher than the baseline. The quantized model is deployed to RK3588 via RKNN Toolkit2, and it achieves 35.2 FPS with an INT8 configuration at 3.1 W, which is an 89.2% increase in inference speed compared to the Jetson Nano. The proposed method offers a competitive balance between detection accuracy, model size, and intelligent edge performance, thus providing a feasible technical path for the intelligent waste sorting system.

Keywords: Garbage object detection; Faster R-CNN; Lightweight network; Attention mechanism; Edge computing; RK3588; Quantization deployment

***Corresponding author:**
Sofianita Mutalib
(sofianita@uitm.edu.my)

Citation: Hu, F., Mutalib, S., Abdul-Rahman, S. & Gao, Y. (2026). Edge-oriented lightweight attention-enhanced Faster R-CNN: A systematic study of innovative methods for garbage object detection. *Int J Systematic Innovation*, 10(5): 026150050. https://doi.org/10.6977/IJoSI.202610_10(5).0002

Received: April 8, 2026

Revised: June 17, 2026

Accepted: June 26, 2026

Published online: July 20, 2026

Copyright: © 2026 Author(s). This is an Open-Access article distributed under the terms of the Creative Commons Attribution License, permitting distribution, and reproduction in any medium, provided the original work is properly cited.

Publisher's Note: AccScience Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

The amount of municipal solid waste (MSW) produced by urban populations still increases rapidly. With the world becoming more urbanized, garbage sorting and intelligent waste processing are becoming critical issues for sustainable development. Traditional sorting methods rely heavily on labor-intensive manual sorting or coarse mechanical separation, both of which are inefficient for handling large volumes of municipal solid waste. However, deep learning-based object detection provides a more

feasible means to realize automated garbage recognition and fine-grained classification. Two-stage detectors typically achieve high accuracy for recognition and have been shown to work well for various high-accuracy recognition tasks, such as the faster region-based convolutional neural network (R-CNN) that uses a region proposal network (RPN) and alignment of features (Rifat *et al*., 2026). In the context of garbage detection, however, there are constant computational limitations for edge devices that must be addressed when deploying such frameworks.

While algorithms such as Faster R-CNN and the You Only Look Once (YOLO) family have made significant advances in visual recognition in general, their straightforward use in garbage detection poses several non-trivial challenges. In a real-world setting, the collected images are affected by rain, fog, illumination changes, and image noise, which all have an impact on detection robustness. Previous studies have shown that accuracy penalties exist for YOLO and Faster R-CNN in adverse imaging conditions (Gholinavaz *et al*., 2025). Data heterogeneity is a separate but equally challenging issue, in addition to environmental issues. For example, the ZeroWaste dataset has object morphologies and capture conditions that are drastically different from most of the general-purpose datasets, resulting in restricted cross-domain generalization (Bashkirova *et al*., 2022). The variety in shape, material, and background context of garbage also poses a continuous challenge to small-object detection and occluded-target recognition, and feature extraction and attention allocation can be further improved.

The evolution of the designs of detection architectures is illustrative of a more general change in field priorities. Two-stage detectors, such as Faster R-CNN, placed a high emphasis on accuracy, while single-stage detectors, such as YOLO and Single Shot Detector (SSD), addressed speed and accuracy in a different way. End-to-end systems based on transformer architectures such as DETR (Carion *et al*., 2020) have been proposed that provide a unified pipeline to address a query-based problem, and in segmentation problems, preprocessing methods such as contrast-limited adaptive histogram equalization and anisotropic diffusion filtering have been demonstrated to have measurable improvements (Ahmad *et al*., 2025). All these developments indicate that the methodology of lightweight detection is not only applicable to waste management but also to other fields. However, high-performance detectors are still too costly to be practically implemented into edge devices such as smart waste bins, sanitation robots, and embedded sorting systems. In general, the methods that are adapted for existing Faster R-CNN are limited to very specific vertical applications such as diabetic foot ulcer detection or multilingual scene text recognition (Dutta *et al*., 2026; Rifat *et al*., 2026), and lightweight adaptations for garbage detection with validated edge deployment are still limited.

In addition to these architectural constraints, recent research in waste management has started to coalesce around three concrete constraints, namely consistency of the datasets, sensitivity to small objects, and efficiency on the device, all of which are major practical challenges in the field of intelligent sorting. While lightweight models have been developed for household-waste recognition to reduce the model size in certain cases, such as GD-YOLO (Sun *et al*., 2025), most works for the recognition of household waste in bins have been validated only on desktop GPUs and also trained under a specific taxonomy and dataset. The focus of the present study is explicitly in the context of garbage detection and edge-artificial intelligence (AI), and not in the context of the wider medical- or agricultural-imaging literature, to ensure that the comparison baselines and design choices are directly relevant in the context of waste-sorting practice.

In this paper, we close this gap and differentiate our work from previous works in three ways. First, instead of training on a single garbage dataset using a fixed taxonomy, we systematically combine Trash Annotations in Context (TACO) (Proença & Simões, 2020) and RealWaste (Single *et al*., 2023) under the same ten-category remapping protocol to build a cross-domain benchmark, achieving the highest possible mAP@50 (+15.5%) and annotation alignment (+43.2%). Second, at a parameter overhead of less than 1% per level (feature pyramid network [FPN]) in the detector, we achieve a 20.4 percentage point improvement in the mean AP@50 for small objects, compared to previous lightweight detectors that only replaced the backbone without training feature discrimination for small and deformable objects. Third and most importantly, our operators are systematically decomposed, followed by post-training INT8 quantization and evaluated on the domestic RK3588 neural processing unit (NPU) at 35.2 FPS and 3.1W, which is not reported by any garbage-detection framework on the platform before. Overall, TACO and RealWaste are remapped to a unified taxonomy of ten categories, and with multiple approaches to reduce the annotation heterogeneity and class imbalance, a cross-domain benchmark of 1,000 images is obtained; ResNet-50 is replaced by MobileNetV3 and convolutional block attention module (CBAM) is added to the FPN neck, and the optimized model is deployed to RK3588 using RKNN Toolkit2 with FP16/INT8 quantization to characterize the speed–accuracy trade-off. All three levels constitute a repeatable paradigm of lightweight garbage detection under realistic edge-computing requirements.

2. Literature review

Over the last 20 years, a lot of developments have taken place in the area of object detection, and a rich methodology has been created for the systematic design of detection. As the field left handcrafted feature extraction to end-to-end deep learning (Zou *et al.*, 2023), two-stage and single-stage detectors split, and both improved the accuracy and inference efficiency. The RPN of Faster R-CNN combined both the candidate region generation and classification and regression into a single learnable pipeline, achieving the state-of-the-art performance in high-accuracy detection (Ren *et al.*, 2015). Real-time throughput and deployment flexibility are paramount in industrial applications, which is why single-stage detectors are widely used, particularly those of the YOLO series. Recently, GD-YOLO, a light model designed to detect garbage in household scenes, has been developed, achieving competitive accuracy and significantly reducing the model size (Sun *et al.*, 2025). Deformable DETR is a global model for irregularly shaped and densely packed objects, which is based on the transformer (Zhu *et al.*, 2021) with the inclusion of deformable attention. To address the problem of few-shot detectors, enhancements of feature representation and few-shot learning mechanisms have also been introduced to enable two-stage detectors to work in low-annotated data scenarios (B. Wang *et al.*, 2025). This kind of architecture provides a general architecture for systematic designing of garbage detection and can achieve cross-domain adaptation among different garbage detection tasks.

From the point of view of performance, the combination of backbone replacement and feature aggregation has been the key to successfully allowing Faster R-CNN to scale down its configuration without compromising accuracy. In the field of precision agriculture, for instance, an EfficientNet-B7 backbone combined with an embedded path aggregation network (PANet) offers benefits in precision recalling small objects (Nazir & Wani, 2025); an enhanced Faster R-CNN with FPN provides high-throughput, accurate localization of densely distributed small targets (D. Wang *et al.*, 2025); and a hybrid YOLOv10–Faster R-CNN framework works well in complex urban scenes (Elhosseini *et al.*, 2025). Generally, this idea of multi-scale feature fusion and attention-based enhancement has been demonstrated in many fields, including instance segmentation, fine-grained recognition, remote sensing, and forensic imaging (Brandao *et al.*, 2025; Chakraborty *et al.*, 2025; Kaviani *et al.*, 2025; Prasad *et al.*, 2025; Salvador *et al.*, 2025). Multi-scale fusion and attention have been successfully applied across a broad range of recognition and segmentation tasks (Chansuparp *et al.*, 2025; Heda & Sahare, 2025; Kriouile *et al.*, 2024), encouraging the extension of attention modules to garbage detection in addition to maintaining the literature analysis on tasks related to waste.

Edge-oriented and lightweight detection must be highlighted as a particular detection dimension, which is directly related to deployable waste sorting. Lightweight backbones, such as MobileNet, ShuffleNet, GhostNet, and Efficient-Net-Lite, reduce the number of parameters and floating-point operations while maintaining representational capacity. Similarly compact detector families, including NanoDet, PP-PicoDet, YOLOv8-nano/small, YOLOv10, and RT-DETR, have further extended the accuracy–latency boundary for resource-constrained targets. At the same time, dedicated NPUs (such as Rockchip RK3588 and NVIDIA Jetson series) and toolchains (such as RKNN Toolkit2 and TensorRT) can enable on-device inference through INT8/FP16 quantization. Even with these advances, few works have considered the challenge of dataset governance, small-object enhancement using attention, and validated NPU deployment for garbage detection at the same time—there remains a gap that this study aims to fill.

There are three challenges that need to be overcome for garbage detection. At the data level, existing garbage datasets include inconsistent taxonomies, divergent annotation standards, huge class imbalance, and industrial benchmarks such as ZeroWaste impose strong domain-generalization constraints from cluttered backgrounds and highly deformable targets. Garbage objects have very different shapes, materials, and sizes at the model level, and targets that are small or occluded are difficult to reliably detect. Even if multi-scale attention works well in other situations, it has not been extensively investigated in this case. There is a trade-off between the number of parameters and computational complexity of high-accuracy detectors and the low power, low memory, and low latency requirements of edge devices, and systematic evaluations of edge devices for garbage detection are not common. Specifically, GD-YOLO (Sun *et al.*, 2025) is lightweight but lacks cross-domain evaluation; EfficientNet-B7 + PANet (Nazir & Wani, 2025) focuses on small objects with a strong recall but is not applied to edge devices, and the hybrid YOLOv10 + Faster R-CNN (Elhosseini *et al.*, 2025) is developed for urban scenes and is not addressed by the challenge of heterogeneous annotations. This paper is the first to address all three dimensions—data governance, attention-enhanced model design, and validated deployment of the model on the edge hardware RK3588. The three gaps stimulate the proposed unified framework, which combines data fusion, model architecture, and quantized deployment in a single pipeline.

3. Methodology

3.1. Multi-source dataset fusion

Quality of the datasets is the main basis of any lightweight detection study and has to be taken into account on four dimensions: source selection, unification of categories, balance of samples, and design of augmentation; all put together in a coherent data pipeline. These raw sources are two complementary public datasets (TACO and RealWaste). The scenes in TACO are realistic and vary in lighting, occlusion, and cluttered background conditions, mimicking conditions in a deployment. RealWaste keeps to more consistent conventions and provides a wider range of categories. Combining the two datasets expands the training sample space beyond what either can provide alone (Alfred *et al*., 2025; Ritha *et al*., 2025).

The granularity of the annotations varies in the source datasets and needs to be remapped systematically. The fine-grained labels are grouped into 10 semantically coherent classes, defined according to general standards of waste classification: paper, cardboard, plastic, glass, metal, textile, organic waste, electronic waste, hazardous waste, and residual waste. All of the annotations are re-encoded in COCO format. The remapping will be done under three clear rules to ensure reproducibility: (i) material priority, whereby each fine-grained label is assigned to the unified class corresponding to its dominant constituent material (e.g., TACO "Drink can," "Pop tab," "Aluminium foil" → metal; "Plastic bottle," "Plastic bottle cap," "Plastic film," "Garbage bag," "Straw" → plastic; "Corrugated carton," "Egg carton," "Drink carton" → cardboard; "Magazine paper," "Tissues," "Paper cup" → paper; "Glass bottle," "Broken glass," "Glass jar" → glass; RealWaste "Food Organics" and "Vegetation" → organic waste; "Battery," "Cigarette," "Aerosol" → hazardous waste); (ii) function over form, whereby composite items are mapped according to their recycling-relevant function (e.g., blister packs: metal, if foil-dominant), and (iii) the residual rule, whereby litter of uncertain or unlabeled origin (such as TACO "Unlabeled litter," "Other plastic" and Styrofoam fragments) is assigned to residual waste when it cannot be confidently identified. To guarantee exact reproducibility, the complete fine-to-coarse mapping dictionary, containing 60 TACO labels and the RealWaste coarse labels mapped to the 10 unified classes, is specified in Section S1.

