Bulletin of Electrical Engineering and Informatics
Vol. 15, No. 3, June 2026, pp. 2191∼2211
ISSN: 2302-9285, DOI: 10.11591/eei.v15i3.10696

Object detection for waste management: a comparative review of models, challenges, and future directions

Owen Tamin[1], Ervin Gubin Moung[2,3], Ali Farzamnia[4]
[1]Faculty of Science and Technology, Universiti Malaysia Sabah, Kota Kinabalu, Malaysia
[2]Data Technologies and Applications (DaTA) Research Group, Faculty of Computing and Informatics, Universiti Malaysia Sabah, Kota Kinabalu, Malaysia
[3]Faculty of Computing and Informatics, Universiti Malaysia Sabah, Kota Kinabalu, Malaysia
[4]School of Computing and Engineering, University of Huddersfield, Huddersfield, United Kingdom

Article Info

Article history:

Received May 16, 2025
Revised Nov 21, 2025
Accepted Mar 31, 2026

Keywords:

Deep learning
Object detection
Region-based convolutional neural networks
Sustainability
Transformer models
You only look once
Waste detection

ABSTRACT

Despite growing interest in automated waste detection, existing surveys either focus on a narrow set of models or lack systematic comparisons across object detection paradigms. This review addresses that gap by examining recent advances in deep learning for waste management, spanning two-stage detectors (Faster region-based convolutional neural network (Faster R-CNN) and Mask region-based convolutional neural network (Mask R-CNN)), single-shot frameworks (you only look once version 1 (YOLO)v1 to YOLOv11), and emerging Transformer-based models (ViT-WM and AL-DETR). Faster R-CNN achieved category-level accuracy of 91.68% and overall accuracy of 89.68%, while Mask R-CNN reported AP values between 26.2% and 34.5% across varied datasets. YOLO models demonstrated strong real-time capability, with YOLOv5 reaching a mAP@0.5 of 92.96% and YOLOv8 achieving 97.63% accuracy with precision and recall above 93%. Transformer-based approaches are especially promising: ViT-WM achieved 98.17% accuracy, the highest among reviewed models, and AL-DETR reported a mAP of 58.9% while integrating active learning (AL) strategies to reduce reliance on extensive labeled data. These results emphasize YOLO's efficiency for real-time waste sorting and the potential of Transformer architectures for handling complex, cluttered environments. Remaining challenges include dataset variability, computational demand, and limited standardized benchmarks. Future research should prioritize developing comprehensive datasets, optimizing Transformers for real-time use, and leveraging AL to enhance generalizability with reduced annotation effort.

Corresponding Author:

Ervin Gubin Moung
Data Technologies and Applications (DaTA) Research Group, Faculty of Computing and Informatics
Universiti Malaysia Sabah
Jalan UMS, Kota Kinabalu 88400, Sabah, Malaysia
Email: ervin@ums.edu.my

1. INTRODUCTION

Waste management has been a persistent global challenge, as most human activities inherently generate waste [1]. Over time, the rate and diversity of waste production have accelerated due to population growth, urbanization, and industrial activities [2]–[4]. However, poor waste management remains a pressing issue,

primarily attributed to a lack of proper planning and ineffective operational practices [5], [6].

Historically, waste management relied on manual sorting and basic mechanical processes, which were labor-intensive and often inaccurate [7]. Traditional methods, such as manual separation and conveyor belt systems, struggled to handle the growing diversity and volume of waste materials. These techniques were not only slow but also lacked the precision needed to effectively categorize and recycle waste materials. As waste generation continued to increase, these methods became less sustainable and efficient. This gap led to the development of more automated and reliable systems, powered by advances in deep learning and computer vision technologies, offering a more efficient alternative for waste detection and sorting [8].

The growing complexity and volume of waste materials, ranging from plastics to organic and electronic waste, necessitate more efficient and automated waste detection systems [9]–[11]. As a result, there has been a notable shift towards deep learning-based solutions in waste management. These technologies have revolutionized the ability to detect and sort waste materials with high speed and accuracy, overcoming the limitations of manual and mechanical sorting methods.

Global waste generation is expected to increase by 70% by 2050, reaching 3.4 billion metric tons annually, with plastics accounting for a significant portion [12]. Improper waste management leads to severe environmental impacts, such as pollution of oceans and landfills, posing risks to biodiversity and human health. To address this crisis, artificial intelligence (AI)-driven approaches have become integral in modern waste management strategies, helping optimize sorting, recycling, and disposal processes [8], [13]. AI-powered solutions, including machine learning and deep learning, have shown promise in automating classification and enhancing accuracy in waste detection.

Among these AI-powered methods, object detection technologies have gained particular attention for their effectiveness in identifying, classifying, and sorting waste materials with high precision and speed [14]. Much like face recognition systems that classify facial features for authentication, object detection models in waste management use advanced algorithms to differentiate waste based on material, shape, and other visual characteristics. In recent years, significant advancements have been made in object detection models, such as you only look once (YOLO) [15], Faster region-based convolutional neural networks (Faster R-CNN) [16], and Transformer-based architectures [17]. These models have demonstrated high accuracy and efficiency in various domains, and their application to waste detection has the potential to revolutionize waste management practices.

Despite advances in object detection, their application to waste detection remains fragmented and under-validated, with heterogeneous datasets, inconsistent metrics, and few deployment-grade evaluations. While numerous surveys have explored applications of object detection for waste management, most either focus primarily on traditional models or do not fully analyze recent advancements, such as Transformer-based architectures and active learning (AL) techniques [18]–[21]. Although newer architectures have begun to appear in waste detection research [22], [23], there remains a significant gap regarding a systematic and comparative evaluation of these emerging methods against traditional, established models. To effectively understand the evolution and impact of these new technologies, it is essential not only to examine their individual strengths but also to critically assess how they perform relative to earlier models such as R-CNN and YOLO variants. Addressing this gap through comparative synthesis can guide researchers and practitioners in selecting the most suitable methodologies for specific waste detection scenarios and outline clear future directions for integrating advanced architectures into practical, scalable waste management solutions.

This review aims to bridge that gap by providing an up-to-date overview of the literature, covering developments from 2019 to 2024, with a focus on the growing role of Transformer-based models and their potential for advancing waste management practices. The primary objectives of this review are: i) to review the latest object detection applications in waste management; ii) to assess the performance of various models, including newer approaches like Transformers, across different waste types; and iii) to explore future research opportunities and trends in this rapidly evolving field.

The contributions of this review are:

a. Unifies two-stage (Faster/Mask R-CNN), one-stage (YOLOv1–v11), and Transformer (DETR/AL-DETR/ViT-WM) results under standard metrics evaluation, reporting YOLOv8 accuracy 97.63%, YOLOv5 mAP@0.5 92.96%, ViT-WM accuracy 98.17%, AL-DETR mAP 58.9%, and 23–28 FPS as a benchmark of real-time suitability.

b. Introduces a taxonomy linking architecture family, supervision, modality, and deployment target to accuracy–latency–power trade-offs and maps all 12 studies to it.

c. Identifies five gaps which are imbalance, occluded objects, temporal robustness, edge power–latency limits, and reproducibility that pairs each with targeted remedies (augmentation, small-object priors, temporal ensembles, quantization, standardized reporting).

This paper is organized into five sections. Section 1 discusses prior research, the motivation behind the study, and its contributions to the field. Section 2 presents the research methodology and provides an overview of recent object detection models in waste detection. Section 3 presents the results and discussions. Section 4 outlines the key challenges and research gaps identified in this study, and also provides recommendations for future research directions. Finally, section 5 presents the main conclusions of the study.

2. METHOD

This review is inspired by the general principles of the preferred reporting items for systematic reviews and meta-Analyses (PRISMA) framework to ensure transparency and reproducibility. It is structured around three key research questions: i) What are the most recent object detection models applied to waste detection, and how have they evolved from 2019 to 2024? ii) How do these models perform in detecting and classifying various types of waste, particularly in complex and cluttered environments? and iii) What are the key challenges and future research opportunities in this field? To address these questions, a comprehensive search strategy was applied using IEEE Xplore, Springer, Scopus, and Science Direct. The initial query was "object detection model in waste detection," limited to publications from 2019 to 2024, and refined with related terms such as "object detection for waste," "machine learning in waste detection," and "deep learning in waste detection." The process was then followed by removing duplicates, screening titles and abstracts, and conducting full-text reviews to assess methodological rigor, empirical contributions, and alignment with the study scope. Only peer-reviewed articles published in English with sufficient methodological detail were included, while studies lacking empirical evidence, scientific rigor, or presenting conflicts of interest were excluded. To minimize publication bias, articles were drawn from a wide range of journals and conferences. The overall review selection process is summarized in the flow diagram presented in Figure 1.

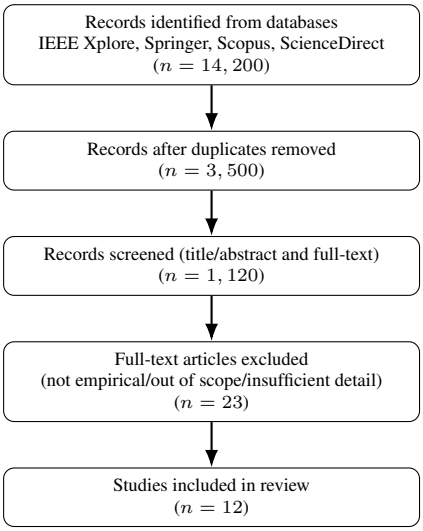

Figure 1. Workflow of article identification and selection in the systematic review

2.1. Review of object detection models in waste detection

This section provides an overview of key object detection models, including region-based convolutional neural networks (R-CNN), single shot detectors (SSD), and Transformer-based models, which have been extensively utilized in waste detection tasks. The conceptual taxonomy and timeline of these models is shown in Figures 2 and 3. The strengths, limitations, and critical analysis of these models in addressing the challenges of waste detection are also discussed.

