Benchmarking computer vision models for automated construction waste sorting

Zhiming Dong, Liang Yuan, Bing Yang, Fan Xue, Weisheng Lu*

Department of Real Estate and Construction, Faculty of Architecture, The University of Hong Kong, Pokfulam, Hong Kong
* Corresponding email: wilsonlu@hku.hk

This is the peer-reviewed post-print version of the paper:
Dong, Z., Yuan, L., Yang, B., Xue, F., & Lu, W. (2025). Benchmarking computer vision models for automated construction waste sorting. *Resources, Conservation and Recycling.* Doi: 10.1016/j.resconrec.2024.108026.
The final version of this paper is available at: https://doi.org/10.1016/j.resconrec.2024.108026.

Abstract

Waste sorting is a critical process in construction waste management system. Computer vision (CV) offers waste sorting automation potential by recognizing waste composition and instructing robots or other mechanical devices accordingly. However, how the plethora of CV models developed perform relative to each other remains underexplored, making model selection challenging for researchers and practitioners. This study aims to benchmark existing CV models towards automated construction waste segregation. Seventeen models were selected and trained with unified configuration, and then their performance was evaluated on the aspect of accuracy, efficiency, and robustness, respectively. In experimental results, BEiT attained top accuracy (58.31% MIoU) while FastFCN had the best efficiency (12.87ms). SAN displayed the least standard deviation (4.41%) for robustness evaluation. This research contributes a reliable reference for CV model selection, advancing automated construction waste sorting research and practices, and ultimately promoting efficient recycling while reducing the environmental impact of construction and demolition waste.

Keywords: Construction Waste Management; Waste Sorting; Computer Vision; Benchmarking; Composition Recognition

1. Introduction

Construction waste, also known as construction and demolition (C&D) waste, is the solid waste stream arising from construction activities such as new building construction, renovation, demolition, foundation excavation, and site formation (Lu et al., 2021). Mainly comprising concrete, bricks, tiles, soil, rock, metals, timber, gypsum, asphalt, plastics, textiles, and vegetation, C&D waste accounts for a large share of total solid waste generated around the world (Yuan et al., 2024). For instance, it contributed 37.1% of solid waste across all sectors in the European Union in 2020 (European Commission, 2023), 67.3% of solid waste in the United States in 2018 (USEPA, 2023), 44.0% of solid waste in Australia in 2020 (Gov.AU, 2021), and 40.0% of solid waste in China in 2017 (Zhao et al., 2023). Most components of C&D waste are chemically stable, making it less hazardous than other solid waste streams (e.g., municipal solid waste and household waste); however, the massive amount of C&D waste generated from construction activities amplifies impacts on environment and society, including landfill space occupation (Yuan et al., 2021), embodied carbon emission (Liu et al., 2023), illegal dumping (Lu, 2019), and so on. Therefore, construction waste management (CWM) is a growing environmental protection concern globally.

Recycling is a prioritized CWM strategy as most C&D waste components are recyclable. Specific recycling methods vary depending on material composition, while construction waste is usually a mixture of materials when generated at source (Yuan et al., 2021). Waste sorting, which means separating mixed waste into different composition groups, is therefore a critical procedure in CWM (Hoornweg & Bhada-Tata, 2012). In practice, contractors either employ workers to manually sort construction waste on site, or they transport it to off-site sorting facilities where, on payment of a levy, the bulk waste is separated using specialized equipment and then the residual mixtures are sorted manually. Manual construction waste sorting, albeit common, has been widely criticized for its low efficiency and the serious health risks it poses to workers (Lu & Chen, 2022).

In promoting a shift from manual sorting to automated separation, computer vision (CV) enables machinery (e.g., robots) to automatically recognize and separate different C&D waste materials (Lu & Chen, 2022). Scholars have developed diverse CV models for automated construction waste sorting. These have achieved promising performance by narrowing specific gaps, such as integrating advanced algorithms to improve recognition accuracy (Davis et al., 2021; Zhou et al., 2022), infer the overall waste content according to surface composition identification (Chen et al., 2021), precisely distinguish distinct waste materials' mixing boundaries (Dong et al., 2022), recognize compositions in complex real-life scenarios (Lu, Chen, et al., 2022; Yong et al., 2023), reach acceptable performance on unbalanced and insufficient data (Na et al., 2022), and optimize feature selection to balance recognition accuracy and speed (Nežerka et al., 2024), making significant contributions to CV-enabled construction waste sorting.