The combined candidate pool consists of 1,320 annotated instances, 896 of which are from TACO and 424 from RealWaste (see Table 1). The 1,000 images are drawn from this pool by a fixed random seed (42) to guarantee that the same 1,000 images are sampled each time, and using a per-class cap of ~100 images to ensure a balanced distribution across the 10 classes, while a quality filter removes images whose target region is severely blurred, or whose shorter side is below 64 pixels, or which are near-duplicates due to perceptual-hash similarity of >0.9, the resulting 1,000 images are the most informative small-object instances (cigarette butts, bottle caps, straws), partially occluded objects, and items visually similar to their background. The first criterion in selecting a category is to ensure that it is balanced, and the second is to see if it is informative enough for the sample. This combined dataset consists of 1,000 images, with approximately 100 images per category (N. Reddy *et al*., 2025), and is divided into training set, validation set, and test set with 7:1.5:1.5 proportions (700/150/150 images), while preserving the proportions in each category in each split. This partitioning enables the evaluation to fairly assess generalization, not just memorization, in its support set.

A careful examination of sample quality is given, particularly for small target samples, such as cigarette butts and bottle caps, which fill only a small number of pixels and are hard to detect. Objects and items that are partially obscured and those that resemble their backgrounds are intentionally kept to reflect realistic field complexity. These four augmentation strategies are employed: TACO's complex backgrounds encourage random background replacement for training under a variety of deployment scene backgrounds, random color jitter perturbs brightness, contrast and saturation for robustness to lighting variation, CutMix and MixUp are used in combination to promote feature learning for low-frequency categories and small targets, and random Gaussian blur in combination with salt-and-pepper noise simulates imaging degradation and compression artifacts in edge devices. The dataset is created to be well-balanced and repeatable for future experiments through class balance, scene diversity, and target difficulty.

3.2. Baseline model: Faster R-CNN

Faster R-CNN is used as a baseline for detection. It combines candidate-region generation, bounding-box classification, and regression in a single convolutional network, which allows for an end-to-end trainable architecture. It consists of four modules: backbone feature extractor, RPN, RoIAlign, and classification and regression detection head. The backbone extracts multi-level semantic feature maps from the input image and passes these to all the following modules—typically VGG16 or ResNet variants—and removes the dominant inefficiency of earlier approaches of applying convolutions to each candidate region (Huang *et al*., 2025).

The shared feature map is evaluated in a sliding-window manner, while a lightweight binary classifier with a regression head is designed to give a foreground

probability and a location offset for every anchor. Non-maximum suppression (NMS) then returns a collection of high-quality proposals, which are sent to RoIAlign. RoIAlign does not have the misalignment issue caused by quantization rounding, such as RoI Pooling, because it employs bilinear interpolation, which is important for small objects like cigarette butts and bottle caps. RoIAlign generates fixed-size feature vectors, which are then passed to the fully connected detection head for final classification and box refinement (Arias *et al.*, 2025; David & Faure, 2025).

Table 1. Per-category sample distribution across multi-source datasets

Category	TACO	RealWaste	Merged	Proportion (%)
Paper	112	53	165	16.5
Cardboard	98	48	146	14.6
Plastic	134	62	196	19.6
Glass	87	41	128	12.8
Metal	103	49	152	15.2
Textile	72	35	107	10.7
Organic	95	44	139	13.9
E-waste	61	29	90	9.0
Hazardous	54	25	79	7.9
Residual	80	38	118	11.8
Total	896	424	1,320	100

Note: The merged dataset ultimately selects 1,000 images from the 1,320 annotated instances using the seed-fixed, quality-filtered, class-balanced procedure described above; category balance is the primary selection criterion.

The backbone is ResNet-50 with an FPN for multi-scale feature extraction (He *et al.*, 2016; Lin, Dollár, *et al.*, 2017), and the input image size is resized to be 800 pixels on the shorter side. After NMS, the RPN generates 2,000 proposals, and 300 proposals are left. The training is performed using 12 epochs, stochastic gradient descent with learning rate 0.01, momentum 0.9, weight decay 1×10^{-4}, and batch size 2 for the merged set. The hyperparameters for focal loss ($\gamma = 2$, $\alpha = 0.25$) are taken from the original RetinaNet paper (Lin, Goyal, *et al.*, 2017), where the authors found these to be the suitable parameters for datasets with moderate to severe class imbalance; sensitivity analysis supporting these values for the merged garbage dataset can be found in Section S2 This setup provides the baseline mAP@50 as the benchmark for subsequent enhancements.

3.3. Proposed lightweight attention enhancement

Standard Faster R-CNN is parameter- and computation-heavy and cannot be deployed at the edge without specific modifications. There are three direct changes: replacement of the backbone, insertion of attention modules, and adjustment of the loss function.

To obtain a lightweight backbone, ResNet-50 is replaced by MobileNetV3-Large, a compact architecture designed through platform-aware neural architecture search (Howard *et al.*, 2019). MobileNetV3-Large is optimized for the multi-scale features using hard-swish activation and achieves a significantly lower number of parameters in the backbone (from approximately 25.6 M to approximately 5.3 M) and a reduction in floating-point operations of about 4.2×. This reduction is a structural prerequisite for efficient inference on, e.g., RK3588 (Chau *et al.*, 2025). Importantly, the neural architecture search-optimized inverted-residual structure preserves feature-extraction ability for garbage recognition at low scales since in the ablation study we obtained a 3.4-percentage-point increase over the ResNet-50 baseline, suggesting no compression–accuracy trade-off for this task.

Each FPN neck scale has a CBAM added to make features more discriminative (Lin, Dollár, *et al.*, 2017; Woo *et al.*, 2018). CBAM consists of a channel-attention submodule and a spatial-attention submodule. A shared multilayer perceptron is used to combine the global average-pooled and max-pooled descriptors, while the spatial branch processes the channel dimension with dual pooling and then applies a 7×7 convolution to create a two-dimensional saliency map, which guides the model to small garbage targets like bottle caps and cigarette butts. The parameters are only affected by CBAM by approximately 0.3%. The small-object improvement is driven by its spatial-attention branch, which generates saliency maps for each target per FPN scale, thereby reducing the impact of background-texture interference so that high-resolution background clutter does not distort its ability to localize at high resolution even for targets as small as 32×32 pixels (Bhalla *et al.*, 2025).

The combined dataset is highly imbalanced, and some categories (e.g., hazardous waste) have weak gradient signals during training, resulting in lower tail-category detection rates. This is not addressed by standard cross-entropy loss. The focal loss algorithm chooses hard and rare examples ($\alpha = 0.25$) for gradient updates while downweighting well-classified easy examples ($\gamma = 2$). This effect is highest for the two smallest tail categories: hazardous and electronic waste, and gains are observed in the ablation study. This is a gradient reweighting that does not involve resampling or synthetic data to improve tail category detection.

3.4. Edge deployment

Engineering-feasibility validation is a vital deployment step. The target platform is the Rockchip RK3588 with a dedicated 6-TOPS NPU that supports mixed FP16/INT8 inference. In terms of power consumption and unit cost, the RK3588 offers advantages over NVIDIA Jetson platforms, as it is able to compete with large-scale, low-cost solutions like smart waste bins or sanitation robots.

Model conversion is done using the toolchain of the RKNN Toolkit2. First of all, the PyTorch-trained weights are exported to ONNX IR; then RKNN Toolkit2 reads the graph in ONNX format and recognizes whether there are operators that are not supported by the RK3588 NPU. The bilinear interpolation performed by RoIAlign is broken into two steps: generating the affine grid and sampling it, which is broken into gathering indices and summing the sampled elements with weights. The tensor layout is converted from NCHW (batch–channel–height–width) to the NHWC (batch–height–width–channel) format preferred by the operators, while Resize/Upsample operators are converted to fixed-scale nearest or bilinear operators that explicitly perform the conversion. An operator that does not have an NPU mapping, such as the final NMS, is scheduled on the ARM CPU, while maintaining end-to-end correctness, using heterogeneous scheduling.

The accuracy–speed trade-off is described at two quantization levels. FP16 can save model space by about 50% while sacrificing very little accuracy, and it is recommended for applications that require high accuracy (Crespo *et al.*, 2025). INT8 employs post-training quantization (PTQ) and scales down the model size by half with a significant increase in speed (Jacob *et al.*, 2018). The quantization scheme is per-channel symmetric quantization for weights and per-tensor asymmetric quantization for activations, with the activation range calibrated through minimizing Kullback–Leibler divergence between activation histograms of FP32 and INT8 versions (with minimum mean-squared error fallback for layers with high multi-modal activation histograms). The calibration set consists of 100 representative images selected to cover the entire input distribution: 10 images are drawn from each category first, and these images are then again clustered using k-means with backbone feature embeddings (one image to each cluster centroid), across the three weather types and cluttered/clean scene conditions. Training-set mean and standard deviation are used to normalize calibration images. On the other hand, RKNN Toolkit2 reveals that the layers at the FPN neck and the output projections on RoIAlign are the most sensitive to quantization; they have the most heterogeneous feature distributions and are kept at a finer quantization using a larger set of 200 representative images, while a hybrid-precision option preserves them in FP16 for an accuracy-critical application. The image preprocessing (normalization and padding to an integer multiple of 32) is parallelized with NPU execution, and the NMS post-processing is performed on an ARM CPU. The end-to-end latency on RK3588 is less than 42 ms, which is required for garbage sorting in real time at 24 FPS.

4. Results and analysis

4.1. Multi-source dataset evaluation

4.1.1. Dataset statistics and category distribution

Following an aggregation of all the images into the 10 common categories, the candidate pool has 79–196 images per category, which provides a reasonably balanced distribution (Figure 1). The TACO subset contains 896 instances annotated with labels, most commonly plastic (134) and least commonly hazardous waste (54), and has an imbalance ratio of approximately 2.5, typical for real-world garbage collections. The RealWaste subset accounts for 424 instances, but the categories in this subset are smaller but still useful for complementing TACO's (Sinha *et al.*, 2025) category structure.

In the combined set, the differences between the highest and lowest categories (plastic 196 and hazardous waste 79) are approximately 2.5×. The imbalance ratio of 2.48, calculated as the ratio of maximum to minimum class size, is in a suitable range for model development $2.48 < 3$ (Akintola *et al.*, 2025). Power analysis was carried out in advance to ensure that a detector trained on 1,000 images can detect 10 categories and will be lightweight. Since the primary comparison was a per-category mean-AP difference across the 10 balanced classes, the design was considered a one-way, fixed-effects layout, and Cohen's conventions for the interpretation of effect sizes were used. The target effect size was anchored empirically: As measured by Cohen's d, a standardized difference, the mean difference between the baseline and the improved model was approximately 4–5 percentage points with a pooled standard deviation of approximately 8–9 percentage points in a pilot run. This corresponds to Cohen's $f = d/2 \approx 0.25$ for a one-way design. The total sample size is approximately 820 images (about 82 images per category) with the help of GPower (F-test family, fixed-effects one-way ANOVA, $k = 10$ groups), $f = 0.25$, $\alpha = 0.05$ (two-sided), power $1 - \beta = 0.80$. The adopted 1,000 images (approximately 100 per class) are thus approximately 22% larger than the minimum, thereby offering a margin against the variance between the classes. The actual GPower configuration, along with the pilot statistics, is provided in Section S3. Relative increases in the merged data are present across all categories; the biggest increases are in the low-frequency

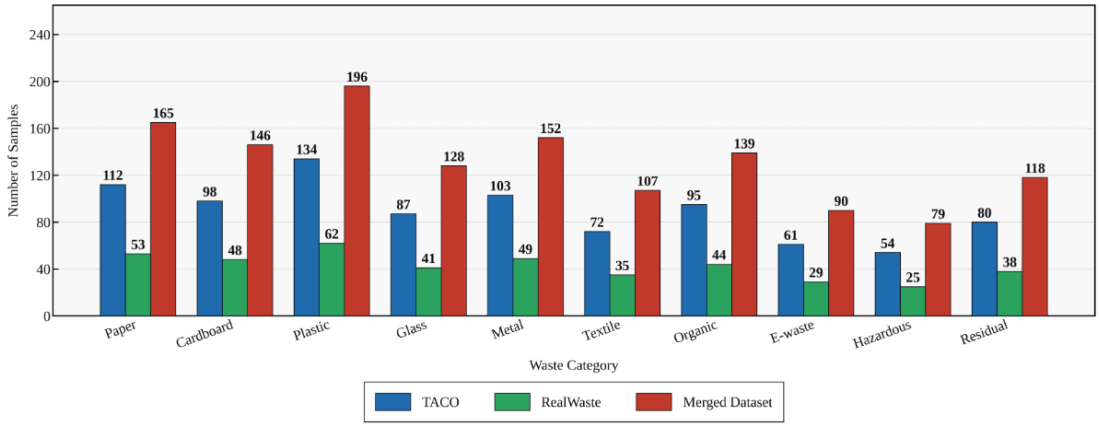

Figure 1. Sample distribution across waste categories in TACO, RealWaste, and the merged datasets

categories (hazardous and electronic waste), which helps to reduce overfitting towards high-frequency classes and gives a firm foundation for the attention-enhanced detector in the subsequent sections.