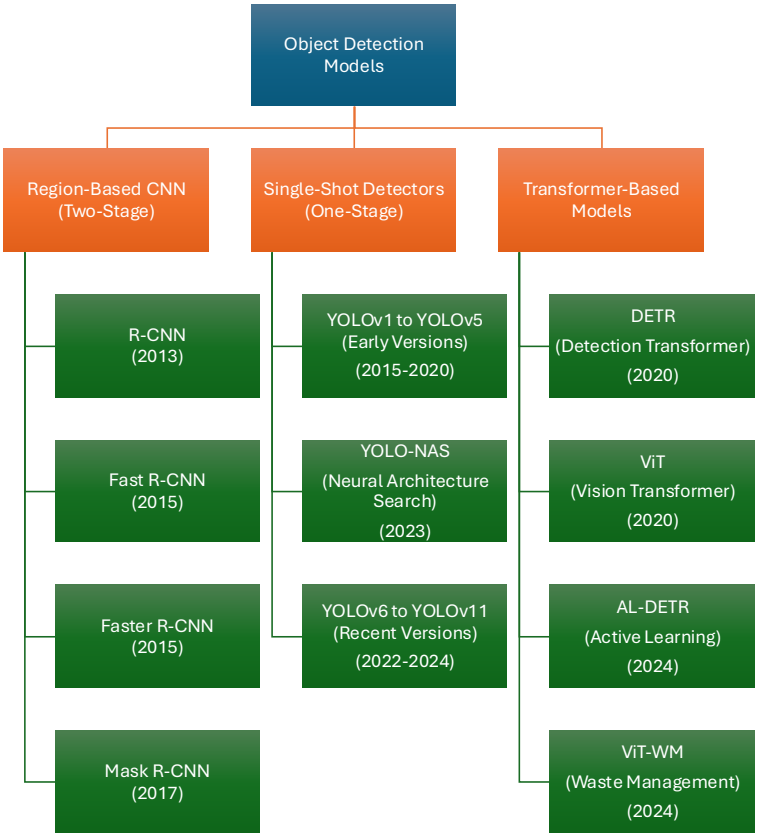

Figure 2. Taxonomy of object detection models

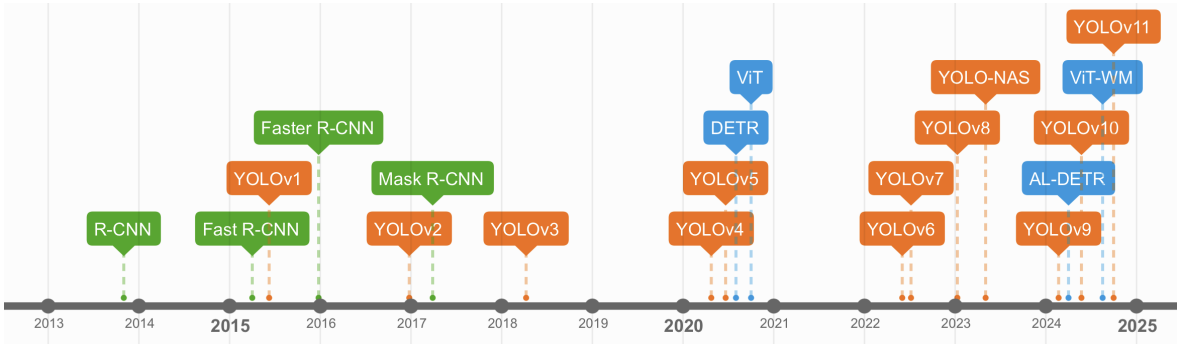

Figure 3. Object detection timeline across R-CNN, YOLO, and Transformer families (2013–2024)

2.2. Region-based convolutional neural networks

R-CNN is a foundational two-stage object detection algorithm that combines selective search for region proposals with deep convolutional neural networks (CNNs) for object classification and bounding box refinement [24]. The R-CNN framework operates through three main components. First, region proposals are generated using a selective search algorithm that identifies approximately 2,000 rectangular bounding boxes as potential object locations within an image. Second, each region proposal is normalized to a fixed size and processed through a CNN to extract meaningful features that capture the visual characteristics of the region. Finally, the extracted features are used in the classification and localization stage, where object classes are predicted, and bounding box coordinates are refined to ensure precise localization of detected objects [25].

2.2.1. Advancements in region-based convolutional neural networks variants

The R-CNN framework has evolved through several notable variants, each addressing specific limitations and extending its functionality to broader applications.

- Fast region-based convolutional neural network (Fast R-CNN): Fast R-CNN significantly improves the computational efficiency of the original R-CNN by sharing convolutional features across all region proposals. Instead of processing each region proposal through the CNN independently, Fast R-CNN computes feature maps for the entire image once. Region-specific features are then extracted using a region of interest (ROI) pooling layer, which reduces redundancy and accelerates the object detection pipeline [26].
- Faster R-CNN: Faster R-CNN further optimizes the object detection process by integrating region proposal generation directly into the CNN architecture. This is achieved through a region proposal network (RPN) that shares convolutional features with the detection network. The RPN simultaneously generates high-quality region proposals and performs object classification, making the model highly efficient and effective for large-scale detection tasks [16].
- Mask region-based convolutional neural network (Mask R-CNN): building upon Faster R-CNN, Mask R-CNN extends the capabilities of the framework to instance segmentation. While Faster R-CNN excels in object detection, Mask R-CNN adds a branch dedicated to predicting pixel-level object masks, enabling precise segmentation of individual objects within an image. Additionally, it introduces improvements in region alignment, ensuring higher accuracy in bounding box and mask predictions [27].

2.2.2. Recent waste detection by region-based convolutional neural networks family

A study on intelligent garbage sorting utilized Faster R-CNN with ResNet50, achieving 89.68% overall accuracy in garbage recognition, where the model was trained on 3,984 images and tested on 3,552. The model classified 23 types of recyclable waste, 6 types of other waste, 4 types of hazardous waste, and 8 types of kitchen waste, providing valuable data for automated waste sorting in diverse environments [28].

The TACO dataset, designed for litter detection and segmentation, uses Mask R-CNN for instance segmentation. On the current version of the dataset, Mask R-CNN achieved an average precision (AP) of 26.2%±1.0% for the TACO_1 dataset and 18.4%±1.5% for the TACO_10 dataset. This dataset, containing 1,500 images and 4,784 annotations, still requires more manual annotations to improve deployment effectiveness in real-world scenarios [29].

In another study, Faster R-CNN and Mask R-CNN with a ResNeXt-101-FPN backbone were employed to detect various waste categories in the COCO dataset. Faster R-CNN outperformed Mask R-CNN in object detection, achieving an AP of 34.5% for instance detection, while Mask R-CNN achieved 30.0% for instance segmentation and 28.2% for material detection [30]. This comparison demonstrates the versatility and utility of R-CNN models in detecting waste, even in complex and cluttered environments.

Mask R-CNN has also been successfully applied to plastic waste sorting using "The Open AI Dataset Project," where it achieved an accuracy of 91.2% and a mean average precision (mAP) of 91.1%. The model excels in tasks requiring precise segmentation, making it a reliable option for applications prioritizing segmentation accuracy over speed [31].

Additionally, Mask R-CNN has been tested on the ZeroWaste dataset, achieving an mAP of 22.8% and a processing speed of 23 frame per second (FPS). While it was slower than fine-tuned DETR models in terms of speed and scalability, Mask R-CNN still proved valuable for processing complex scenes and providing accurate predictions, especially in cases where real-time performance is less critical [32]. Table 1 presents recent studies on waste detection using R-CNN family models, highlighting datasets, model variants, and application focus. Key insights include performance trade-offs, dataset challenges, and suitability for real-world waste sorting tasks.

2.2.3. Critical analysis on region-based convolutional neural networks

While R-CNN and its variants, such as Fast R-CNN and Faster R-CNN, have been fundamental in advancing object detection tasks, they are not without limitations. The original R-CNN model suffers from high computational costs due to the independent feature extraction for each region proposal. Even though Fast R-CNN addresses this by sharing convolutional features, it still requires significant processing time for complex images.

Faster R-CNN offers considerable improvements by introducing the RPN, which eliminates the reliance on external region proposal methods. However, the performance of Faster R-CNN can degrade in real-

time applications due to its slower inference speed, particularly in high-density environments like waste detection. Furthermore, the architecture can struggle with detecting small or occluded objects, a common challenge in waste detection tasks. On the other hand, Mask R-CNN, deliver high accuracy and detailed segmentation, but this comes at the cost of slower inference times (200–350 ms per image) and greater computational demands, making them less practical for real-time conveyor belt systems. Instead, these models are better suited for controlled environments such as lab-based waste analysis or smart-bin systems.

Table 1. Recent waste detection studies using R-CNN family models

Ref.	Dataset	Model	Application focus	Implications and insights
[28]	Custom	Faster R-CNN (ResNet50)	Multi-category classification (recyclable, hazardous, kitchen, and other)	Demonstrated generalization across categories. Ideal for large-scale sorting in diverse environments.
[29]	TACO	Mask R-CNN	Litter detection and instance segmentation	Lower precision indicated dataset inadequacy; highlighted need for better and balanced annotations.
[30]	COCO	Faster R-CNN, Mask R-CNN (ResNeXt-101-FPN)	Comparative between detection and segmentation	Faster R-CNN showed better detection than Mask R-CNN in complex scenes. Trade-offs between accuracy and segmentation.
[31]	Open AI Dataset project	Mask R-CNN	Plastic waste sorting (segmentation focused)	High segmentation accuracy made it ideal for sorting tasks where speed is secondary.
[32]	ZeroWaste	Mask R-CNN	Waste detection in cluttered environments	Moderate accuracy and slower speed. Still useful where detailed scene understanding matters more than real-time performance.

Although Mask R-CNN enhances precision through pixel-level segmentation, the rationale behind its boundary delineations remains opaque. Visualization techniques such as saliency maps, Grad-CAM, and class activation mapping (CAM) offer partial insights by highlighting image regions influencing model decisions, yet these methods provide limited interpretability. This lack of transparency becomes especially problematic in real-time applications like waste detection systems, where quick and reliable decisions are essential but may be hindered by the computational demands of interpretability tools.

Given the interpretability challenges and high computational cost of R-CNN-based methods, there is a growing demand for faster and more efficient alternatives. This need has driven the development of single-shot detectors, which streamline the detection pipeline by eliminating the region proposal stage. These models offer a compelling solution for real-time waste detection in dynamic environments.

2.3. Single-shot detectors

The SSD is a feed-forward convolutional network designed for efficient object detection. It generates a fixed-size collection of bounding boxes and corresponding scores to indicate the presence of object class instances within those boxes. The final detections are produced after a non-maximum suppression step. SSD employs a base network, typically a truncated version of a high-quality image classification architecture, to extract feature maps from the input image. Unlike region-based detectors that rely on region proposals, SSD directly predicts bounding boxes and class scores from feature maps, streamlining the detection process [33].