A large number of CV models have been proposed, yet how these models perform in comparison to one another remains underexplored. While some model development studies have conducted cross-comparisons to highlight the strengths of their proposed models, these are primarily with limited comparison targets (Lu, Chen, et al., 2022; Nežerka et al., 2024) on task-specific datasets (Demetriou et al., 2023; Dong et al., 2022). The lack of systematic and unified comparisons between existing models creates a significant research gap. Academically, this gap hinders researchers from identifying the advantages and drawbacks of various CV methods for CWM, slowing progress in the development of more robust and accurate CV methods for automated construction waste sorting. Practically, this absence of systematic comparison complicates the selection process for system developers and practitioners, potentially leading to suboptimal performance in automated waste sorting systems, which in turn exacerbates waste management issues and increases environmental impact.

Based on the research gap, the primary research question guiding this study is: How do existing CV models for construction waste sorting perform against each other in real-life CWM scenarios. Therefore, the research objective is to systematically benchmark existing CV models developed for automating construction waste sorting across various measurement dimensions under a unified standard. Such benchmarking benefiting from a large image dataset that presents the visual features of construction waste bulks in real-life scenarios, the result is compared across various measurement dimensions, including accuracy, efficiency, and robustness, to provide model selection suggestions. Furthermore, additional valuable opinions are revealed and discussed, such as challenges ahead and future improvement opportunities. The remainder of the paper is organized as follows. Section 2 reviews previous related studies, Section 3 introduces the research methodology and Section 4 describes the benchmarking experiment procedure and result. Section 5 discusses the research contributions and shortcomings, and conclusions are drawn in Section 6.

2. Literature review

2.1 Computer vision models proposed for construction waste sorting

Waste sorting plays an essential role in CWM, separating the mixed waste by individual composition and allowing the implementation of appropriate recycling and reuse strategies (Hoornweg & Bhada-Tata, 2012). Still common today, manual sorting has been criticized for its low efficiency and the health hazards it presents to workers (Lu & Chen, 2022). To address these issues, scholars have proposed various strategies to automate the sorting of C&D waste. Of these technologies, one of the most promising is the implementation of CV models to help robotic arms detect and separate different construction and demolition waste materials (Lu & Chen, 2022).

CV models exhibiting a wide range of techniques, architectures, and applications have been developed for automated construction waste sorting. Some studies focus on classification tasks using deep convolutional neural networks (Davis et al., 2021) or hybrid approaches combining visual and physical features (Chen et al., 2021). Others tackle object detection challenges with improved YOLOv5 models (Zhou et al., 2022) or real-time detection using single-stage and two-stage detectors (Demetriou et al., 2023). Semantic segmentation is another prominent approach, with models like the boundary-aware transformer (Dong et al., 2022) and DeepLabv3 (Lu, Chen, et al., 2022; Yong et al., 2023) being applied to recognize waste compositions in complex real-world scenarios. Additionally, some studies address specific challenges, such as handling unbalanced and insufficient data (Na et al., 2022), comparing feature extraction methods (Nežerka et al., 2024), or calculating waste volume (Jiang et al., 2022).

2.2 Performance comparison of CV models for waste sorting

Given the broad landscape of CV models and their implementation in C&D waste sorting, it is crucial to evaluate and compare their performance so that researchers and practitioners in the construction industry can identify the most effective approaches as a guide for their future efforts. The studies we reviewed employ a variety of evaluation metrics, which can be categorized into three main groups: accuracy metrics, efficiency metrics, and robustness metrics. Accuracy metrics, such as overall accuracy and intersection over union (IoU), measure the correctness of the model's predictions compared to the ground truth. For instance, Davis et al. (2021) and Chen et al. (2021) both reported an accuracy of 94% for their respective models. Yong et al. (2023) achieved an accuracy of 96.3% and an IoU of 74.6%. Efficiency metrics, such as computing time and frames per second (FPS), evaluate the computational efficiency of the model, which is critical for real-time applications. Demetriou et al. (2023) reported an inference speed of less than 30ms for their model. Robustness metrics, including deviation and standard deviation, assess the model's ability to correctly identify positive instances on data with diverse distributions (Dong et al., 2020).