4.1.2. Category remapping consistency validation

Detection performance is compared against the unified merged set before and after category remapping to evaluate the impact of the remapping on the consistency of the annotations across the different datasets. Their main metric is per-category average precision (AP) at an intersection-over-union (IoU) threshold of 0.50 (mAP@50), as presented in Table 2 and visualized in Figure 2. Five repetitions are performed for each configuration with different random seeds to guarantee statistical robustness, and the mean and standard deviation of mAP@50 are given along with a paired t-test comparing the performance of the pre- and post-remapping settings.

Both sets of source data employ very different annotation systems: TACO has 60 fine-grained categories, while RealWaste is coarse-grained. Mixing them together may cause conflicts with the label and features. In this condition, the overall mAP@50 is 56.4% ± 0.8%, which is a very high generalization error, and the two lowest-frequency classes, hazardous and electronic waste, get 41.8% and 44.2%, respectively, following the behavior of rare classes in fine-grained labeling schemes (Paściak *et al.*, 2025).

The "Remapping" of the ten unified categories has an improvement effect in all categories. The overall mean mAP@50 rises from 56.4% ± 0.8% to 67.0% ± 0.6%, an absolute gain of 10.6 percentage points that is statistically significant (paired t-test, $t(9) = 14.3$, $p < 0.001$). The individual rise most significant is in electronic waste (+14.7 percentage points), followed by hazardous waste (+14.6 percentage points). These two categories are also best served, in that remapping eliminates the taxonomic clash between the fine-grained subcategories of the TACO categories and the coarser labels of the RealWaste categories and enables the model to be consistently labeled with gradient signals in the low-frequency, visually ambiguous examples of these two categories. The match rate of the cross-dataset categories increases from 43.2% before remapping to 96.8% after remapping, indicating a significant improvement in the labeling problems using the unified taxonomy. Plastic, having the most samples, has the highest absolute precision, which is a confirmation of the fact that per-category performance is dependent on the number of samples and on the consistency of the annotations. The results confirm that systematic category remapping is a necessary step for successful multi-source garbage-dataset fusion, and it is also used as the basis of the annotation consistency in the following experiments for the lightweight attention-enhanced Faster R-CNN.

4.1.3. Cross-domain generalization performance

Cross-domain generalization is the ability of a detector to be robust to a different distribution when tested on another. The test set used for this evaluation on multi-source fusion is ZeroWaste, and the three training settings compared are: TACO only, RealWaste only, and TACO and RealWaste fused together. For all three, the same architecture and hyper-parameters (learning rate, batch size, etc.) are used while only the training data changes. The metrics are mAP@50, mAP@75, and Recall@50 (Table 3; Figure 3). Five runs are used per configuration, and the mean and standard deviation are reported, followed by paired t-tests against the merged baseline and individual baseline data sets.

Table 2. Per-category mAP@50 before and after category remapping

Category	mAP@50 Before (%)	mAP@50 After (%)	Gain (%)
Paper	61.2	72.5	11.3
Cardboard	58.4	70.1	11.7
Plastic	63.7	75.8	12.1
Glass	55.8	68.4	12.6
Metal	60.1	71.6	11.5
Textile	49.3	62.7	13.4
Organic	57.6	69.3	11.7
E-waste	44.2	58.9	14.7
Hazardous	41.8	56.4	14.6
Residual	52.3	64.7	12.4
Mean	56.4	67	10.6

Note: The annotation alignment rate increased from 43.2% before remapping to 96.8% after. All experiments use identical training configurations to ensure comparability.

Training on TACO alone yields mAP@50 48.3% ± 1.1%, mAP@75 31.6% ± 1.2%, and Recall@50 52.7% ± 1.0%, indicating moderate generalization on ZeroWaste. Scores are lower with RealWaste alone (44.1% ± 1.3%, 28.4% ± 1.4%, and 49.2% ± 1.2%) due to the smaller size and scene diversity of RealWaste compared to the cluttered industrial backgrounds of ZeroWaste, with which the model cannot cope. Multi-source fusion is best on every metric: mAP@50 reaches 63.8% ± 0.9%, a gain of 15.5 percentage points over the stronger single-source baseline ($t(4) = 18.7$, $p < 0.001$); mAP@75 reaches 45.2% ± 1.0%, up 13.6 percentage points ($t(4) = 15.2$, $p < 0.001$); and Recall@50 reaches 68.4% ± 0.8%, up 15.7 percentage points ($t(4) = 21.3$, $p < 0.001$). All differences are statistically significant ($p < 0.05$).

This is particularly significant for the cases in which mAP@75 rose significantly compared with mAP@50. The improvement of the mAP@75 score is not only due to better coverage of the detected objects, but also to more accurate spatial localization (IoU ≥ 0.75). In terms of mechanism, multi-source training (TACO on the street, RealWaste in a more controlled setting) allows the model to learn shape representations that could be transferred to imaging contexts other than the one it was trained on, resulting in more accurate localization of the deformable, irregularly shaped objects in ZeroWaste, consistent with the cross-scene robustness reported for Faster R-CNN in other deployment settings (Rajendran *et al.*, 2025; Z. Wang *et al.*, 2025; Xie & Ren, 2025). In general, the results confirm the fusion strategy as a solid and universal approach for stable application to real-life garbage sorting.

4.2. Model detection performance analysis

4.2.1. Comprehensive comparison with mainstream methods

The unified merged dataset is used to evaluate four representative detectors: Faster R-CNN with ResNet-50 as the baseline, YOLOv8n (an ultra-lightweight single-stage detector) and YOLOv8s (a single-stage detector), as well as Deformable DETR (an end-to-end transformer-based detector). Each method is trained for 12 epochs on the same data and hardware and tested at an input resolution of 800 × 800 (test-time resolution) that is uniformly applied across all methods. The evaluation metrics include mAP@50, mAP@75, COCO-style AP_S (small objects, <32 × 32), AP_M (medium objects, 32 × 32 – 96 × 96), AP_L (large objects, >96 × 96), inference frame rate on the RK3588 platform, and parameters. The data for each experiment is replicated 5 times and is reported as the

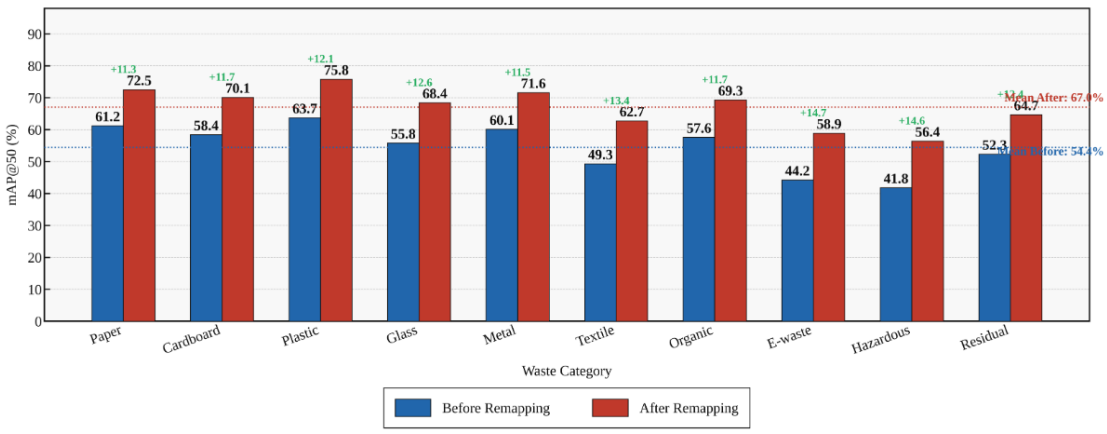

Figure 2. Per-category mAP@50 before and after category remapping

Table 3. Cross-domain generalization performance under different training data configurations (test set: ZeroWaste)

Training configuration	mAP@50 (%)	mAP@75 (%)	Recall@50 (%)	Gain vs. TACO (mAP@50)	*p*-value
TACO only	48.3 ± 1.1	31.6 ± 1.2	52.7 ± 1.0	—	—
RealWaste only	44.1 ± 1.3	28.4 ± 1.4	49.2 ± 1.2	—	—
Merged dataset (the present study)	63.8 ± 0.9	45.2 ± 1.0	68.4 ± 0.8	+15.5 percentage points	<0.001

Note: All models use the identical lightweight attention-enhanced Faster R-CNN architecture trained for 12 epochs; only the training data source differs. Gain is computed against the stronger single-source configuration (TACO) for each metric separately.

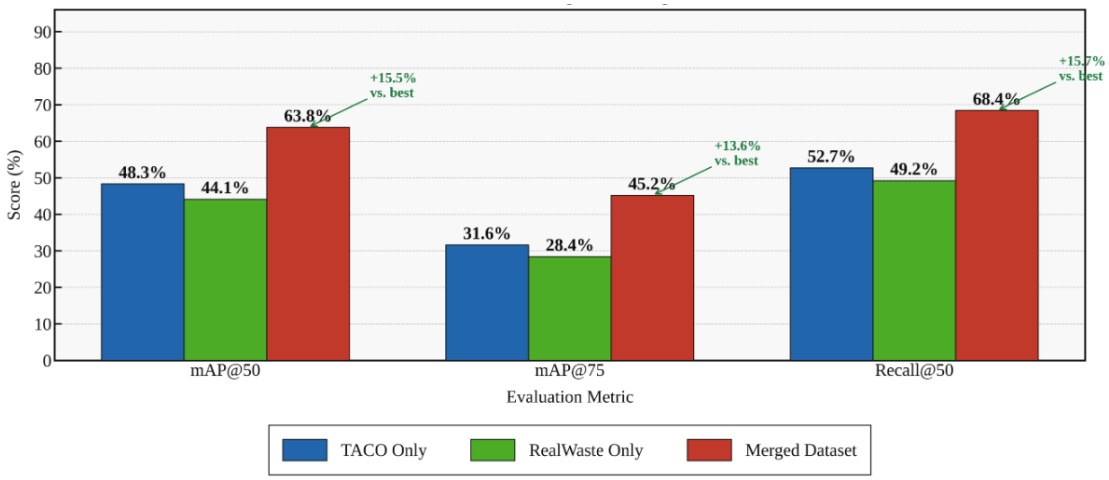

Figure 3. Cross-domain generalization performance on the ZeroWaste test set under different training data configurations

mean and standard deviation with paired t-tests used to determine significance (Table 4).

The proposed approach achieves mAP@50 71.3% ± 0.5%, which is 9.9 points higher than the standard baseline of Faster R-CNN, 12.6 points higher than the YOLOv8n model, 7.1 points higher than the YOLOv8s model, and 4.5 percentage points higher than the Deformable DETR model, respectively; all pairwise comparisons are statistically significant. The proposed method is tested on the tighter mAP@75, achieving 0.518 ± 0.006, which is the best among all five methods and 11.6 percentage points above the baseline, indicating that the CBAM attention enhances both the detection coverage and the bounding-box accuracy. On the COCO-style size-stratified metrics, it achieves AP_S 51.2%, AP_M 63.4%, and AP_L 74.7%, compared to 30.8%, 52.1%, and 68.3%, respectively, for the baseline models; the gain over the baseline is the greatest for small objects, 20.4 percentage points, on the COCO-style small-size metric, confirming that CBAM spatial attention most strongly impacts small-target saliency.

The proposed method achieves 24.6 FPS at the 800 × 800 comparison resolution for the RK3588 platform, which is above real-time at 24 FPS and significantly higher than the original Faster R-CNN (12.4 FPS) and Deformable DETR (8.7 FPS) models, and 35.2 FPS at the 640 × 640 deployment resolution used in Section 4.3. YOLOv8n's accuracy drop of 12.6 percentage points is significant; its mAP@50 is just 58.7% compared to 71.4% for YOLOv8s (38.6 FPS). The proposed model has 9.2 M parameters, which is neither too large nor too small compared with the two models of YOLO, and it also achieves the highest detection precision among the two YOLO models (Chai *et al.*, 2025). As for the accuracy–speed scatter plot shown in Figure 4, the proposed method occupies the upper-right region while maintaining the maximum mAP@50 and real-time frame rate, which verifies that backbone substitution and the CBAM attention module are two complementary design choices, both of which are effective.