During training, SSD uses an input image and ground truth bounding boxes for each object. The unique training process involves assigning ground truth information to specific outputs within a predefined set of default bounding boxes. These default boxes vary by location, aspect ratio, and scale, enabling the model to handle objects of diverse shapes and sizes. This approach eliminates the need for region proposal generation and allows for real-time object detection with high accuracy.

2.3.1. You only look once

The YOLO family of object detection algorithms offers a unified framework that directly predicts bounding boxes and class probabilities from the entire input image in a single evaluation. YOLO divides the image into a grid and assigns each grid cell the responsibility of detecting objects within its bounds. By formulating detection as a regression problem, YOLO achieves exceptional speed while maintaining competitive accuracy.

Over the years, YOLO has evolved through numerous versions, each introducing enhancements to improve performance, accuracy, and ease of use [34]. Below is a brief summary of key developments in each

YOLO version:

a. YOLOv1 (2015): unified object detection by simultaneously predicting bounding boxes. The input image is divided into an $S \times S$ grid, with each cell predicting B bounding boxes and C class probabilities. Despite its speed, it suffered from localization errors and limited multi-object detection [35].
b. YOLOv2 (2016): introduced anchor boxes, batch normalization, high-resolution classifiers, dimension clusters, and multi-scale training. It significantly improved accuracy while maintaining speed, enabling detection of up to 9000 categories [36].
c. YOLOv3 (2018): added multi-scale predictions, a new backbone (Darknet-53), and logistic regression for objectness scores. It also adopted binary cross-entropy for classification, allowing multi-label predictions [37].
d. YOLOv4 (2020): introduced a CSPDarknet-53 backbone, PANet for feature aggregation, and a suite of training enhancements like mosaic augmentation, self-adversarial training (SAT), and CIoU loss. It balanced accuracy and speed effectively [38].
e. YOLOv5 (2020): developed by Ultralytics in PyTorch, it leveraged YOLOv4 enhancements and added tools for ease of deployment. It became popular due to its seamless integration with modern machine learning workflows [39].
f. YOLOv6 (2022): optimized for industrial use, YOLOv6 introduced an efficient architecture and training pipeline, focusing on real-world deployment scenarios [40].
g. YOLOv7 (2022): enhanced feature optimization and model efficiency, making it one of the fastest object detectors with high accuracy across diverse datasets [41].
h. YOLOv8 (2023): further refined with an emphasis on modular design, ease of customization, and improved performance in edge and cloud applications [42].
i. YOLO-NAS (2023): the latest iteration, integrating neural architecture search (NAS) for automated architecture optimization, achieving state-of-the-art accuracy and efficiency [43].
j. YOLOv9 (2024): introduced the RT-DETR-based architecture, fusing Transformer-based decoding with YOLO's real-time efficiency. YOLOv9 achieved significant performance gains with better feature alignment and dynamic anchor assignment [44].
k. YOLOv10 (2024): unified the detection, segmentation, and pose estimation tasks in a single model. YOLOv10 emphasized lightweight performance and scalability for both mobile and edge devices while maintaining competitive accuracy across all tasks [45].
l. YOLOv11 (2024): focused on further reducing latency with an upgraded backbone and enhanced feature pyramid network. YOLOv11 integrated quantization-aware training and advanced knowledge distillation, setting new benchmarks for ultra-fast inference in low-power environments [46].

2.3.2. Recent waste detection by you only look once family

One study evaluates the use of YOLO-v8 for real-time recycling waste detection and classification. The model was trained on datasets collected from Malacca, Selangor, and the garbage classification dataset, which consists of 4,039 images of four types of recycled waste. The model was tested with different train-test splits (70:30, 80:20, 90:10). This study demonstrates the effectiveness of YOLO-v8 in classifying recycling waste and highlights its potential for future applications in recycling and other domains [47].

In another study, an improved YOLOv4 model was developed for detecting floating garbage in challenging conditions such as fluctuating illumination, complex backgrounds, and occlusion. The proposed model achieved a mAP of 89% and successfully identified five types of garbage: plastic bottles, aluminum cans, plastic bags, styrofoam, and plastic containers. The performance of this model was compared with other YOLO versions, including YOLOv3-tiny, YOLOv4-tiny, YOLOv3, and YOLOv4, with significant improvements in detection accuracy [48].

The YOLOv5 model was used for plastic waste detection, leveraging both red, green, blue (RGB) and red, green, near-infrared (RGNIR) images. The model underwent a 10-fold cross-validation approach, and the best performance was achieved when the datasets were fused. This fusion of visible and near-infrared spectra resulted in an impressive mAP@0.5 of 92.96%$\pm$2.63%, mAP@0.5:0.95 of 69.47%$\pm$3.11%, and a weighted metric score (WMS) of 71.82%$\pm$3.04%.

YOLOv8 was also found to be suitable for real-time plastic waste sorting in "The Open AI Dataset Project," achieving an accuracy of 86.7% and a mAP of 92.2%. With a shorter inference time of 80–160 ms,

YOLOv8 offers rapid detection, making it highly efficient for real-time applications where both accuracy and processing speed are essential [31]. Table 2 outlines key studies leveraging YOLO-based models for waste detection, emphasizing dataset choices, model types, and practical insights for real-time waste management applications.

Table 2. Recent waste detection studies using YOLO family models

Ref.	Dataset	Model	Application focus	Implications and insights
[31]	The Open AI dataset project	YOLOv8	Plastic waste sorting	Achieved high accuracy and low inference time, suitable for real-time applications in waste sorting systems.
[47]	Garbage classification dataset	YOLOv8	Recycling waste detection	Demonstrated real-time waste classification, with potential for large-scale deployment in recycling.
[48]	Floating Garbage	YOLOv4	Floating garbage detection	Achieved high mAP in challenging environments, showing YOLOv4's robustness for real-world waste detection.

2.3.3. Critical analysis on you only look once family

YOLO has evolved through several iterations, with each new version introducing improvements in accuracy and efficiency. The key strength of YOLO lies in its speed, making it ideal for real-time applications like waste detection. However, earlier versions (YOLOv1 and YOLOv2) suffered from localization errors and struggled with detecting small objects or objects in dense environments.

YOLOv3 and later versions, including YOLOv5 and YOLOv8, have significantly improved accuracy, with advancements like multi-scale predictions and better backbone networks (e.g., Darknet-53). These versions perform well in cluttered and real-world environments, such as waste sorting systems. However, they still face challenges in detecting highly occluded objects or distinguishing between objects with similar visual characteristics. YOLOv5 has been noted for its practical implementation and ease of deployment, making it a popular choice for real-world waste detection tasks. However, its trade-off between accuracy and speed could still be a concern in applications requiring very high precision, especially in environments with overlapping or partially obscured waste.

Despite the recent release of YOLOv9, YOLOv10, and YOLOv11, no published studies have yet applied these newer versions in the context of waste detection. Their capabilities and advantages for this specific application domain remain unexplored, presenting an opportunity for future research. As these models continue to evolve with enhanced architectural improvements, evaluating their potential in waste detection could lead to better performance in challenging and diverse environmental conditions.

While YOLO models excel in real-time inference with frame rates often exceeding 30–120 FPS, they may trade off fine-grained accuracy in cluttered or overlapping waste scenarios compared to Mask R-CNN. Their efficiency makes them highly suitable for conveyor belt sorting systems and drone-based floating waste detection, but occasional misclassifications occur when objects are occluded or visually similar. Although YOLO is lightweight enough for edge devices, interpretability remains limited, with Grad-CAM and related visualization tools offering only partial insight into prediction decisions.

2.4. Transformer-based models

The Transformer is a novel network architecture, designed to address the limitations of traditional sequence-to-sequence models that rely on recurrence and convolutions [49]. Instead, the Transformer architecture relies entirely on an attention mechanism, allowing it to draw global dependencies between input and output without the need for recurrent or convolutional operations. This shift in architecture offers several advantages, including significant improvements in parallelization during training and enhanced computational efficiency. The primary benefit of Transformer models lies in their multi-headed self-attention mechanism, which enables them to effectively capture long-range dependencies in input sequences. This attention-based approach has proven to be faster to train than models based on recurrent or convolutional layers, establishing the Transformer as a key advancement in deep learning architectures.

2.4.1. Transformer variants

There are two notable variants of Transformer models in this domain: the detection Transformer (DETR), designed for object detection, and the vision transformer (ViT), which is primarily used for image classification. Both models leverage a Transformer encoder-decoder architecture for their respective tasks.

- DETR: simplifies object detection by removing components like spatial anchors and non-maximum suppression. It uses a Transformer encoder-decoder and a set-based global loss with bipartite matching for parallel object predictions, leveraging learned object queries and global context for direct end-to-end predictions. Its architecture consists of three components: a CNN backbone for feature extraction, a Transformer encoder-decoder, and a feedforward network for final predictions [50].
- ViT: applies a standard Transformer encoder to image classification by treating images as sequences of patches. These patches are linearly embedded into vectors, and positional encoding are added to preserve spatial relationships between them. The ViT architecture processes these patch embeddings through a stack of Transformer encoder layers with multi-headed self-attention and feed-forward networks. The final output is passed through a classification head to predict class probabilities, achieving competitive performance in image recognition tasks [51].

2.4.2. Recent waste detection by Transformer family

Recent advancements in Transformer-based models for waste detection have demonstrated their potential in automating waste sorting systems. One notable approach is the AL-DETR model, which integrates AL strategies into the DETR framework. AL-DETR addresses the challenge of limited annotated data by iteratively labeling high-uncertainty samples, thus enhancing the model's learning process. By utilizing non-local visual correlations, the model significantly improves waste detection accuracy. It has demonstrated a minimum improvement of 9.23% in detection accuracy over other methods and has surpassed state-of-the-art (SOTA) AL approaches by 1.01% in mAP with reduced annotations. In practical applications, a robotic sorting platform equipped with gripping and suction capabilities successfully validates the model's efficiency in kitchen waste detection and sorting [52].

In the realm of plastic waste sorting, the fine-tuned DETR model has shown exceptional performance when applied to the ZeroWaste dataset. This dataset, which contains crowded scenes with multiple objects, is particularly challenging for traditional models. The fine-tuned DETR achieved an mAP of 25.1% and processed at a speed of 28 FPS, all while reducing computational cost (GFLOPs 86). These characteristics make it an ideal model for large-scale waste sorting systems that require autonomous operation without human intervention. Furthermore, compared to other models like Mask R-CNN and TridentNet, the fine-tuned DETR outperformed them by 5 FPS and 13 FPS, respectively [32].