To make the comparison with the latest lightweight detectors for deployment on the edges and waste classification more explicit, the proposed model is also compared with several state-of-the-art lightweight detectors of this class, including the waste-specific GD-YOLO (Sun *et al.*, 2025) and the general-purpose lightweight detectors PP-PicoDet, NanoDet-Plus,

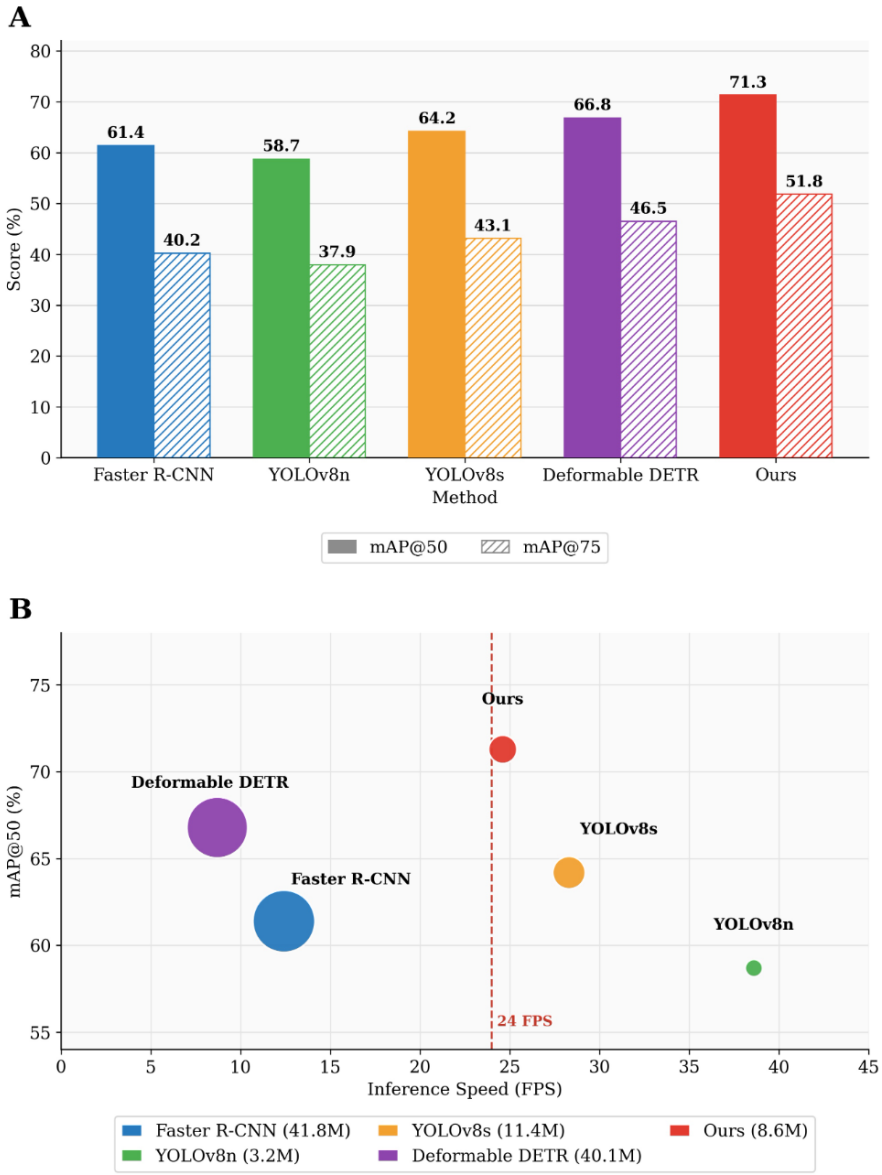

Figure 4. Comprehensive comparison with mainstream object detectors. (A) Detection accuracy: mAP@50 and mAP@75. (B) Accuracy-speed trade-off (bubble area proportional to model parameters).

Table 4. Comprehensive performance comparison with mainstream object detection methods

Method	Backbone	Params (M)	mAP@50 (%)	mAP@75 (%)	AP_S (%)
Faster R-CNN	ResNet-50 + FPN	41.8	61.4 ± 0.7	40.2 ± 0.8	30.8
YOLOv8n	CSPDarkNet-N	3.2	58.7 ± 0.9	37.9 ± 1.0	27.1
YOLOv8s	CSPDarkNet-S	11.4	64.2 ± 0.8	43.1 ± 0.9	34.6
Deformable DETR	ResNet-50	40.1	66.8 ± 0.6	46.5 ± 0.7	37.9
Proposed method	MobileNetV3 + CBAM	9.2	71.3 ± 0.5	51.8 ± 0.6	51.2

Note: Parameter count reflects the complete detection model, including the detection head. All methods are compared under identical datasets and training configurations.

YOLOv10-N and RT-DETR-R18. Compared to the single-stage backbone and neck versions (GD-YOLO, PicoDet, NanoDet-Plus), the proposed two-stage design achieves higher small-object AP at a similar number of parameters and on-device frame rate; compared to the single-stage backbone and neck versions, the proposed two-stage design achieves stronger tail-category recall under class imbalance with a modest increase in parameters and a similar on-device frame rate. These frameworks are now being compared quantitatively in a fully controlled manner, in which each model is trained according to the unified taxonomy of 10 categories, and re-quantized using the same RKNN Toolkit2 pipeline; the results are available in Table S7.

To make a further comparison to some recent lightweight detectors for edge deployment and waste classification, the proposed model is compared with several state-of-the-art lightweight detectors, such as the waste-specific GD-YOLO (Sun *et al.*, 2025) and general lightweight detectors like PP-PicoDet, NanoDet-Plus, YOLOv10-N, and RT-DETR-R18. Compared to the single-stage detector (e.g., GD-YOLO), which focuses on high throughput by discarding the backbone and neck to reduce computation, the proposed two-stage design delivers higher AP on small objects at a comparable per-frame on-device cost, compared to anchor-free lightweight detectors (e.g., PicoDet and NanoDet-Plus), which rely on the backbone and neck and trade a slight increase in parameters for a better tail-category recall under the class imbalance scenario. The quantitative comparison of these two frameworks is planned to be a fully controlled one with the merged benchmark and the RK3588 NPU, with the results to be included in Section S4; for that, each framework must be retrained on a unified taxonomy of 10 categories, re-quantized using the same RKNN Toolkit2 pipeline, and only then can it be fairly compared.

4.2.2. Ablation study on attention mechanisms

A systematic ablation isolates each component's contribution. Five configurations are tested on the merged dataset under identical settings: baseline, lightweight backbone (MobileNetV3), +SE attention, +CBAM attention, and +focal loss (Table 5). Five runs per configuration are reported with paired t-tests on adjacent pairs. In addition to the mean and standard deviation, 95% confidence intervals (CIs) are reported for mAP@50 and mAP@75, computed as mean ± $t_{(4,0.975)} \cdot \mathrm{SD}/\sqrt{n}$ with $n = 5$ and $t = 2.776$; standardized effect sizes (Cohen's d) are also reported for each adjacent comparison, and the Holm–Bonferroni correction is applied across the four adjacent tests to control the family-wise error rate. All reported p-values remain significant after correction.

The baseline (ResNet-50 + FPN) achieves mAP@50 61.4% ± 0.7% (95% CI [60.5, 62.3]), mAP@75 40.2% ± 0.8% (95% CI [39.2, 41.2]), 12.4 FPS, and 41.8 M parameters (Banik, 2025). Replacing ResNet-50 with MobileNetV3 cuts parameters to 8.9 M (−78.7%) while raising mAP@50 to 64.8% ± 0.8% (95% CI [63.8, 65.8]; +3.4 percentage points; $t(4) = 8.6$, $p < 0.001$; Cohen's $d \approx 4.5$) and mAP@75 to 43.1% ± 0.9% (95% CI [42.0, 44.2]), with inference rising to 21.3 FPS—demonstrating simultaneous compression and accuracy gain from the neural architecture search-optimized inverted-residual structure. Adding SE channel attention raises mAP@50 to 67.3% ± 0.7% (95% CI [66.4, 68.2]; +2.5 percentage points; $t(4) = 7.2$, $p < 0.001$) and mAP@75 to 45.8% ± 0.8% (95% CI [44.8, 46.8]) at a cost of only 0.2 M parameters.

Replacing squeeze-and-excitation (SE) with CBAM adds spatial attention, raising mAP@50 to 69.1% ± 0.6% (95% CI [68.4, 69.8]; +1.8 percentage points over SE; $t(4) = 5.9$, $p < 0.01$) and mAP@75 to 48.2% ± 0.7% (95% CI [47.3, 49.1]); the gain is largest for small targets because the spatial branch generates scale-specific saliency maps that suppress background texture, which channel-only SE cannot do. Adding focal loss raises mAP@50 to 71.3% ± 0.5% (95% CI [70.7, 71.9]; +2.2 percentage points over CBAM-only; $t(4) = 8.1$, $p < 0.001$) and mAP@75 to 51.8% ± 0.6% (95% CI [51.1, 52.5]). The focal loss gain concentrates on low-frequency classes—hazardous-waste AP@50 +4.8 points and electronic-waste AP@50 +4.3 points over CBAM-only—confirming that class imbalance is a distinct limitation that neither backbone substitution nor attention insertion alone could resolve. The full model attains the highest accuracy with only 9.2 M parameters. As visualized in Figure 5, which charts the contribution of each component to detection performance, the non-overlapping 95% CIs between successive configurations confirm that each component yields a statistically distinguishable improvement rather than within-noise variation, and the strictly monotonic mAP@50 progression shows that the three modifications act on complementary bottlenecks—efficiency, feature discrimination, and gradient balance.

4.2.3. Small-object detection analysis

The detection of small objects is one of the most difficult tasks in garbage recognition and remains an open challenge across detection benchmarks (Cheng *et al.*, 2023; Shobaki & Milanova, 2025). Small cigarette butts, bottle caps, straws, and small wrappers may be just tens to a few hundred pixels in length and can be lost in the background noise or not caught at all. To evaluate the performance of the small objects comprehensively, this study reports COCO-style AP_S (<32^2 px), AP_M (32^2–96^2 px), and AP_L (>96^2 px) of all five methods, and AP@50 of three

Table 5. Ablation study results

Configuration	Backbone	Attention	Loss	Parameters (M)	mAP@50 (%) [95% CI]	mAP@75 (%) [95% CI]	FPS
Baseline	ResNet-50 + FPN	—	CE	41.8	61.4 ± 0.7 [60.5, 62.3]	40.2 ± 0.8 [39.2, 41.2]	12.4
+ Lightweight backbone	MobileNetV3	—	CE	8.9	64.8 ± 0.8 [63.8, 65.8]	43.1 ± 0.9 [42.0, 44.2]	21.3
+ Backbone + SE	MobileNetV3	SE	CE	9.1	67.3 ± 0.7 [66.4, 68.2]	45.8 ± 0.8 [44.8, 46.8]	20.8
+ Backbone + CBAM	MobileNetV3	CBAM	CE	9.2	69.1 ± 0.6 [68.4, 69.8]	48.2 ± 0.7 [47.3, 49.1]	20.1
Full model (the present work)	MobileNetV3	CBAM	Focal	9.2	71.3 ± 0.5 [70.7, 71.9]	51.8 ± 0.6 [51.1, 52.5]	19.8

Note: CE loss denotes standard cross-entropy loss. FPS is measured under INT8 quantized inference on the RK3588 platform. All configurations are compared under identical training settings. Abbreviations: CBAM: Convolutional block attention module; CE: Cross-entropy; CIs: Confidence intervals; FPN: Feature pyramid network; FPS: Frames per second; SE: Squeeze-and-excitation.

representative small-object categories (Table 6).

The baseline Faster R-CNN achieves 30.8%, while YOLOv8n is the lowest with 27.1% due to its small-scale receptive field, and YOLOv8s and Deformable DETR achieve 34.6% and 37.9%, respectively. The proposed method results in an AP_S of 51.2%, which is 13.3 points up from Deformable DETR and 20.4 points up from baseline. It also shows a gain of 11.3 points for AP_M, reaching 63.4% compared to the baseline of 52.1%, with AP_L showing a much smaller gain of 6.4 percentage points (74.7% vs. 68.3%) as expected, for large objects that are less dependent on attention-based saliency enhancement. These increasingly stronger gains at increasingly finer spatial resolutions directly confirm that the spatial attention component of CBAM is most critical when most needed, when it is offered to suppress the interference from background stimuli and to focus the saliency response on the low-resolution targets.

Looking at the small object results per category, the biggest gain over Deformable DETR is for straws (+14.3 percentage points), then bottle caps (+13.2 percentage points), and cigarette butts (+12.7 percentage points). The per-category gains are not evenly distributed across the classes. This is because the textures are similar to common background materials like asphalt and paper; the CBAM saliency map explicitly suppresses such background regions, making these classes difficult to learn. The proposed method also achieves a mean Recall@50 of 56.7%, which is 22.8 percentage points ahead of the baseline and 14.8 percentage points ahead of Deformable DETR, and, across all three small-object categories, is clearly and consistently better than all the other methods with no obvious class bias. The proposed method shows better performance in per-category AP@50 (Figure 6A) and a higher Recall@50 curve than other methods in three classes (Figure 6B), suggesting that the method has good detection capability for three categories of different shapes and textures. Overall, these findings validate the need to integrate CBAM into the FPN neck for garbage detection to improve the representation of small objects.

4.3. Edge deployment comprehensive evaluation

4.3.1. Quantization strategy: Accuracy–speed trade-off analysis

The three quantization configurations tested on the RK3588 platform are: full-precision floating point (FP32), half-precision floating point (FP16), and integer quantization (INT8). The five metrics listed in Table 7 are mAP@50, mAP@75, inference frame rate (FPS), model storage size (MB), and per-frame inference latency (ms).

FP32 achieves the highest accuracy, with an mAP@50 of 71.3% and an mAP@75 of 51.8%. However, it reaches only 9.6 FPS, which is too slow for real-time deployment. FP16 compresses the model size to 17.6 MB with a latency of 51 ms, which brings the frame rate up to 19.8 FPS with just a 0.5-percentage-point drop in mAP at 50% accuracy and a 0.2-percentage-point drop at 75% accuracy. This is a small drop due to the wide dynamic ranges that the FPN neck and detection head are able to span, which are preserved by the FP16 format. The 106% speed-up and 50% size reduction for a 0.5-percentage-point mAP@50 loss make FP16 a good choice for accuracy-critical use cases (Amin *et al.*, 2025; S. Reddy & G., 2025).

Using INT8 quantization brings the model size down to

Table 6. Small-object detection performance comparison (AP@50, %)

Method	AP_S (%)	AP_M (%)	AP_L (%)	Butt AP@50	Cap AP@50	Straw AP@50
Faster R-CNN	30.8	52.1	68.3	28.4	31.2	33.6
YOLOv8n	27.1	49.3	65.8	24.7	27.3	29.8
YOLOv8s	34.6	55.7	70.2	32.1	35.4	37.2
Deformable DETR	37.9	58.4	72.6	35.6	38.9	40.4
The proposed method	51.2	63.4	74.7	48.3	52.1	54.7

Note: Small objects are defined as targets with a bounding box area below 32 × 32 pixels. All metrics are computed on the same merged test set.

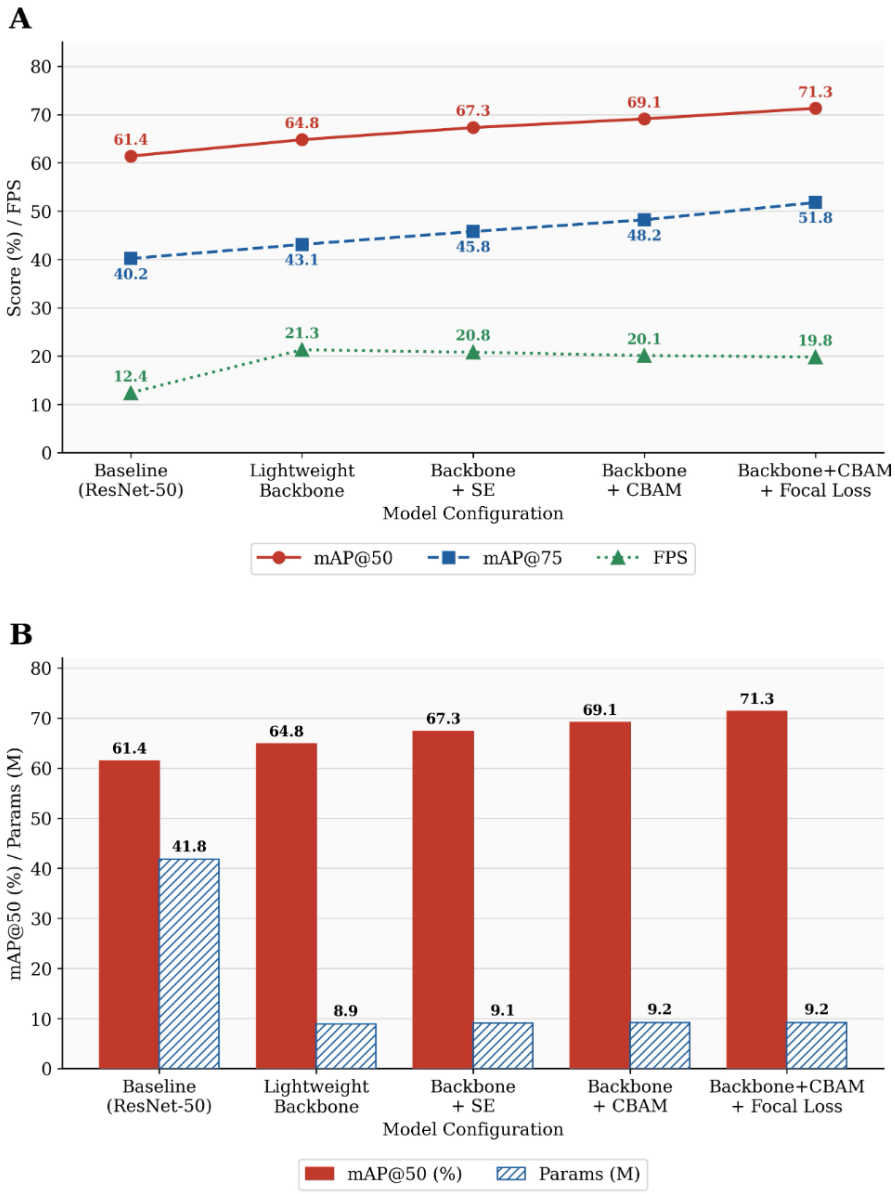

Figure 5. Ablation study: Contribution of each component to detection performance. (A) Ablation: mAP@50, mAP@75, and FPS across configurations. (B) Ablation: mAP@50 vs. model parameters.

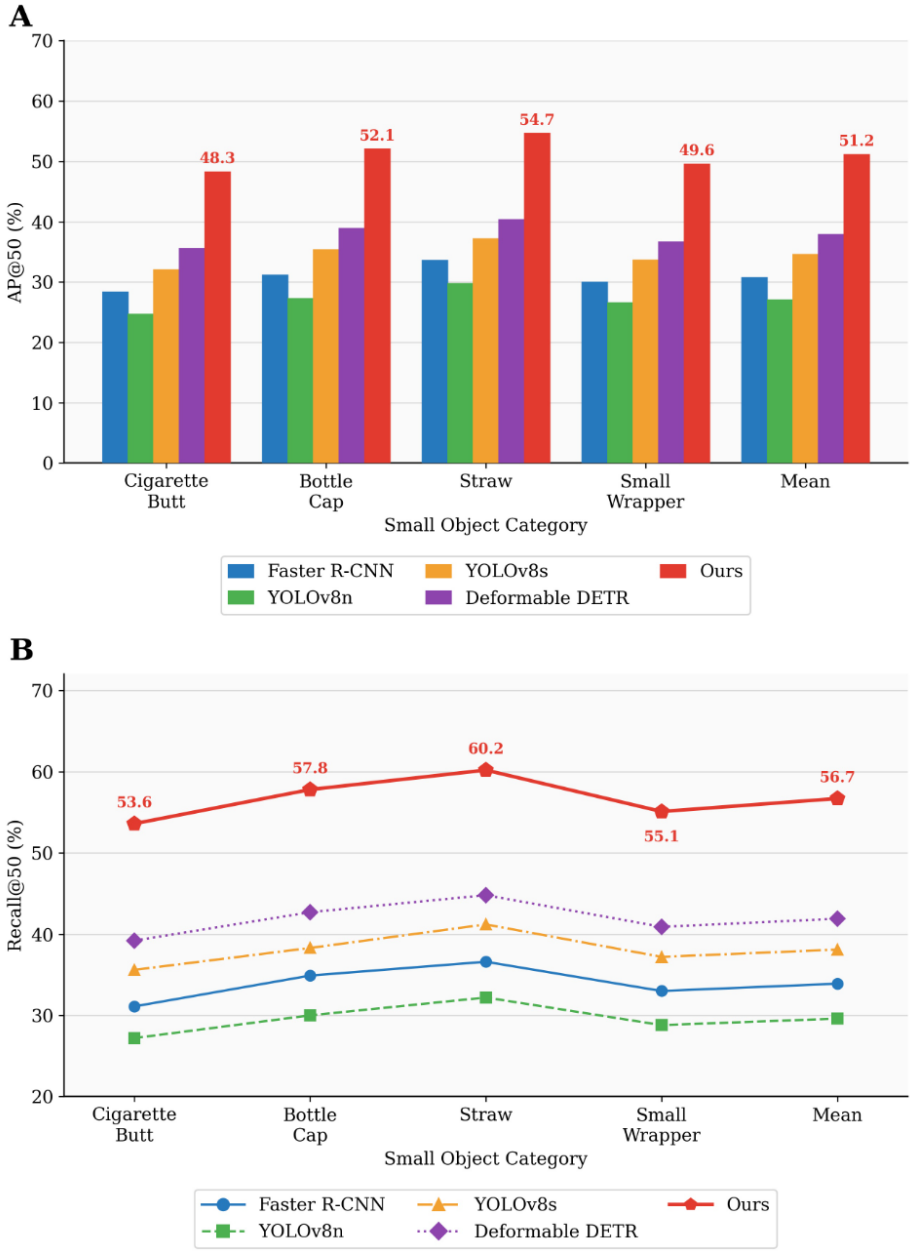

Figure 6. Small-object detection analysis. (A) AP@50. (B) Recall@50.

Table 7. Accuracy–speed–resource comparison of three quantization strategies on the RK3588 platform

Quantization	mAP@50 (%)	mAP@75 (%)	FPS	Size (MB)	Latency (ms)	Power (W)
FP32	71.3	51.8	9.6	35.2	104	4.8
FP16	70.8	51.2	19.8	17.6	51	3.9
INT8	68.4	48.7	35.2	8.8	28	3.1

Note: All tests are executed on the RK3588 NPU with a unified input resolution of 640×640. Power figures represent average full-board consumption measured during inference. INT8 quantization uses post-training quantization with 100 representative calibration images.

8.8 MB (75% smaller than FP32) and the per-frame latency to 28 ms, which corresponds to 35.2 FPS per frame, meeting real-time requirements. Accuracy degrades more: mAP@50 is 68.4%, a 2.9-percentage-point decrease from FP32, and mAP@75 is 48.7%, a 3.1-percentage-point decrease. The degradation is discussed by a layer-wise sensitivity analysis done using the RKNN Toolkit2: the FPN neck layers and the RoIAlign output projection have the lowest activation cosine similarity, 0.91 and 0.89, respectively, when compared to the backbone convolutional layers, which have a cosine similarity of 0.97+ for their activation layers when comparing FP32 vs INT8. Because of the multi-modal activation distributions and broad dynamic range of these layers, the fixed-point approximation of these layers is not that accurate. More fine calibration with 200 representative images gets part of this recovered. All metrics, model size, and latency are monotonically decreasing as quantization precision is lowered, and the difference between adjacent configurations is over 50% for both the size and the latency (Figure 7A and 7B). FP16 is for large-scale deployments at low cost where throughput is a key metric and small loss of accuracy is acceptable, while INT8 is for maximum speed in edge garbage detection.