Another innovative approach is the vision Transformer wastemanagement (ViT-WM) model, customized for smart waste detection. This model was fine-tuned using the enhanced TrashNet dataset, which contains 20,000 images across seven classes, including a newly added biodegradable class. The model achieved an impressive accuracy of 98.17% with a loss of 7.93%. Performance was further validated using real-world images, where the ViT-WM model demonstrated high labeling accuracy. These results highlight the immense potential of Transformer-based models in waste management, paving the way for future optimizations that enhance scalability, modularity, and integration with internet of things (IoT) systems for smarter waste detection and sorting solutions [53]. Table 3 highlights recent advancements in Transformer-based models for waste detection, noting the datasets, model types, and key outcomes relevant to automated sorting and classification systems.

Table 3. Recent waste detection studies using Transformer family models

Ref.	Dataset	Model	Application focus	Implications and insights
[32]	ZeroWaste	Fine-tuned DETR	Plastic waste sorting	Achieved mAP of 25.1% and processing speed of 28 FPS, outperforming other models (Mask R-CNN and TridentNet) in speed. Ideal for large-scale, autonomous sorting.
[52]	ZeroWaste	AL-DETR	Kitchen waste detection	Demonstrated improvement in detection accuracy with AL strategies, surpassing state-of-the-art methods in mAP. Practical validation in robotic sorting platforms.
[53]	TrashNet	ViT-WM	Smart waste detection	Achieved 98.17% accuracy in waste classification, demonstrating high labeling accuracy and potential for scalable smart waste management.

2.4.3. Critical analysis on Transformer-based models

Transformer-based models, particularly those used in object detection (e.g., DETR), have shown promising results in tasks requiring long-range dependencies and complex scene understanding. One of the main advantages of Transformer models is their attention mechanism. This mechanism allows them to capture

global context and inter-object relationships, an essential feature in waste detection, where objects can overlap or have varying shapes.

However, Transformer models like DETR suffer from high computational requirements, especially during training. This can make them less suitable for real-time applications in resource-constrained environments such as waste management systems, where speed is often prioritized. Additionally, their performance can degrade when applied to smaller datasets, as they require larger amounts of labeled data to train effectively. Although recent improvements in DETR-like models (e.g., deformable DETR) aim to address these issues, they still require more resources and longer inference times compared to lightweight models like YOLO.

Transformers models revolutionize object detection by using attention mechanisms to capture long-range dependencies and relationships within an image. While they offer high performance, they are also often difficult to interpret. The multi-head attention mechanism allows the model to attend to different parts of the image, but understanding which specific parts of the image influenced a particular prediction is challenging. For instance, DETR's use of set-based global loss and object queries means that the decision-making process involves interactions across the entire image. Visualizing attention maps may help shed light on this process, but like YOLO, the interpretability of Transformer models remains an ongoing challenge. Furthermore, their complexity and high computational cost can make it difficult to apply interpretability techniques at scale in real-time systems.

3. RESULTS AND DISCUSSION

This section presents an analysis of the performance of object detection models used for waste detection based on the reviewed literature in section 2.1. It is important to emphasize that direct comparisons among models are limited due to variations in datasets, pre-processing techniques, and evaluation metrics. Instead, the discussion emphasizes key trends and observations that can guide future research directions. Table 4 summarizes the performance metrics and key insights for various deep learning models applied in waste detection tasks. It serves as a concise reference to support the comparative discussion of model capabilities, trends, and practical applications.

Table 4. Comparative analysis of deep learning models in waste detection

Model family	Performance metrics	Observed patterns, trends, and insights
R-CNN variants (Two stage models: Faster R-CNN and Mask R-CNN) [28]–[32]	Faster R-CNN: accuracy ∼89.68–91.68%, AP ∼29.1%–34.5%; Mask R-CNN: AP ∼18.4%–30%, mAP ∼22.8%–91.1%, and accuracy up to 91.2%	Generally high accuracy in multi category and precise segmentation tasks. Effective in scenarios prioritizing precision and detailed segmentation, but at the cost of slower inference speed. Performance is highly dependent on dataset quality and annotation density.
YOLO variants (Single stage models: YOLOv4, YOLOv5, and YOLOv8) [31], [47]	YOLOv4 (improved): mAP ∼89%; YOLOv5: mAP ∼93%; YOLOv8: accuracy ∼86–97% and mAP ∼90–97%	Superior in real time applications due to high inference speed. Newer versions show consistent improvements. Despite the release of YOLOv9-v11, their capabilities in waste detection remain untested, offering opportunities for future research.
Transformer-based models (DETR, AL-DETR, and ViT WM) [32], [52], [53]	AL-DETR: mAP ∼58.9%; DETR (fine tuned): mAP 22.8% and 23 FPS; ViT WM: accuracy ∼98%	Rapidly emerging trend, particularly ViT based models achieving high classification accuracy. Limitations remain in speed and deployment complexity. Ideal for applications needing high interpretability.

3.1. Common performance metrics in object detection

Performance metrics in object detection serve as key indicators for assessing the effectiveness of models under various operational scenarios. The most commonly used metrics include accuracy, precision, recall, F_1-score, AP, and mAP. Each metric offers distinct insights, and their usage often depends on the nature of the task, dataset imbalance, or the specific focus of the research.

a. Accuracy is straightforward and widely understood but may not be suitable for imbalanced datasets, where high accuracy can be misleading if the model simply predicts the majority class correctly.
b. Precision focuses on how many of the predicted positive samples are truly positive.
c. Recall assesses how many of the actual positives are correctly detected by the model.

d. F_1-score, the harmonic mean of precision and recall, provides a balanced metric that is especially useful when there is a trade-off between false positives and false negatives.
e. AP evaluates the precision-recall trade-off for a specific class. It is commonly used in benchmark datasets such as COCO or PASCAL VOC.
f. mAP is the mean of AP values across all classes and is regarded as the standard for object detection performance. Two common variants are:

- mAP@0.5: uses an Intersection-over-Union (IoU) threshold of 0.5.
- mAP@0.5:0.95: averages mAP across multiple IoU thresholds (from 0.5 to 0.95 with a step of 0.05), offering a stricter evaluation of localization accuracy.

The choice of evaluation metrics often depends on the nature of the detection task. For instance, studies targeting classification across a limited number of waste categories may emphasize overall accuracy as the primary metric. In contrast, assessments of complex detection models, particularly in multi-object or cluttered environments, often prioritize mAP due to its finer granularity and clearer interpretability. For real-time applications, mAP is typically considered alongside speed-related metrics such as FPS to ensure a balance between accuracy and performance.

Among all, mAP has become the most common and reliable benchmark for object detection performance, particularly for multi-class detection tasks. Nonetheless, many researchers still report additional metrics like accuracy, precision, and F_1-score to provide a comprehensive evaluation, especially when targeting real-world deployment.

Therefore, understanding the research objective is essential for interpreting performance metrics appropriately. A model achieving high accuracy might perform poorly in terms of precision or recall, depending on the task and data distribution. As shown in Table 4, different studies prioritize different metrics based on their specific research goals and dataset characteristics.

3.2. Performance analysis of object detection models

The reviewed studies demonstrate the application of a wide range of object detection models, including both traditional frameworks such as Faster R-CNN and Mask R-CNN, and modern architectures like YOLOv5, YOLOv8, ViT-WM, and AL-DETR. The performance of each model is highly dependent on the specific datasets and tasks it was designed for, as well as the complexity of the detection scenarios. Variability in datasets, such as different types of waste or environmental conditions, plays a crucial role in determining model effectiveness. As such, comparing models across varying conditions is essential for understanding their practical applications and limitations in real-world waste detection.

Faster R-CNN, for instance, has been employed effectively in waste detection tasks. This model achieved an accuracy of 91.68% per category and an overall accuracy of 89.68%, indicating its capacity for accurate classification in structured datasets [28]. Similarly, Mask R-CNN was applied in multiple studies, yielding performance metrics such as an AP ranging from 26.2% to 34.5%, with some outliers reaching over 90% in more structured environments [29], [30]. These results reflect the inherent challenges posed by dataset variability and task complexity, particularly in unstructured or real-world settings.

In contrast, YOLO-based models have consistently demonstrated both high accuracy and computational efficiency, making them well-suited for real-time applications. YOLOv8, for example, achieved an accuracy of 97.63% with precision and recall values exceeding 93% [47]. These findings highlight the robustness and adaptability of YOLO models in waste detection scenarios, especially where fast inference and deployment on resource-constrained systems are required.

Recently, emerging architectures such as ViT-WM and AL-DETR have introduced new possibilities in the domain. ViT-WM achieved the highest accuracy among the reviewed models at 98.17% [53], while AL-DETR reported a mAP of 58.9% [52]. These Transformer-based and AL approaches are particularly promising for tackling complex, cluttered environments and dynamic waste detection challenges. This trend highlights the growing potential of Transformer architectures for precision tasks, while reinforcing the consistent real-time advantages offered by YOLO-based detectors in practical deployments.

In general, single-stage detectors such as YOLOv5 and YOLOv8 consistently demonstrate superior inference speed with only a slight compromise in accuracy compared to two-stage detectors like Faster R-CNN and Mask R-CNN. A clear trade-off emerges between YOLO and Mask R-CNN in cluttered or multi-object scenes. YOLOv5 offers superior speed and strong accuracy, making it ideal for conveyor belt sorting or outdoor monitoring, but it can struggle with overlapping objects. In contrast, Mask R-CNN, although slower, provides

finer segmentation and higher precision in these cluttered scenarios, making it better suited for controlled environments where detailed separation is required.

Meanwhile, Transformer-based models like ViT-WM and AL-DETR provide strong capabilities in handling intricate, cluttered scenes but demand significantly more computational resources. These comparisons emphasize the importance of aligning model selection with operational constraints such as processing speed, accuracy, hardware capabilities, and environmental complexity.

Despite promising results, several failure cases were noted across studies. YOLO-based models often misclassified overlapping or occluded items, particularly plastic bags or transparent waste. Mask R-CNN, while effective in segmentation, struggled with computational cost, limiting its real-time applicability. Transformer-based models such as ViT-WM achieved high accuracy but required powerful GPUs, making them impractical for low-power devices.

Moreover, practical deployments bring additional challenges, including varying lighting conditions, object occlusions, and adverse weather. Although not all reviewed studies explicitly account for such factors, some models demonstrate inherent robustness depending on their architecture and training datasets. For instance, YOLO models are frequently preferred in outdoor environments due to their rapid inference speed, although they may face limitations when detecting overlapping objects. In contrast, Transformer-based models, while computationally intensive, tend to perform better in visually complex scenes. This further reinforces the need to balance accuracy and efficiency with the realities of field deployment.