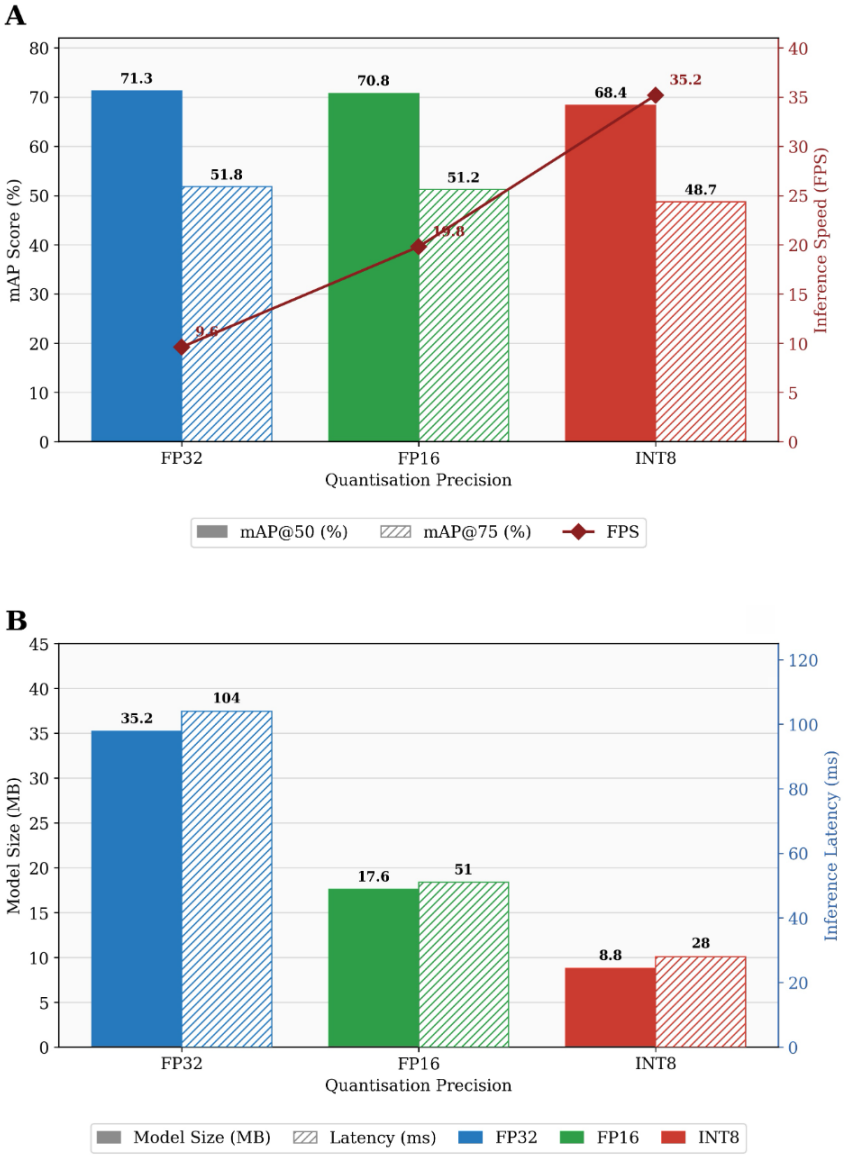

Figure 7. Quantization strategy: Accuracy–speed trade-off on RK3588. (A) mAP@50, mAP@75, and FPS under three quantization levels. (B) Model size and inference latency under three quantization levels.

4.3.2. RK3588 NPU inference benchmark

Systematic benchmarks are performed using both the RK3588 and the NVIDIA Jetson Nano chips to make a multi-dimensional comparison of deployment performance with FP16 and INT8 quantization. There are four metrics: inference frame rate (FPS), per-frame latency (ms), runtime memory consumption (MB), and average full-board power consumption (W) (Table 8).

In the inference speed aspect, the Jetson Nano scores just 11.2 FPS at FP16 (89 ms latency) and 18.6 FPS at INT8 (54 ms latency), both of which are under real time. The RK3588 delivers 19.8 FPS at FP16 (51 ms) and 35.2 FPS at INT8 (28 ms), an 89.2% speed-up over the Jetson Nano's INT8 configuration. This is because the RK3588 NPU has dedicated INT8 units that are not available on Jetson Nano; the INT8 speed-up over FP16 on Jetson Nano is dependent on TensorRT's dynamic quantization path, and therefore lower than the INT8 speed-up over FP16 on the RK3588.

Another notable thing about memory is that it uses it sparingly. The RK3588 is 55% and 58% lower at FP16 and INT8, respectively, than the Jetson Nano. This is important in the case of embedded smart waste bins equipped with limited DRAM. The two also have different power consumption: the RK3588 is 3.9 W for FP16 and 3.1 W for INT8, while the Jetson Nano is 6.8 W and 6.2 W, respectively, a 42.6% and a 50.0% reduction (Marzouk & Elkholy, 2025). The INT8 configuration on the RK3588 operates at 35.2 FPS with 124 MB of memory usage and 3.1 W of power consumption, maintaining good performance in all areas. As shown in Figure 8A and 8B, on all metrics, the RK3588 configurations outperform the Jetson Nano configurations, with the INT8 advantage being the most significant. These results confirm the RK3588 as an excellent hardware platform for scalable, cost-effective edge deployment of intelligent garbage-sorting systems.

Table 8. Inference benchmark comparison: RK3588 vs. Jetson Nano

Platform	Precision	FPS	Latency (ms)	Memory (MB)	Power (W)
Jetson Nano	FP16	11.2	89	412	6.8
Jetson Nano	INT8	18.6	54	298	6.2
RK3588	FP16	19.8	51	186	3.9
RK3588	INT8	35.2	28	124	3.1

Note: All tests use input resolution 640 × 640 with batch size 1. Jetson Nano uses the TensorRT inference engine; RK3588 uses RKNN Toolkit2. Power figures represent net inference power after subtracting idle baseline board consumption.

4.3.3. Power consumption and real-time performance evaluation

The proposed method is evaluated in five representative scenarios: static clean background, waste bin indoors, street scene, highly cluttered environment, and the overall mean, using the RK3588. The four measurements taken are: inference frame rate (FPS), average full-board power consumption (W), Precision@50, and Recall@50 (Table 9; Kulambayev & Olzhayev, 2025).

In the static-background scenario, the feature maps being processed by the NPU are sparse and have very few active areas, resulting in the highest frame rate (38.4 FPS) and the lowest power (2.9 W). As the complexity of the scene increases, so does the frames per second, or fps, drop: When the scene is an indoor waste bin, 35.2 FPS is recorded; when the scene is a street scene, 33.8 FPS; and when the scene is a very cluttered area, 29.6 FPS. More importantly, even the lowest frame rate (29.6 FPS) for the cluttered scenario is above the real-time target (24 FPS), showing good deployment capability in all tested scenarios. The mean of its five scenarios is 34.3 FPS, 20 FPS lower than 50 FPS, which is the demand of real-time and power requirements; it is much lower than the 8–15 W power consumption of mobile GPUs, which highlights the energy efficiency of RK3588 NPU for this application.

The accuracy decreases as a function of scene complexity, but it decreases slowly and steadily. The overall mean Precision@50 and Recall@50 are 69.7% and 55.3%, respectively. The lowest scores (64.2% and 50.8%) come from the cluttered environment, and the highest scores (74.8% and 59.4%) come from the static-background scenario. They perform well in both precision and recall for the scenario-to-scenario range, which is roughly 10.6 points and 8.6 points, respectively. This suggests that this approach does not plateau or fail as the test gets harder. As seen in Figure 9A, the frame rate remains above the 24-FPS reference line in all cases, and the power curve in Figure 9B is below 3.5 W, indicating that throughput and power efficiency do not have to be in opposition. When the real-time throughput (34.3 FPS mean) and low power consumption (3.2 W mean) are considered simultaneously, the proposed approach, under the RK3588 INT8 configuration, is suitable for practical edge applications like smart waste bins and sanitation robots.

5. Discussion

5.1. Effectiveness of the proposed contributions

This work covers three aspects, data fusion, model design, and edge deployment, and experiments back several independent ideas of each design aspect. Specifically, multi-source integration with unified remapping not only

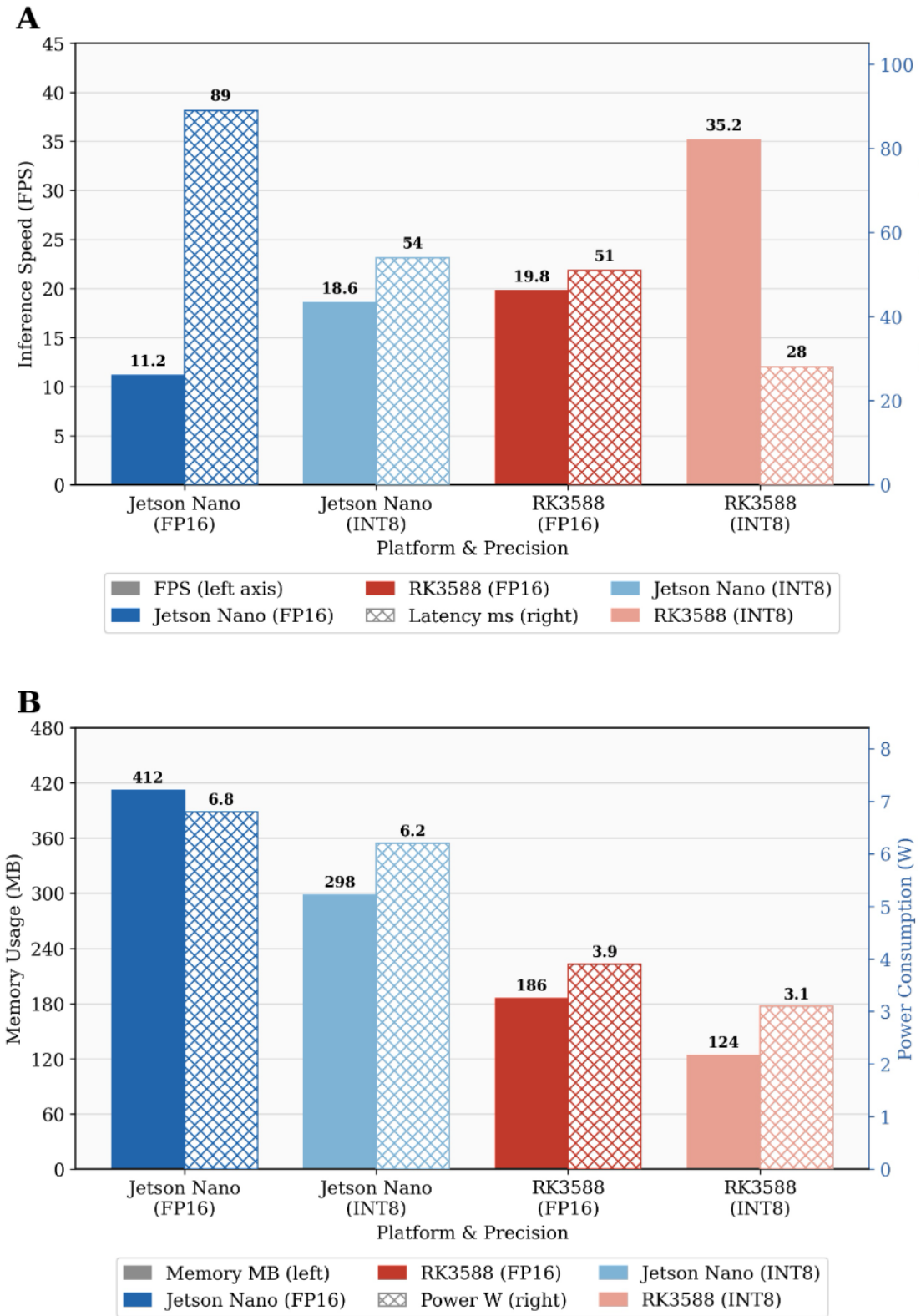

Figure 8. RK3588 NPU inference benchmark vs. Jetson Nano. (A) FPS and latency comparison across platforms. (B) Memory usage and power consumption across platforms.

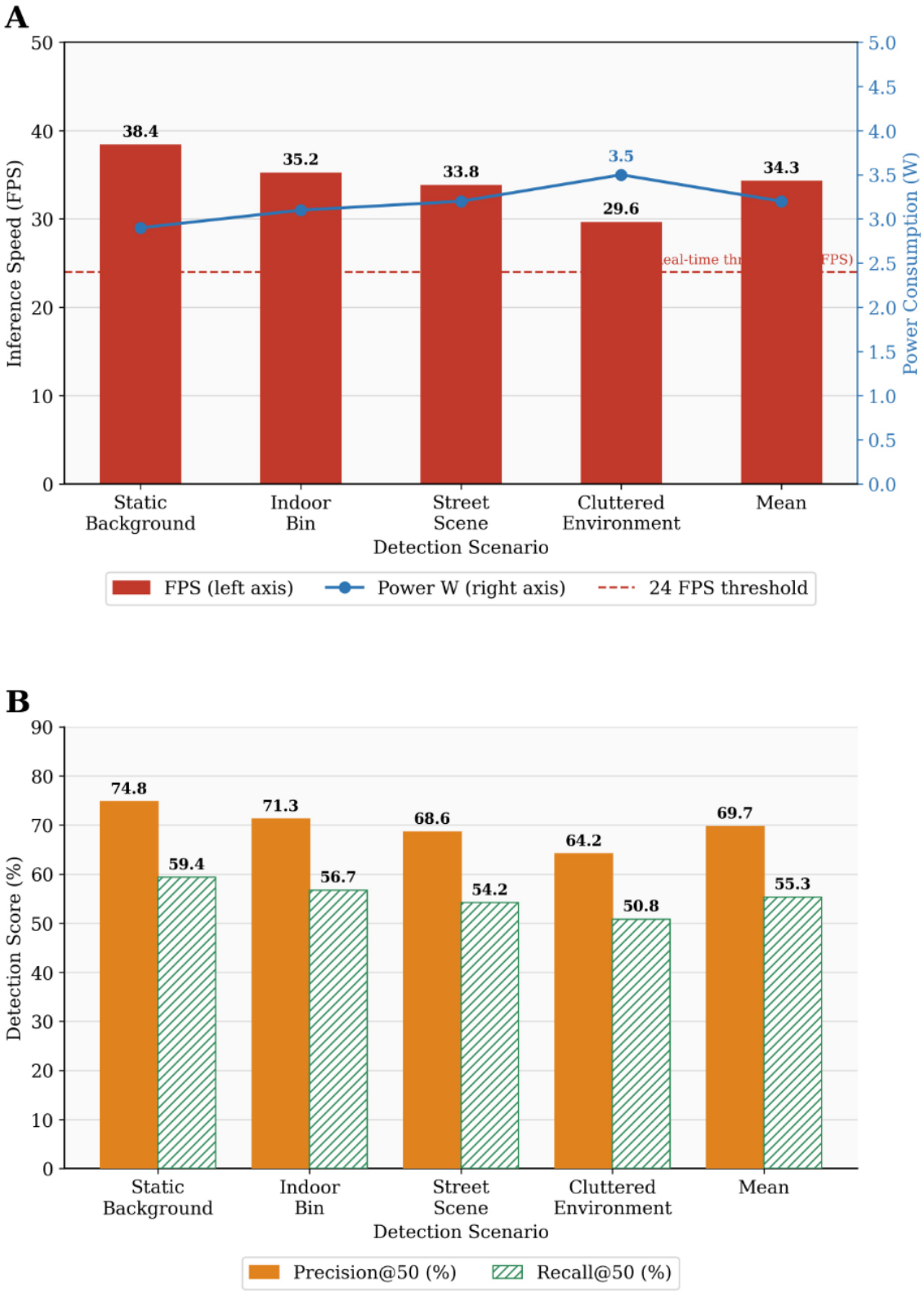

Figure 9. Power consumption and real-time performance across typical waste detection scenarios on RK3588 (INT8). (A) FPS and power consumption across scenarios. (B) Precision@50 and Recall@50 across scenarios.

helps to close the domain gap to 96.8% but also improves ZeroWaste mAP@50 by 15.5 percentage points over single-source training ($t(4) = 18.7$, $p < 0.001$), demonstrating that systematic data governance is an independent prerequisite for cross-domain generalization that cannot be replaced by architectural improvements (Kwon *et al.*, 2024). The remapping protocol, which mapped 60 fine-grained TACO labels and the coarse Real-Waste labels to unified classes, can be used as a methodological model for other multi-source garbage data sets in the future.

Table 9. Power consumption and real-time performance across five scenarios on RK3588 (INT8)

Scenario	FPS	Power (W)	Precision@50 (%)	Recall@50 (%)
Static clean background	38.4	2.9	74.8	59.4
Indoor waste bin	35.2	3.1	71.3	56.7
Street scene	33.8	3.2	68.6	54.2
Highly cluttered environment	29.6	3.5	64.2	50.8
Overall mean	34.3	3.2	69.7	55.3

Note: All tests are conducted on the RK3588 platform under INT8 quantization with input resolution 640 × 640 and batch size 1. Power figures represent average full-board measured consumption. The real-time threshold is defined as ≥ 24 FPS.

The ablation is the quantification of three modifications at the model level with statistical rigor. This is because replacing ResNet-50 with MobileNetV3 discards redundant high-frequency texture information, while retaining the most informative channels for garbage, resulting in a 78.7% reduction in the number of parameters and a 3.4-percentage-point increase in mAP@50 ($t(4) = 8.6$, $p < 0.001$). Despite the addition of only 0.3% of parameters, CBAM increases mAP@50 by 1.8 percentage points ($t(4) = 5.9$, $p < 0.01$) and small-object mean AP@50 by 20.4 percentage points, as the branch of CBAM generates a saliency map per scale that helps remove background clutter from the image, such as asphalt or paper. To tackle a different limitation, focal loss combats low-frequency class weak gradient contributions, which boosts hazardous- and electronic-waste AP@50 by 4.8 and 4.3 percentage points, respectively, when compared to CBAM-only (Dandıl *et al.*, 2024). The strictly monotonic progression of mAP@50 ensures that the three modifications are complementary and not redundant.

Taking the overall picture, mAP@50 has improved by 9.9 percentage points from baseline and shows an improved accuracy–efficiency trade-off. FP16 can increase the speed of inference by about 50%, with only a slight drop of 0.5 percentage point in mAP@50, and maintain the inference velocity at a five-scenario mean of 34.3 FPS at 3.2 W, while INT8 can maintain the inference velocity at a five-scenario mean of 34.3 FPS at 3.2 W while gracefully degrading with increasing scene complexity. Layer-wise sensitivity analysis helps determine the layer causing loss of INT8 accuracy, allowing for targeted and precise recalibration, an approach echoed in other industrial fault-diagnosis deployments (Yang, 2025). Combined, these results demonstrate the engineerability of the proposed algorithm–hardware co-optimization roadmap on RK3588 and a replicable lab-to-real-deployment path.

In addition to waste sorting, the methodological pattern discussed here has been followed by other research efforts that rely on artificial neural networks (ANNs) for perception, forecasting, and control in emerging materials, energy systems, and low-power devices, all of which are characterized by rigorous data governance, attention-based feature enhancement, and quantization-aware edge deployment. In the power and energy sector, for example, lightweight ANN models have been employed in short-term PV power forecasting by Patch Time Series Transformers and multi-modal fusion (Lv *et al.*, 2026) and in bidding optimization strategies of virtual power plants with wind and solar power generation sources under uncertainty through information-gap decision theory and DR optimization (Lei *et al.*, 2026). The cross-domain promise of these advances highlights the data consistency, attention, and INT8-deployment paradigm proven here for garbage detection that can extend to the broader application of ANN-based sensing and decision-making for next-generation energy materials and embedded AI hardware, thus enriching the potential readership and downstream applicability of this study.

5.2. Limitations and future directions

Although the method performs well in terms of accuracy, efficiency, and edge deployment, several limitations remain. The scale is somewhat limited by multi-strategy augmentation: although the range of images is sufficient in this dataset, generalization may be limited in larger-scale deployments because of variations in deformations of plastic bags, color and transparency variations in glass bottles, and changes in outdoor backgrounds with weather and time of day. The other way to overcome the annotation bottleneck is to leverage the vast amount of unlabeled garbage images using semi-supervised learning or self-supervised pre-training (Bejo *et al.*, 2024) and to

gather scene-specific data from authentic smart bins and sanitation robots to narrow the domain gap in cluttered environments.

Sequential channel then spatial processing is possible at the model level and causes a measurable latency at large feature-map resolutions, which is relevant for frame rates well above 35 FPS. Coordinate attention/efficient channel attention may be a parallel or less demanding approximation that does not impair discriminability. There are also noticeable flaws in MobileNetV3's ability to deal with classes that are very similar, such as clear plastic versus glass, paper versus cardboard, which are the most common misclassifications in cluttered scenes; knowledge distillation with soft labels from a larger teacher network might help to refine these fine granularities without increasing the number of parameters used at deployment time, and material-level subcategories (e.g., opaque plastic versus transparent plastic) could offer more supervision for downstream uses related to recycling.

At the deployment level, RK3588 NPU does not natively support all pipeline operators, as it involves engineering overhead and reduces portability to other edge platforms if the RoIAlign bilinear interpolation is decomposed into NPU-friendly operations. Since RKNN Toolkit2 is under development, consider using quantization-aware training instead of PTQ: The goal is to add robustness against quantization error by training the network, rather than calibrating it after training, thus reducing the INT8 mAP@50 loss and potentially even eliminating the need for calibration per layer (Islam *et al*., 2024). Finally, only static image detection is possible so far, while adding the temporal consistency and multi-object tracking for video streams, where temporal dependency is introduced by conveyor belts, rotating bins, and crossing trajectories, is the most critical future direction with joint optimization of the detection accuracy and the continuity of tracking and energy consumption under video conditions as the key objective.

6. Conclusion

In this paper, a lightweight, attention-aware Faster R-CNN framework is presented and evaluated to address the three major problems of garbage object detection: data heterogeneity, model complexity, and deployment, which synergize all three innovations, namely, data-level, model-level, and deployment-level innovations. On the data level, TACO and RealWaste are combined into a unified dataset of 1,000 images with 10 common categories; the data is remapped to a cross-domain benchmark, and the mAP@50 in ZeroWaste is increased by 15.5 percentage points compared with single-source training, demonstrating that systematic data governance is an independent prerequisite for generalization. At the model level, CBAM is incorporated into the FPN neck, and MobileNetV3 (CBAM) replaces ResNet-50, reducing the number of parameters by 78.7%, boosting the full model's mAP@50 by 71.3%, 9.9 percentage points higher than the original model, and all improvements are statistically significant ($p < 0.001$) over five runs. When deployed on RK3588 with RKNN Toolkit2, INT8 PTQ enables an inference speed of 35.2 FPS (89.2% faster and 50.0% more power-efficient than the Jetson Nano) per layer-wise sensitivity analysis, and a five-scenario mean of 34.3 FPS at 3.2 W meets real-time requirements all around. In summary, the findings provide a consistent and statistically valid paradigm for lightweight garbage detection in edge-computing contexts and a potential way for scaling intelligent waste-sorting systems. Future work will add semi-supervised learning to expand the training set of 1,000 images and jointly optimize quantization-aware training and streaming detection and multi-object tracking as the most promising direction.

Acknowledgments

None.

Funding

This work was supported by the Scientific Research Project of the Anhui Provincial Department of Education (Grant nos.: 2025AHGXZK30896 and 2024AH052013).

Conflict of interest

The authors declare they have no competing interests.

Author contributions

Conceptualization: Fei Hu, Sofianita Mutalib
Formal analysis: Fei Hu, Yongli Gao
Investigation: Fei Hu, Yongli Gao
Methodology: Fei Hu, Sofianita Mutalib, Shuzlina Abdul-Rahman
Writing–original draft: Fei Hu
Writing–review & editing: Sofianita Mutalib, Shuzlina Abdul-Rahman, Yongli Gao

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

Data are available from the corresponding author upon reasonable request.

References

Ahmad, Z. S. M., Aziz, A. A. N., Lim, S. H., Hasanuddin, Z. S., Hadi, N. A., Hussin, M. S., & Mohd Zain, N. (2025). Impact of image enhancement using contrast-limited adaptive histogram equalization (CLAHE), anisotropic diffusion, and histogram equalization on spine X-ray segmentation with U-Net, Mask R-CNN, and transfer learning. *Algorithms, 18*(12), Article 796.

https://doi.org/10.3390/a18120796

Akintola, G. A., Obiwusi, Y. K., & Olatunde, O. Y. (2025). Integrated deep learning paradigm for comprehensive lung cancer segmentation and classification using Mask R-CNN and CNN models. *Franklin Open, 11*, Article 100278.

https://doi.org/10.1016/j.fraope.2025.100278

Alfred, R., Leo, J., & Kaijage, F. S. (2025). Detectron2-enhanced Mask R-CNN for precise instance segmentation of rice blast disease in Tanzania: Supporting timely intervention and data-driven severity assessment. *Smart Agricultural Technology, 12*, Article 101301.

https://doi.org/10.1016/j.atech.2025.101301

Amin, J., Gul, N., & Sharif, M. (2025). Dual-method for semantic and instance brain tumor segmentation based on UNet and Mask R-CNN using MRI. *Neural Computing and Applications, 37*(14), 1–19.

https://doi.org/10.1007/s00521-025-11013-y

Arias, R. F. J., Castaño, M. J. M., & Espriella, L. D. M. F. M. (2025). An application of deep learning models for the detection of cocoa pods at different ripening stages: An approach with Faster R-CNN and Mask R-CNN. *Computation, 13*(7), Article 159.

https://doi.org/10.3390/computation13070159

Banik, D. (2025). Enhanced detection and segmentation of ground-glass opacities in SARS-CoV-2 patients using Mask R-CNN on chest CT images. *Multimedia Tools and Applications, 84*(31), 1–16.

https://doi.org/10.1007/s11042-025-20720-6

Bashkirova, D., Abdelfattah, M., Zhu, Z., Sherrill, Z., Calli, B., Shi, B., & Saenko, K. (2022). ZeroWaste dataset: Towards deformable object segmentation in cluttered scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 21298–21307). IEEE.

https://doi.org/10.1109/cvpr52688.2022.02047

Bejo, K. S., Ibrahim, F. M., & Hanafi, M. (2024). Automatic paddy planthopper detection and counting using Faster R-CNN. *Agriculture, 14*(9), Article 1567.

https://doi.org/10.3390/agriculture14091567

Bhalla, S., Kushwaha, R., & Kumar, A. (2025). HydR-CNN: Advancing underwater object detection using a multi-stage framework with hybrid R-CNN and pyramid vision transformer with augmented convolution. *The European Physical Journal Plus, 140*(6), Article 556.

https://doi.org/10.1140/epjp/s13360-025-06498-4

Brandao, I., Fidalgo, B., & González, S. R. (2025). Storm damage and planting success assessment in *Pinus pinaster* Aiton stands using Mask R-CNN. *Forests, 16*(11), Article 1730.