3.3. Influence of dataset variability

Dataset variability emerged as a critical factor influencing model performance. Table 5 shows the unified dataset characteristics summary based on these studies. A closer inspection of the reviewed studies reveals a direct correlation between dataset quality, particularly in terms of balance, size, and representativeness of the model performance. For instance, in [47], the YOLOv8 model achieved exceptionally high performance metrics (accuracy of 97.63%, precision of 95.30%, and recall of 93.03%), largely attributed to the well-curated dataset that was both diverse and sufficiently annotated.

In contrast, studies utilizing more diverse or imbalanced datasets, such as the TACO dataset used in [29], reported significantly lower AP scores, ranging from 18.4% to 26.2%. This is likely due to several compounding factors: the wide range of waste categories and an uneven distribution of labeled images. Such conditions hinder the model's ability to generalize across categories, leading to reduced accuracy.

Critically, while dataset variability presents a challenge, it also offers an opportunity. Models that perform well on diverse and imbalanced datasets are more likely to succeed in real-world deployments, where conditions are far less controlled. Therefore, low performance scores should not necessarily be viewed as a limitation of the model, but rather a reflection of the model's exposure to realistic and complex scenarios. To that end, future benchmarking efforts should incorporate not only standardized datasets but also stress-test models under varying degrees of environmental complexity, lighting, occlusion, and object overlap.

Moreover, the lack of common benchmarking datasets across studies hinders meaningful model-to-model comparisons. To mitigate dataset limitations, several strategies can be employed. Data augmentation (e.g., rotation, flipping, scaling, and color jittering) can increase dataset diversity and improve generalization to unseen conditions. Synthetic data generation using GANs or simulation can help balance underrepresented categories and reduce collection costs.

3.4. Task-specific observations

The complexity of specific tasks also significantly impacts performance metrics. Waste detection encompasses a spectrum of tasks, from simple object detection and classification to more complex operations such as instance segmentation and real-time detection. Each task presents unique computational and algorithmic challenges, which influence model selection and performance expectations.

For instance, models used for instance segmentation, like Mask R-CNN in [30], tend to exhibit lower AP values compared to those used solely for object detection. In [30], instance segmentation yielded an AP of 30.0%, and material detection an AP of 28.2%. This performance gap is expected, as instance segmentation tasks involve finer-grained localization and boundary delineation, which are inherently more complex and error-prone, especially in cluttered scenes or when detecting irregular-shaped waste materials such as plastic bags or fragmented glass.

In contrast, object detection models optimized for real-time applications, such as YOLOv5 and fine-tuned DETR [32], demonstrated lower APs (22.8%) but compensated with high-speed inference (up to 23

FPS). This trade-off between detection accuracy and processing speed is crucial in real-world applications, particularly for dynamic environments such as conveyor belts in recycling facilities, where decisions must be made within milliseconds.

Table 5. Unified dataset characteristics summary for reviewed waste-detection studies

Study	Dataset size	Waste categories	Environment	Scene complexity	Annotation type
Domestic garbage [28]	14,587 images (training subset cited: 3,984)	23 recyclable, 6 other, 4 hazardous, 8 kitchen (total 41)	Mobile phone & web images; varied definition/sizes; lighting NR	NR	VOC2007 format; variable image quality noted
TACO [29]	1,500 images; 4,784 inst.	60 categories (28 super); TACO-10 derived (9 super + Other)	Outdoor/"in-the-wild"; high-res mobile; diverse backgrounds	Small objects common (e.g., many $< 64 \times 64$); class imbalance	Instance segmentation; crowdsourced; hierarchical taxonomy
TrashCan [30]	7,212 images	Trash, ROV, Bio (plant/animal types), Unknown; material/instance variants	Underwater ROV footage; long-term archive (JAMSTEC)	Overgrowth/decay; clutter; multiple non-trash distractors	Instance segmentation; 21 annotators; ∼1,500 hours; COCO-converted polygons
AI-Hub (KR) plastics [31]	∼800k available; balanced subset 5,000 (1,250/class)	PE, PET, PS, PP	Recycling facility; realistic conveyor context; lighting NR	Multiple objects likely; crowding NR	COCO JSON; rich metadata (container, transparency, shape, size)
ZeroWaste [32]	∼3,000 images (scenes)	Cardboard, Soft plastic, Rigid plastic, Metal (class imbalance: Cardboard ∼12k objs)	MRF conveyor; controlled lighting; dual cameras + diffusers	Multiple objects/scene (typ. 4–12; up to 13+); crowding	COCO; real MRF captures; strong labeling quality
Custom dataset [47]	4,039 images (test); 10,057 with augmentation	Paper, glass, metal, plastic	Field captures (Malacca and Selangor) + public dataset; lighting NR	NR	Bounding boxes; standard labeling; augmentation applied
Floating debris dataset [48]	NR	Styrofoam, plastic bag, bottle, container and can	Rivers; varied brightness/positions by design	NR explicitly; motion/background clutter implied	Bounding boxes; transfer learning from MS-COCO
Kitchen waste dataset [52]	NR (new dataset; 8 classes)	8 kitchen-waste classes (large objects focus)	Industrial sorting scenario; background complexity	Complex backgrounds; large-object focus	New dataset; DETR-compatible; labels expanded via AL
Smart-city dataset [53]	N/A (survey / system-design focus)	Urban waste types (conceptual)	City-scale sensing (IoT); varied conditions (conceptual)	N/A	N/A (integrative/system-focused study)

Notes: NR = not reported. "Inst." = instances. Where studies provided only partial counts (e.g., subsets/augmented sets), both are noted. This table standardizes fields across heterogeneous sources to aid side-by-side comparison.

Critically analyzing this trade-off reveals that no single model architecture universally outperforms across all tasks. For high-precision segmentation in controlled environments (e.g., smart bins or lab-based setups), models like Mask R-CNN or ViT-WM may be preferable despite their slower inference. Conversely, for real-time deployment where speed and responsiveness are paramount, lightweight versions of YOLO or DETR, possibly further optimized via quantization or edge-acceleration techniques, are better suited.

Furthermore, it's worth noting that models designed for multi-task learning used for handling both detection and classification or segmentation simultaneously are still underexplored in the waste detection domain. Such architectures could potentially enhance overall system performance by leveraging shared features and reducing redundancy. Future work could benefit from integrating multi-task learning strategies to simultaneously detect, classify, and segment waste materials in a unified pipeline.

In summary, the suitability of a model for waste detection is highly task-dependent. Researchers and practitioners must clearly define the objectives of their waste detection systems whether prioritizing accuracy, speed, segmentation precision, or hardware efficiency before selecting or developing models. A deeper understanding of these trade-offs will ensure that waste detection systems are not only technically robust but also

practically deployable in diverse environmental and industrial contexts.

3.5. Scalability for industrial-scale waste sorting

Evaluating scalability is a key consideration for deploying waste detection models at an industrial scale, where high throughput and real-time processing are critical. The scalability of models like YOLOv5 and ViT-WM, which balance high accuracy with computational efficiency, makes them strong candidates for such applications. However, ensuring that these models can handle the demands of large-scale deployment requires optimizations in both processing speed and hardware requirements.

For instance, real-time waste sorting systems in industrial settings need models that can process data quickly. YOLO models are particularly well-suited for these environments due to their ability to perform fast inference (e.g., up to 23 FPS in some configurations). To further enhance scalability, model optimizations such as quantization and pruning can reduce model size and computational complexity without sacrificing performance, making them more efficient for edge devices.

In practical terms, waste detection systems are being deployed in scenarios such as conveyor belt sorting in recycling plants, where inference speed is critical for keeping up with high waste throughput. Drone-based detection has also been explored for monitoring coastal or landfill waste, where lightweight YOLO models are often preferred due to hardware constraints and the need for real-time feedback. These deployment contexts highlight the importance of selecting models not only based on accuracy but also on operational feasibility. In terms of hardware, deploying these models on edge devices for industrial-scale waste sorting requires powerful yet energy-efficient computing resources. A systematic evaluation of computational requirements shows that YOLO models are generally more lightweight, achieving real-time inference on mid-tier GPUs or edge devices with memory usage often below 2–4 GB. In contrast, R-CNN and Transformer-based models typically demand 8–16 GB or more, which restricts their feasibility in low-power or resource-constrained facilities.

To address these demands, high-performance GPUs or specialized AI chips (e.g., TPUs) are often necessary to accelerate inference in real-time applications. For edge deployments, low-power embedded systems equipped with AI accelerators can reduce energy consumption while maintaining reliable operation. Ultimately, close collaboration between hardware manufacturers and waste management stakeholders will be essential to define appropriate specifications and fine-tune models for optimal performance across diverse, large-scale, and resource-limited settings. Finally, integrating waste detection models with IoT devices and edge computing platforms can further enhance scalability, enabling faster, more accurate waste sorting and recycling processes in industrial environments. This integration can support smart city initiatives by promoting more efficient and sustainable waste management practices, aligning with broader environmental goals.

3.6. Operational integration: robotics and environmental impact

Deploying waste-detection models requires coupling perception with actuation, sustainability accounting, and human–system interaction within real facilities. Regarding robotics, integration challenges frequently arise with deformable items such as thin films and bags that collapse under suction. Depth sensors can be confused by wet or greasy organics that clog vacuum lines and entangle materials (e.g., stringy plastics, cables) that trigger multi-object picks. In production cells, teams typically budget end-to-end latency across vision, planning, and approach to grasp success rate, picks per hour, and mis-sort rate remain within shift targets without increasing jams or maintenance load.

From a sustainability perspective, we recommend evaluating deployments with an ISO 14040/44-aligned, gate-to-gate life cycle assessment (LCA) that reports a clear functional unit (e.g., per ton of inbound waste). Rather than only quoting model accuracy, operators benefit from energy-anchored metrics such as energy per inference, energy per pick, and total kWh per ton processed, converted to $kgCO_2e$ using local grid factors. These should be reported alongside diversion improvements (e.g., recovery-rate uplift or contamination reduction) to reveal trade-offs between added electricity, air consumption, and downstream environmental gains.

For operator-facing software in material recovery facilities (MRFs), user interface design should minimize cognitive load and mesh with shift workflows. Confidence-aware overlays (e.g., traffic-light encoding) with adjustable thresholds help supervisors tune precision–recall per class; progressive disclosure surfaces only critical alarms (jams, suction faults, high mis-sort) while detailed analytics (per-class precision/recall, per-camera FPS) stay in drill-down views. Color-blind–safe palettes, large touch targets usable with gloves, and multilingual iconography support diverse crews. Lightweight, two-click feedback to confirm or correct

low-confidence detections can continuously populate an active-learning queue, improving performance without interrupting throughput. Audit trails (event timelines of detections, interventions, and downtime) enable traceability and shift-to-shift comparisons.

3.7. Benchmarking framework

To enable fair, reproducible, and deployment-oriented comparisons, we consolidate protocol, metrics, pre-processing, statistics, scoring, implementation timeline, and validation into a single framework. First, use stratified, scene-disjoint splits (70% train, 20% val, 10% test) and 10-fold cross-dataset validation. Train each model with three seeds and report mean±95% confidence interval (CI). Fix 300 epochs for YOLO/SSD, 50 for two-stage R-CNNs, 150 for DETR-like models (early stopping on validation mAP@0.5:0.95, patience=20). Constrain hyperparameters to a small grid (LR $\in \{10^{-2}, 10^{-3}, 5 \times 10^{-4}\}$; weight decay $\in \{0, 10^{-4}\}$; input size $\in \{640, 800\}$) and log all choices. For latency, set batch size as 1, standardize image size, warm up 200 iterations, then time 1000 inferences without I/O; report p50/p95 latency and FPS (inverse of median latency). Record hardware (CPU/GPU/RAM), drivers, and framework versions. Measure energy via on-device telemetry and/or external wattmeter at 1 Hz; compute Joules per inference and kWh per ton for end-to-end cells. Report picks/hour and mis-sort rate for sorting cells using a fixed grasp-selection heuristic.

Next is to report COCO-style mAP@0.5:0.95, mAP@0.5, per-class AP, class-averaged precision, recall and F_1; efficiency (median/p95 latency, FPS, parameters, model size, FLOPs/image); resources/costs (peak GPU memory, average power, J/inference, CAPEX of test hardware, OPEX from electricity rate and duty cycle); robustness deltas under blur/low light/occlusion/domain shift; operator KPIs (picks/hour, mis-sort %, alarm rate). Then, convert all datasets to COCO JSON with a unified taxonomy (e.g., {plastic, paper, metal, glass, organics, other}) and publish the class map. Remove near-duplicates with perceptual hashing (Hamming$\leq$5) and enforce scene-disjoint splits. Standard input: short side=640 px (letterbox), RGB float32, ImageNet mean/std; additionally report two-stage results at 800 px. Train-time augmentations at horizontal flip (0.5), scale jitter (0.8–1.2), color jitter (0.2 each), light cutout; mosaic disabled by default (report separately). Relabel objects $< 4 \times 4$ px to background and record counts. If any class has < 200 train instances, use instance-balanced sampling or focal loss and document the choice.

For statistical testing, provide 10,000-sample bootstrap CIs over images. For paired comparisons on the same test set, use Wilcoxon signed-rank (or paired t-test if normality passes Shapiro–Wilk); control family-wise error with Holm–Bonferroni and report effect sizes (Cliff's δ or Cohen's d). For robustness, apply two-way repeated-measures ANOVA (or aligned rank transform) across model×stressor with corrected post-hoc tests. Release seeds, configs, commit hashes, and per-image outputs. For validation, hold out an unseen site/domain and report absolute metrics and deltas; perform cross-dataset generalization (A→B, B→A); augmentation, NMS/decoder, and quantization/pruning; re-run latency on three hardware tiers (edge SoC, mid-GPU, data-center GPU) with identical software; simulate human-in-the-loop feedback for low-confidence detections (active-learning queue) and quantify one-iteration gains.

3.8. Ethical framework for artificial intelligence waste management

An ethical framework for AI-enabled waste systems should integrate bias detection and mitigation, algorithmic fairness, privacy protection, environmental impact assessment, participatory governance, and staged implementation guidance. The objective is to ensure that computer vision and robotics deployments in waste monitoring and sorting are not only accurate and efficient, but also fair, privacy-preserving, environmentally responsible, and responsive to stakeholder needs.

Bias should be investigated through systematic dataset and label audits that quantify long-tail class imbalance, co-occurrence skew (e.g., plastics with specific backgrounds), and site/domain skew across facilities and lighting conditions. Double annotation can be used to estimate label noise, and performance should be stratified by contextual subgroups G (e.g., facility, shift, and lighting bin) with per-class AP/mAP reported alongside 95% bootstrap confidence intervals. Spurious correlations should be probed via counterfactual tests (background swaps, copy–paste of objects, and photometric jitter) and stressors (blur, glare, wetness, and occlusion). Mitigation may include reweighting or rebalancing, focal loss for rare classes, targeted active-learning queries in underrepresented subgroups, and class-specific thresholds or abstention. All changes should be pre-registered and evaluated with paired statistical tests as specified in subsection 4.7.

Fairness should be reported alongside utility. Recommended measures include the subgroup mAP gap $\Delta\mathrm{mAP} = \max_{g,g'} |\mathrm{mAP}_g - \mathrm{mAP}_{g'}|$, equal opportunity via true-positive-rate parity at IoU $\geq \tau$, equalized odds using disparities in both TPR and FPR, calibration parity via group-wise expected calibration error (ECE),

and selective-risk parity at fixed prediction coverage. These metrics should be presented together with latency and throughput to expose trade-offs between fairness and efficiency.

Privacy risks should be reduced through on-device inference, logging structured detections and telemetry rather than raw video frames, and automatic redaction of faces, license plates, and name badges before any storage or transfer. A data protection impact assessment should document purpose limitation, role-based access control, transport-layer security (TLS 1.3), encryption at rest (e.g., AES-256), immutable audit logs, and default retention periods not exceeding 30 days for video and 180 days for redacted crops unless a legitimate operational need is documented. Clear signage and worker notices should be provided, and dataset/model cards should disclose data sources, known risks, and mitigation steps.

Environmental performance should be quantified with an ISO 14040/44-aligned, gate-to-gate lifecycle assessment using the functional unit "per ton of inbound waste." Average electrical power $\bar{P}$ and median per-image latency ℓ should be measured to compute energy per inference $E_{\text{inf}} = \bar{P}\,\ell$ and cell-level energy intensity (kWh/ton). Emissions should be reported as $\mathrm{CO_2e/ton} = \text{kWh/ton} \times \mathrm{PUE} \times \mathrm{EF}_{\text{grid}}$, and interpreted jointly with diversion uplift and contamination reduction. Recommended impact-reduction measures include quantization and pruning, edge or accelerator deployment, duty-cycling, and scheduling high-throughput training jobs on low-carbon-intensity grids.

A participatory process should map operators, municipal sanitation departments, robotics vendors, line workers and union representatives, NGOs, and local communities. Responsibilities should be formalized using a RACI matrix across safety, privacy, fairness, and deployment decisions, with an independent ethics board including external members. User-centered design should be validated through glove-friendly, color-blind-safe, multilingual interfaces that minimize cognitive load and support progressive disclosure (e.g., high-salience alarms with drill-down analytics). Governance should include public fairness and energy dashboards, incident reporting, and clear grievance and redress mechanisms.

Deployment should proceed through five stages with explicit entry and exit criteria. Stage 0 (scoping) defines KPIs such as mAP@0.5:0.95, picks per hour, mis-sort rate, fairness gaps, and kWh/ton, and completes privacy/LCA plans. Stage 1 (sandbox) conducts controlled benchmarking, failing the gate if $\Delta\mathrm{mAP} > 5$ percentage points or if calibration gaps exceed 3 percentage points. Stage 2 (shadow mode) operates without actuation while completing the DPIA and validating redaction and retention. Stage 3 (human-in-the-loop) enables selective abstention to operators with weekly fairness and carbon reviews. Stage 4 (scale-out) hardens security, performs quarterly re-audits, and enforces rollback triggers for mis-sort drift, fairness regression, or privacy incidents.

In a municipal material-recovery facility, a transparent-plastic subgroup gap was reduced from 7.2 to 2.6 percentage points through targeted augmentation and class weighting; only redacted crops were stored with a 28-day retention limit, and the system traded an additional 4.8 kWh/ton for a 6.1 percentage-point reduction in contamination. In a coastal litter monitoring program using UAVs, on-device inference with geofenced recording preserved privacy while maintaining habitat-level parity (beach, mangrove, and rip-rap $\Delta\mathrm{mAP} < 3$ percentage points) and providing hotspot reports to community cleanup groups.

3.9. Emerging trends in the literature

Recent advancements in the field of object detection highlight a growing interest in leveraging state-of-the-art deep learning architectures for environmental applications. Transformer-based models, such as ViT-WM, are gaining attention for their ability to capture long-range dependencies and contextual relationships in complex scenes. This ability is particularly beneficial in cluttered environments like coastal or urban waste sites. Simultaneously, AL frameworks like AL-DETR are emerging as promising solutions to reduce the reliance on large annotated datasets. By prioritizing high-uncertainty samples, these methods enhance learning efficiency, making them well-suited for domains where labeled data is limited or costly to obtain.

Another noticeable trend is the increasing interest in deploying lightweight and efficient models on edge devices and IoT systems, reflecting a broader movement toward real-time, in-situ environmental monitoring. This aligns with global smart city initiatives and the drive for sustainable, automated waste management solutions. Collectively, these trends indicate a shift toward more intelligent, adaptive, and scalable approaches for addressing environmental challenges through AI-driven object detection.

3.10. Summary of research questions addressed

This subsection revisits the research questions (RQs) stated in section 1 and summarizes how each has been addressed based on the experiments and analysis.

a. RQ1: What are the most recent object detection models applied to waste detection, and how have they evolved from 2019 to 2024?
   This study reviewed and selected several recent object detection models developed between 2019 and 2024 for their applicability to waste detection tasks. The models include the R-CNN, YOLO family, and Transformer-based models. These models were selected based on their relevance in both the general object detection domain and the specific challenges of waste detection. The evolution of these models from traditional convolutional approaches to advanced Transformer-based architectures has significantly impacted detection accuracy and efficiency, directly addressing RQ1 by identifying state-of-the-art models that reflect the latest advancements in the field.
b. RQ2: How do these object detection models perform in detecting and classifying various types of waste, particularly in complex and cluttered environments?
   Performance evaluations, presented in subsection 4.2, explored how well each model detects and classifies various waste items in real-world environments, which often feature complex and cluttered backgrounds. The analysis included a variety of waste categories, such as plastics, metals, and organic materials. The results showed that models like YOLOv8 achieved the best detection accuracy, particularly in cluttered environments with partial occlusions and varying lighting conditions. This directly addresses RQ2 by demonstrating the practical performance of these models in challenging waste detection scenarios.
c. RQ3: What are the key challenges and future research opportunities in the field of waste detection using object detection models?
   Several challenges remain in the field of waste detection, including the need for larger and more diverse datasets, handling background clutter, dealing with occlusion, and improving detection accuracy in low-light conditions. Additionally, the application of lightweight models for real-time detection and the integration of multimodal sensors (such as RGNIR) present promising future research opportunities. These challenges and opportunities are outlined in detail, addressing RQ3 by providing insights into the limitations and directions for future work in the field.