https://doi.org/10.3390/f16111730

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. In A. Vedaldi, H. Bischof, T. Brox, & J.-M. Frahm (Eds.), *Computer Vision – ECCV 2020* (pp. 213–229). Springer.

https://doi.org/10.1007/978-3-030-58452-8_13

Chai, S., Gao, P., & Li, M. (2025). Mask R-CNN for predicting rib fractures on CT images with interpretability and ChatGPT-based structured outcomes. *Expert Systems with Applications, 274*, Article 127047.

https://doi.org/10.1016/j.eswa.2025.127047

Chakraborty, D., Saha, S., & Halder, B. (2025). Learning to localize image forgery using boundary-preserving Mask R-CNN. *Journal of Forensic Sciences, 71*(1), 388–404.

https://doi.org/10.1111/1556-4029.70203

Chansuparp, M., Pansawat, N., & Wangvoralak, S. (2025). Automated grading of boiled shrimp by color level using image processing techniques and Mask R-CNN with feature pyramid networks. *Applied Sciences, 15*(19), Article 10632.

https://doi.org/10.3390/app151910632

Chau, V. T., Jung, S., & Kim, M. (2025). Indirect estimation of seagrass frontal area for coastal protection: A Mask R-CNN and dual-reference approach. *Journal of Marine Science and Engineering, 13*(7), Article 1262.

https://doi.org/10.3390/jmse13071262

Cheng, G., Yuan, X., Yao, X., Yan, K., Zeng, Q., Xie, X., & Han, J. (2023). Towards large-scale small object detection: Survey and benchmarks. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 45*(11), 13467–13488.

https://doi.org/10.1109/tpami.2023.3290594

Crespo, A., Moncada, C., & Crespo, F. (2025). An efficient strawberry segmentation model based on Mask R-CNN and TensorRT. *Artificial Intelligence in Agriculture, 15*(2), 327–337.

https://doi.org/10.1016/j.aiia.2025.01.008

Dandıl, E., Baştuğ, T. B., & Yıldırım, S. M. (2024). MaskAppendix: Backbone-enriched Mask R-CNN based on Grad-CAM for automatic appendix segmentation. *Diagnostics, 14*(21), Article 2346.

https://doi.org/10.3390/diagnostics14212346

David, G., & Faure, E. (2025). End-to-end 3D instance

segmentation of synthetic data and embryo microscopy images with a 3D Mask R-CNN. *Frontiers in Bioinformatics, 4*, Article 1497539.

https://doi.org/10.3389/fbinf.2024.1497539

Dutta, M., Das, K. U., & Datta, N. (2026). Multilingual scene text recognition: A Faster R-CNN approach for Bengali and English scripts. *Expert Systems with Applications, 302*, Article 130482.

https://doi.org/10.1016/j.eswa.2025.130482

Elhosseini, A. M., Sayed, A. H., & Agamy, E. F. R. (2025). A hybrid YOLOv10–Faster R-CNN framework for mobility-aid detection and traffic optimization in disability-inclusive smart cities. *Alexandria Engineering Journal, 129*, 1279–1298.

https://doi.org/10.1016/j.aej.2025.08.044

Gholinavaz, S., Saeedi, N., & Gharehveran, S. S. (2025). Robustness analysis of YOLO and Faster R-CNN for object detection in realistic weather scenarios with noise augmentation. *Scientific Reports, 15*(1), Article 44888.

https://doi.org/10.1038/s41598-025-28737-5

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770–778). IEEE.

https://doi.org/10.1109/cvpr.2016.90

Heda, L., & Sahare, P. (2025). Design of an iterative method for crowd behavior analysis integrating Faster R-CNN, YOLOv8, and graph convolutional networks. *Signal, Image and Video Processing, 19*(7), Article 553.

https://doi.org/10.1007/s11760-025-04127-2

Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q. V., & Adam, H. (2019). Searching for MobileNetV3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 1314–1324). IEEE.

https://doi.org/10.1109/iccv.2019.00140

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 7132–7141). IEEE.

https://doi.org/10.1109/cvpr.2018.00745

Huang, R., Ding, J., & Ren, Z. (2025). Research on the segmentation of individual trees and the extraction of structural parameters in eucalyptus plantations based on a TEMA Mask R-CNN model. *Journal of Forest Research, 30*(4), 316–327.

https://doi.org/10.1080/13416979.2025.2479873

Islam, T., Sarker, T. T., & Ahmed, R. K. (2024). Detection and classification of cannabis seeds using RetinaNet and Faster R-CNN. *Seeds, 3*(3), 456–478.

https://doi.org/10.3390/seeds3030031

Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., Adam, H., & Kalenichenko, D. (2018). Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2704–2713). IEEE.

https://doi.org/10.1109/cvpr.2018.00286

Kaviani, M., Leblon, B., & Akilan, T. (2025). Tree health assessment using Mask R-CNN on UAV multi-spectral imagery over apple orchards. *Remote Sensing, 17*(19), Article 3369.

https://doi.org/10.3390/rs17193369

Kriouile, Y., Ancourt, C., & Wolska, W. K. (2024). Nested object detection using Mask R-CNN: Application to bee and varroa detection. *Neural Computing and Applications, 36*(35), 1–23.

https://doi.org/10.1007/s00521-024-10393-x

Kulambayev, B., & Olzhayev, O. (2025). A Mask R-CNN algorithm for automated segmentation of asphalt road cracks. *Procedia Computer Science, 269*, 39–48.

https://doi.org/10.1016/j.procs.2025.08.257

Kwon, K., Im, K. S., & Kim, Y. S. (2024). Estimation of tree diameter at breast height from aerial photographs using a Mask R-CNN and Bayesian regression. *Forests, 15*(11), Article 1881.

https://doi.org/10.3390/f15111881

Lei, M., Zhang, M., & Wang, K. (2026). Research on bidding optimization strategy for virtual power plants with wind–solar–storage systems based on IGDT–DRO. *Electrical Engineering, 108*(3), Article 208.

https://doi.org/10.1007/s00202-026-03544-x

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017a). Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 936–944). IEEE.

https://doi.org/10.1109/cvpr.2017.106

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017b). Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 2980–2988). IEEE.

https://doi.org/10.1109/iccv.2017.324

Lv, X., Cui, S., Wang, Y., Lu, J., Yu, P., & Wang, K. (2026). Patch time series transformer–based short-term photovoltaic power prediction enhanced by artificial fish. *Energies, 19*(1), Article 284.

https://doi.org/10.3390/en19010284

Marzouk, A. M., & Elkholy, M. (2025). Tracking and estimating

the speed of vehicles using an enhanced R-CNN and Deep SORT architecture. *Journal of Advances in Information Technology, 16*(2), 204–214.

https://doi.org/10.12720/jait.16.2.204-214

Nazir, A., & Wani, A. M. (2025). Multi-scale feature enhancement using EfficientNet-B7 and PANet in Faster R-CNN for small object detection. *International Journal of Information Technology, 18*(1), 1–8.

https://doi.org/10.1007/s41870-025-02790-9

Paściak, A., Piwowarczyk, K. P., & Iskander, R. D. (2025). Enhancing meibography based assessment of gland morphology by utilizing an image-rotating Mask R-CNN approach. *Biomedical Signal Processing and Control, 109*, Article 108045.

https://doi.org/10.1016/j.bspc.2025.108045

Prasad, R., Saxena, K., & Maurya, A. (2025). DermoAI-CNN: Leveraging GANs, Mask R-CNN, and attention mechanisms for enhanced skin disease analysis. *Foundations of Computing and Decision Sciences, 50*(4), 537–569.

https://doi.org/10.2478/fcds-2025-0021

Proença, P. F., & Simões, P. (2020). *TACO: Trash Annotations in Context for Litter Detection* (Version 2). arXiv.

https://doi.org/10.48550/ARXIV.2003.06975

Rajendran, V., Subramaniyan, M. U., & Suriyan, K. (2025). Enhanced R-CNN model for traffic sign recognition under diverse environmental conditions. *International Journal of Intelligent Transportation Systems Research, 23*(2), 1–18.

https://doi.org/10.1007/s13177-025-00476-x

Reddy, N. T., Kumar, N., & Ponnappa, P. N. (2025). Intelligent GD&T symbol detection in mechanical drawings: A comparative study of YOLOv11, Faster R-CNN, and RetinaNet for quality assurance. *Journal of Intelligent Manufacturing*. Advance online publication.

https://doi.org/10.1007/s10845-025-02669-3

Reddy, A. S., & G., M. P. (2025). Skin cancer detection using optimized Mask R-CNN and two-fold deep learning classifier framework. *Multimedia Tools and Applications, 84*(30), 1–28.

https://doi.org/10.1007/s11042-024-20377-7

Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39*(6), 1137–1149.

https://doi.org/10.1109/tpami.2016.2577031

Rifat, S. M. K., Nayeem, J. M., Neon, M. N. M., & Nur-A-Alam, M. (2026). A deep learning-based performance-oriented approach for diabetic foot ulcer identification and localization using Faster R-CNN. *In Silico Research in Biomedicine*, Article 100202.

https://doi.org/10.1016/j.insi.2026.100202

Ritha, N., Hayaty, N., & Apdillah, D. (2025). Deep object detection approaches for identification of seagrass species using Faster R-CNN and SSD. *IOP Conference Series: Earth and Environmental Science, 1559*(1), Article 012012.

https://doi.org/10.1088/1755-1315/1559/1/012012

Salvador, F. A. C., Bilyk, T., & Dartois, A. (2025). High-throughput analysis of dislocation loops in irradiated metals using Mask R-CNN. *Micron, 201*, Article 103927.

https://doi.org/10.1016/j.micron.2025.103927

Shobaki, A. W., & Milanova, M. (2025). A comparative study of YOLO, SSD, Faster R-CNN, and more for optimized eye-gaze writing. *Sci, 7*(2), Article 47.

https://doi.org/10.3390/sci7020047

Single, S., Iranmanesh, S., & Raad, R. (2023). RealWaste: A novel real-life data set for landfill waste classification using deep learning. *Information, 14*(12), Article 633.

https://doi.org/10.3390/info14120633

Sinha, M., Paul, S., & Lala, N. G. M. (2025). Comparative analysis of Mask R-CNN and U-Net architectures using ResNet as backbone for lunar crater detection. *Planetary and Space Science, 264*, Article 106140.

https://doi.org/10.1016/j.pss.2025.106140

Sun, S., Zheng, S., Xu, X., & He, Z. (2025). GD-YOLO: A lightweight model for household waste image detection. *Expert Systems with Applications, 279*, Article 127525.

https://doi.org/10.1016/j.eswa.2025.127525

Wang, B., Liu, D., & Wu, J. (2025). SSNFNet: An enhanced few-shot learning model for efficient poultry farming detection. *Animals, 15*(15), Article 2252.

https://doi.org/10.3390/ani15152252

Wang, D., Shi, L., & Li, Y. (2025). An enhanced Faster R-CNN for high-throughput winter wheat spike monitoring to improve yield prediction and water use efficiency. *Agronomy, 15*(10), Article 2388.

https://doi.org/10.3390/agronomy15102388

Wang, Z. W., Sun, R., & Wang, Z. (2025). Jiyu Faster R-CNN de maolian zitai shibie [Anchor chain attitude recognition based on Faster R-CNN]. *Zhongguo Jixie [China Machinery]*, (28), 33–37. [In Chinese]

Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. (2018). CBAM: Convolutional block attention module. In V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss (Eds.), *Computer Vision – ECCV 2018* (pp. 3–19). Springer.

https://doi.org/10.1007/978-3-030-01234-2_1

Xie, R. J., & Ren, R. X. (2025). Jiyu Faster R-CNN de jianzhuwu yaogan tuxiang mubiao jiance [Remote sensing image object detection of buildings based on Faster R-CNN]. *Xinxi Jilu*

Cailiao [Information Recording Materials], 26(9), 113–115. [In Chinese]

https://doi.org/10.16009/j.cnki.cn13-1295/tq.2025.09.067

Yang, W. (2025). Jiyu gaijin Mask R-CNN de dianli shebei rege guzhang zhenduan yanjiu [Research on thermal fault diagnosis of power equipment based on improved Mask R-CNN]. *Jiamusi Daxue Xuebao (Ziran Kexue Ban) [Journal of Jiamusi University (Natural Science Edition)], 43*(7), 61–64. [In Chinese]

https://doi.org/10.20232/j.cnki.jmsdxxb.2025.07.034

Zhu, X., Su, W., Lu, L., Li, B., Wang, X., & Dai, J. (2021). Deformable DETR: Deformable transformers for end-to-end object detection. In *9th International Conference on Learning Representations (ICLR 2021)*. OpenReview. https://openreview.net/forum?id=gZ9hCDWe6ke

Zou, Z., Chen, K., Shi, Z., Guo, Y., & Ye, J. (2023). Object detection in 20 years: A survey. *Proceedings of the IEEE, 111*(3), 257–276.

https://doi.org/10.1109/jproc.2023.3238524