4. CHALLENGES, RESEARCH GAPS, AND FUTURE DIRECTIONS

To further advance the practical application of object detection models in waste detection, it is essential to explicitly identify existing challenges, clearly define current research gaps, and strategically outline promising future research directions.

4.1. Challenges

Despite the promising performance of object detection models in waste detection, several technical and practical challenges remain. One major concern is the variability of datasets used to evaluate different models. Many studies rely on diverse datasets with varying complexities and imbalances, leading to inconsistent comparisons and limiting the generalizability of model performance across different environments. Additionally, energy constraints present a significant barrier, particularly for Transformer-based models, which demand considerable computational resources. Deploying these models in real-time applications, especially on edge devices with limited power, remains a major challenge. Ethical and regulatory considerations further complicate the deployment of AI-based waste detection systems. The risk of false detections in automated waste sorting processes may disrupt recycling workflows and lead to improper disposal. This calls for increased transparency, fairness, and accountability in AI system design.

4.2. Research gaps

A key research gap is the limited exploration and validation of Transformer-based models, such as ViT-WM, in waste detection tasks. While these models have demonstrated high performance in general computer vision applications, their suitability, optimization, and real-world effectiveness in waste management remain largely underexplored. Moreover, the lack of a standardized global benchmark dataset poses a significant obstacle. Without such datasets, it becomes difficult to conduct fair evaluations and ensure reproducibility across studies. The absence of unified benchmarking protocols (e.g., consistent evaluation across datasets such as TACO, TrashNet, or ZeroWaste) further compounds this issue, as results are often not directly comparable across studies. This gap highlights the need for collaborative efforts to establish comprehensive and balanced datasets that can serve as a foundation for more consistent model development and evaluation.

4.3. Future directions

To advance the field, several future research directions are recommended. Systematic studies on Transformer-based models should be conducted to evaluate their potential in waste detection scenarios. This includes efforts to optimize these models for both accuracy and computational efficiency. The creation of standardized, global benchmark datasets must be prioritized. Collaboration with environmental agencies, governments, and research institutions will be vital to ensure data diversity and quality, ultimately enhancing the reliability of research outcomes.

Energy-efficient strategies, such as model quantization, pruning, and the adoption of edge AI techniques, should be explored to facilitate real-time deployment in resource-constrained environments. Likewise, optimizing inference speed while maintaining high accuracy is critical for practical applications. Future work should also investigate multi-modal fusion approaches, such as combining RGB images with hyperspectral, thermal, or near-infrared modalities, to improve detection robustness under varying environmental conditions.

In parallel, future research should also prioritize interpretability techniques such as saliency maps, Grad-CAM, attention visualization, and decision rationale explanations. These approaches can provide transparency into how models detect and classify waste, improving trust among operators and stakeholders. Embedding interpretability as a core design principle will not only enhance accountability but also facilitate troubleshooting, bias detection, and informed decision-making in real-world waste management applications.

Finally, future research should emphasize integration with IoT and edge computing platforms for scalable, real-time waste management solutions. Open-source datasets and cross-institutional collaborations will be vital for transparency and reproducibility. Progress will also depend on interdisciplinary synergies: robotics for automated sorting, environmental science for sustainability alignment, and human–computer interaction (HCI) for user-friendly, trustworthy systems. Equally important is embedding research within ethical frameworks guided by fairness, accountability, transparency, and sustainability. Adopting best practices from established guidelines, such as the EU Trustworthy AI principles, will help ensure future waste detection models are not only technically efficient but also socially responsible and environmentally sustainable.

5. CONCLUSION

This review has provided a systematic analysis of recent advancements in object detection models for waste detection, covering traditional two-stage methods (Faster R-CNN and Mask R-CNN), single-shot detectors (YOLOv1 to YOLOv11), and emerging Transformer-based architectures (ViT-WM and AL-DETR). These innovations led to significant improvements in accuracy, robustness, and inference speed. However, several challenges remained, such as dataset variability, high computational demands, ethical concerns, and limitations in real-time deployment. Two critical research gaps were also identified: the limited exploration and application of Transformer-based models specifically within the waste detection domain, and the lack of universally accepted, standardized benchmark datasets essential for consistent and reliable model evaluation. To guide future research efforts, this review explicitly recommends actionable priorities: i) collaboratively developing comprehensive, standardized, and publicly accessible benchmark datasets; ii) systematically optimizing and rigorously validating Transformer architectures specifically tailored to complex waste detection tasks; iii) focusing research on energy-efficient and real-time inference techniques for effective edge deployment; and iv) emphasizing model interpretability, ethical transparency, and regulatory compliance to build trust and reliability in deployed systems. Finally, it encourages interdisciplinary and cross-sector collaboration among academia, industry, and governmental entities to ensure meaningful progress toward scalable, effective, and sustainable waste management solutions.

ACKNOWLEDGMENTS

The authors would like to acknowledge the Ministry of Higher Education (MOHE) Malaysia through Fundamental Research Grant Scheme [FRGS/1/2020/ICT06/UMS/02/1] and Universiti Malaysia Sabah (UMS) for continuous support and resources made available throughout this whole research work. Special thanks to the Faculty of Science and Technology and Faculty of Computing and Informatics, UMS for the tremendous computing facilities support. The authors also wish to thank Penerbit UMS (UMS Press) for funding the publication of this work.

FUNDING INFORMATION

The article processing charge is funded by Penerbit UMS (UMS Press).

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Owen Tamin	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓			
Ervin Gubin Moung	✓	✓	✓	✓		✓	✓	✓		✓	✓	✓	✓	✓
Ali Farzamnia	✓									✓			✓	

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal Analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject Administration
Fu : **Fu**nding Acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

Data availability is not applicable to this paper as no new data were created or analyzed in this study.

REFERENCES

[1] P. H. Brunner and H. Rechberger, "Waste to energy–key element for sustainable waste management," *Waste Management*, vol. 37, pp. 3–12, 2015, doi: 10.1016/j.wasman.2014.02.003.
[2] A. J. Chandler *et al.*, *Municipal solid waste incinerator residues*, Elsevier, 1997.
[3] S. E. Vergara and G. Tchobanoglous, "Municipal solid waste and the environment: a global perspective," *Annual Review of Environment and Resources*, vol. 37, no. 1, pp. 277–309, 2012, doi: 10.1146/annurev-environ-050511-122532.
[4] E. Amasuomo and J. Baird, "The concept of waste and waste management," *Journal of Management and Sustainability*, vol. 6, no. 4, pp. 88-96, 2016, doi: 10.5539/jms.v6n4p88.
[5] M.A. Hannan, M. Arebey, R.A. Begum, and H. Basri, "An automated solid waste bin level detection system using a gray level aura matrix," *Waste Management*, vol. 32, no. 12, pp. 2229–2238, 2012, doi: 10.1016/j.wasman.2012.06.002.
[6] A. Malakahmad and N. D. Khalil, "Solid waste collection system in Ipoh city," in *2011 International Conference on Business, Engineering and Industrial Applications*, Kuala Lumpur, Malaysia, 2011, pp. 174-179, doi: 10.1109/ICBEIA.2011.5994236.
[7] O. M. Poulsen *et al.*, "Sorting and recycling of domestic waste. Review of occupational health problems and their possible causes," *Science of the Total Environment*, vol. 168, no. 1, pp. 33–56, 1995, doi: 10.1016/0048-9697(95)04521-2.
[8] M. Abdallah, M. A. Talib, S. Feroz, Q. Nasir, H. Abdalla, and B. Mahfood, "Artificial intelligence applications in solid waste management: A systematic research review," *Waste Management*, vol. 109, pp. 231–246, 2020, doi: 10.1016/j.wasman.2020.04.057.
[9] O. Tamin, E. G. Moung, J. A. Dargham, F. Yahya, and S. Omatu, "A review of hyperspectral imaging-based plastic waste detection state-of-the-arts," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 13, pp. 3407–3419, 2023, doi: 10.11591/ijece.v13i3.pp3407-3419.
[10] C. Polprasert and T. Koottatep, *Organic waste recycling: technology and management*, IWA publishing, 2007, doi: 10.2166/9781780408217.
[11] P. Kiddee, R. Naidu, and M. H. Wong, "Electronic waste management approaches: An overview," *Waste Management*, vol. 33, no. 5, pp. 1237–1250, 2013, doi: 10.1016/j.wasman.2013.01.006.
[12] S. Kaza, L. Yao, P. Bhada-Tata, and F. Van Woerden, *What a waste 2.0: a global snapshot of solid waste management to 2050*, World Bank Publications, 2018.
[13] B. Fang *et al.*, "Artificial intelligence for waste management in smart cities: a review," *Environmental Chemistry Letters*, vol. 21, no. 4, pp. 1959–1989, 2023, doi: 10.1007/s10311-023-01604-3.
[14] Z.-Q. Zhao, P. Zheng, S.-T. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 11, pp. 3212–3232, 2019, doi: 10.1109/TNNLS.2018.2876865.
[15] P. Jiang, D. Ergu, F. Liu, Y. Cai, and B. Ma, "A Review of Yolo algorithm developments," *Procedia Computer Science*, vol. 199, pp. 1066–1073, 2022, doi: 10.1016/j.procs.2022.01.135.
[16] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017, doi: 10.1109/TPAMI.2016.2577031.

[17] A. Khan *et al.*, "A survey of the vision transformers and their CNN-transformer based variants," *Artificial Intelligence Review*, vol. 56, pp. 2917–2970, 2023, doi: 10.1007/s10462-023-10595-0.

[18] H. Abdu and M. H. M. Noor, "A survey on waste detection and classification using deep learning," *IEEE Access*, vol. 10, pp. 128151–128165, 2022, doi: 10.1109/ACCESS.2022.3226682.

[19] D. O. Melinte, A.-M. Travediu, and D. N. Dumitriu, "Deep convolutional neural networks object detector for real-time waste identification," *Applied Sciences*, vol. 10, no. 20, pp. 1-18, 2020, doi: 10.3390/app10207301.

[20] W. Lu and J. Chen, "Computer vision for solid waste sorting: A critical review of academic research," *Waste Management*, vol. 142, pp. 29–43, 2022, doi: 10.1016/j.wasman.2022.02.009.

[21] T.-W. Wu, H. Zhang, W. Peng, F. Lü, and P.-J. He, "Applications of convolutional neural networks for intelligent waste identification and recycling: A review," *Resources, Conservation and Recycling*, vol. 190, p. 106813, 2023, doi: 10.1016/j.resconrec.2022.106813.

[22] K. Huang, H. Lei, Z. Jiao, and Z. Zhong, "Recycling waste classification using vision transformer on portable device," *Sustainability*, vol. 13, no. 21, pp. 1-14, 2021, doi: 10.3390/su132111572.

[23] F. S. Alrayes *et al.*, "Waste classification using vision transformer based on multilayer hybrid convolution neural network," *Urban Climate*, vol. 49, p. 101483, 2023, doi: 10.1016/j.uclim.2023.101483.

[24] Sumit, S. Bisht, S. Joshi, and U. Rana, "Comprehensive review of r-cnn and its variant architectures," *International Research Journal on Advanced Engineering Hub (IRJAEH)*, vol. 2, no. 04, pp. 959–966, 2024, doi: 10.47392/IRJAEH.2024.0134.

[25] B. Cheng, Y. Wei, H. Shi, R. Feris, J. Xiong, and T. Huang, "Revisiting rcnn: On awakening the classification power of faster rcnn," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 473–490, doi: 10.1007/978-3-030-01267-0_28.

[26] R. Girshick, "Fast r-cnn," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, 2015, pp. 1440-1448, doi: 10.1109/ICCV.2015.169.

[27] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2961–2969, doi: 10.1109/ICCV.2017.322.

[28] Z. Nie, W. Duan, and X. Li, "Domestic garbage recognition and detection based on Faster R-CNN," in *2020 2nd International Conference on Electronics and Communication, Network and Computer Technology (ECNCT)*, Chengdu, China, 2021, vol. 1738, no. 1, p. 012089, doi: 10.1088/1742-6596/1738/1/012089.

[29] P. F. Proença and P. Simões, "Taco: Trash annotations in context for litter detection," *arXiv preprint*, 2020, doi: 10.48550/arXiv.2003.06975.

[30] J. Hong, M. Fulton, and J. Sattar, "Trashcan: A semantically-segmented dataset towards visual detection of marine debris," *arXiv preprint*, 2020, doi: 10.48550/arXiv.2007.08097.

[31] J. Son and Y. Ahn, "AI-based plastic waste sorting method utilizing object detection models for enhanced classification," *Waste Management*, vol. 193, pp. 273–282, 2025, doi: 10.1016/j.wasman.2024.12.014.

[32] T. T. Nguyen, T. T. Luu, and P. T. A. Tong, "Fine-tuning DETR: Toward holistic process in plastic waste sorting system," *Waste Management*, vol. 179, pp. 154–162, 2024, doi: 10.1016/j.wasman.2024.03.015.

[33] W. Liu *et al.*, "Ssd: Single shot multibox detector," in *European Conference on Computer Vision*, Amsterdam, The Netherlands, Springer, 2016, pp. 21–37, doi: 10.1007/978-3-319-46448-0_2.

[34] J. Terven, D.-M. Córdova-Esparza, and J.-A. Romero-González, "A comprehensive review of yolo architectures in computer vision: From yolov1 to yolov8 and yolo-nas," *Machine Learning and Knowledge Extraction*, vol. 5, no. 4, pp. 1680–1716, 2023, doi: 10.3390/make5040083.

[35] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 779-788, doi: 10.1109/CVPR.2016.91.

[36] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, pp. 6517-6525, doi: 10.1109/CVPR.2017.690.

[37] L. Zhao and S. Li, "Object detection algorithm based on improved YOLOv3," *Electronics*, vol. 9, no. 3, p. 537, 2020.

[38] S.-J. Ji, Q.-H. Ling, and F. Han, "An improved algorithm for small object detection based on YOLO v4 and multi-scale contextual information," *Computers and Electrical Engineering*, vol. 105, p. 108490, 2023.

[39] G. Jocher *et al.*, "ultralytics/yolov5: v3. 0," *Zenodo*, 2020, doi: 10.5281/zenodo.3983579.

[40] S. N. Saydirasulovich, A. Abdusalomov, M. K. Jamil, R. Nasimov, D. Kozhamzharova, and Y.-I. Cho, "A YOLOv6-based improved fire detection approach for smart city environments," *Sensors*, vol. 23, no. 6, pp. 1-18, p. 3161, 2023.

[41] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada, 2023, pp. 7464-7475, doi: 10.1109/CVPR52729.2023.00721.

[42] G. Jocher, A. Chaurasia, and J. Qiu, "YOLO by Ultralytics," 2023, [Online]. Available: https://github.com/ultralytics/ultralytics. (Accessed: 2025-01-03).

[43] C. Gupta, N. S. Gill, P. Gulia, A. Kumar, H. Karamti, and D. M. Moges, "An optimized YOLO NAS based framework for realtime object detection," *Scientific Reports*, vol. 15, no. 1, p. 32903, 2025, doi: 10.1038/s41598-025-17919-w.

[44] C.-Y. Wang, I.-H. Yeh, and H.-Y. M. Liao, "Yolov9: Learning what you want to learn using programmable gradient information," in *European Conference on Computer Vision*, Springer, 2024, pp. 1–21, doi: 10.1007/978-3-031-72751-1_1.

[45] A. Wang *et al.*, "Yolov10: Real-time end-to-end object detection," in *NIPS '24: Proceedings of the 38th International Conference on Neural Information Processing System*, 2024, pp. 107984–108011, doi: 10.5555/3737916.3741345.

[46] Ultralytics, "Ultralytics: Cutting-edge YOLO models in PyTorch," 2024, [Online]. Available: https://github.com/ultralytics/ultralytics. (Accessed: 2025-04-12).

[47] M. A. M. Rastari, R. Roslan, R. Hamzah, N. H. I. Teo, F. E. Shahbudin, and K. A. F. A. Samah, "Recycle waste detection and classification model using YOLO-V8 for real-time waste management," in *2024 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)*, Kota Kinabalu, Malaysia, 2024, pp. 372–377, doi: 10.1109/IICAIET62352.2024.10730703.

[48] N. A. Zailan, M. M. Azizan, K. Hasikin, A. S. M. Khairuddin, and U. Khairuddin, "An automated solid waste detection using the optimized YOLO model for riverine management," *Frontiers in Public Health*, vol. 10, p. 907280, 2022, doi:

10.3389/fpubh.2022.907280.
[49] A. Vaswani *et al.*, "Attention is all you need," *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA., 2017.
[50] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision*, Springer, 2020, pp. 213–229, doi: 10.1007/978-3-030-58452-8_13.
[51] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, et al., "A survey on vision transformer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 87–110, 2022.
[52] H. Qin *et al.*, "Active Learning-DETR: Cost-Effective Object Detection for Kitchen Waste," *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–15, 2024, doi: 10.1109/TIM.2024.3368494.
[53] R. Kashaf, E. P. Alegre, T. Prova, and S. Aggarwal, "Automated Waste Management using a Customized Vision-based Transformer Model," in *2024 IEEE World AI IoT Congress (AIIoT)*, Seattle, WA, USA, 2024, pp. 300-309, doi: 10.1109/AIIoT61789.2024.10578946.

BIOGRAPHIES OF AUTHORS

Owen Tamin earned a Bachelor of Applied Science (Hons) in Mathematical Modelling from Universiti Sains Malaysia in 2019 and completed his Master's degree in Computer Science at Universiti Malaysia Sabah in 2024. He is currently a Graduate Research Assistant and pursuing a Ph.D. in Mathematics. His research focuses on developing a novel scattered data interpolation scheme using cubic triangular patches and machine learning for highly accurate RGB image interpolation. He has published several articles in high-impact journals in the areas of machine learning, applied mathematics, and deep learning. He has also contributed to book chapters and international conference proceedings, particularly in the field of artificial intelligence and intelligent systems. He can be contacted at email: owentamin1996@gmail.com.

Dr. Ervin Gubin Moung is a senior lecturer at the Faculty of Computing and Informatics, Universiti Malaysia Sabah (UMS), where he earned his Bachelor's (2008), Master's (2013), and Ph.D. (2018) in Computer Engineering. His research interests generally fall under the category of computer vision and pattern recognition, with a current focus on public health & smart health, agriculture & food security, and environmental sustainability. Over his career, he has secured and led research projects totaling over MYR 1.3 million in grants from agencies such as PETRONAS, the Malaysian Ministry of Higher Education, and UMS's internal research fund. His notable projects include the PETRONAS Advanced Analytics Framework for Resource Optimization (PETRO-AFRO Framework), algae performance prediction tools, and spectrally adapted algorithms for assessing beef cattle health. Since July 2022, he has also served as Deputy Director at UMS's Research Management Centre, overseeing all operational activities. His contributions extend to innovations like hyperspectral imaging for plastic waste detection, solar-powered black pepper dryers, and sentiment analysis tools for business analytics. With 88 publications, 569 citations, and a Scopus h-index of 13, he is a respected researcher committed to advancing computer engineering and interdisciplinary research excellence. He can be contacted at email: ervin@ums.edu.my.

Ali Farzamnia received the B.Eng. degree in Electrical Engineering (Telecommunication Engineering) from Islamic Azad University, Urmia, Iran, in 2005, the M.Sc. degree in Electrical Engineering (Telecommunication Engineering) from the University of Tabriz, in 2008, and the Ph.D. degree in Electrical Engineering (Telecommunication Engineering) from the Universiti Teknologi Malaysia (UTM), in 2014. He is currently a lecturer at the School of Computing and Engineering, University of Huddersfield, United Kingdom. He has secured several research grants and collaborated with numerous research partners. His research interests include wireless communications, signal processing, network coding, information theory, and biomedical signal processing. He is a member of IET and is a Chartered Engineer (C.Eng.) in the U.K. He can be contacted at email: a.farzamnia@hud.ac.uk.