Although existing methods have generally employed multi-dimensional metrics to evaluate performance, the scope of such comparisons remains limited, and the standards are inconsistent. Zhou et al. (2022) proposed a novel computer vision method for construction waste object detection, comparing it with four baseline methods on a self-collected dataset. The results showed that their method achieved a mean Average Precision of 0.9480. Similarly, Dong et al. (2022) introduced a construction waste composition recognition method based on a vision transformer, comparing its performance with four baseline methods on a dataset collected from waste disposal facilities in Hong Kong, with an improvement of 5.48% over the baseline methods. Davis et al. (2021) developed a construction waste material classification method, evaluating its performance by comparing three different hyperparameter configurations on a self-collected dataset, which yielded 94% accuracy. In summary, the primary focus of current computer vision research in automated construction waste sorting is on model development, with evaluations aimed at emphasizing the strengths of the proposed models (Demetriou et al., 2023). Consequently, variations in datasets, experimental setups, and evaluation metrics persist (Nežerka et al., 2024), leading to a significant research gap due to the absence of systematic and unified comparisons among existing models.

The innovation of this research lies in conducting a systematic and comprehensive benchmarking study to address the aforementioned limitations and enable meaningful comparisons. This study utilized both a consistent, representative dataset and standardized evaluation metrics for unified comparison. It tested a wide range of C&D waste sorting models to provide a fair and reliable assessment of their performance. By establishing a common ground for comparison, this benchmarking study offers valuable insights into the strengths and limitations of different approaches, identifies areas for improvement, and guides future research efforts in the field of CV-based construction waste sorting.

3. Methodology

In this section, the methodology of conducting the benchmark study will be illustrated. The overall technical roadmap includes four steps.

1) *Dataset preparation*. A dataset includes totally 4,396 common local construction waste images is prepared for evaluation.
2) *Model selection*. Seventeen models with different popular architecture or technique are selected.
3) *Model training*. All models are trained to convergence with unified standard and same configuration.
4) *Performance evaluation*. The performance of each model is evaluated on the aspect of accuracy, efficiency, and robustness.

3.1 Dataset preparation

With the assistance of staff from the Hong Kong Environmental Protection Department (HKEPD), this research gathered a comprehensive dataset of construction waste truckloads in Hong Kong (Chen et al., 2022). The raw data is sourced from sensing systems installed at multiple government disposal facilities, which record different information including overhead images taken by cameras (DS-2CD2025FWD-IHONG KONG 4 mm from Hikvision Digital Technology) positioned at the facility's entrance toll gates, along with detailed attributes of truck-loaded waste, such as gross vehicle weight, net waste weight, and waste loading height, recorded by additional sensors. All images captured in October 2019 (5,366 images in total) were collected for this research.

Within Hong Kong's CWM system, materials are predominantly classified into one of two categories: inert or non-inert (HKEPD, 2011). Inert materials encompass construction debris, rubble, earth, bitumen, and concrete, while non-inert materials typically consist of bamboo, timber, vegetation, packaging waste, and other organic substances. To facilitate enhanced granularity, the dataset incorporates 10 categories derived from common local construction waste, with seven representing material types and three serving as auxiliary classifications, as summarized in Table 1. With the taxonomy, all images were annotated with pixel-level semantic labels through crowdsourcing services. Each annotation was then manually checked. Finally, by pre-filtering a few low-quality annotations, an annotated dataset of 4,396 images was constructed. The waste dataset comprises 4,017 images for the training subset, 200 images for the validation subset, and 179 images for the evaluation subset.

Table 1. The waste dataset information categories

Name	Category	Description	Palette
Rock	Inert material	Rock/Stone/Rubble/Debris	
Gravel	Inert material	Gravel/Concrete/Bricks	
Earth	Inert material	Earth/Slurry/Mud	
Packaging	Non-inert material	Packaging/Fabric/Plastic	
Wood	Non-inert material	Wood/Cardboard	
Others	Non-inert material	Others (non-inert)	
Mixed	Mixed material	Mixed	
Background	Auxiliary	Background	
Grip	Auxiliary	Grip	
Truck	Auxiliary	Truck	

3.2 Model selection

As waste objects depicted in images often possess no fixed shape or location, image classification and object detection methods struggle to obtain sufficiently accurate information (Dong et al., 2020) for downstream processing of C&D waste sorting. Consequently, semantic segmentation methods merit exploration for application to C&D waste sorting (Lu & Chen, 2022). For semantic segmentation models, various architectures and techniques to enhance performance have been extensively explored. In this study, the segmentation method is broadly divided into three categories based on architecture: convolutional neural network (CNN)-based, attention-based, and transformer-based (Csurka et al., 2022). Within each architectural category, researchers have proposed diverse techniques to improve performance. For each technique in this study, we select one or two representative methods for benchmarking, enumerated in Table 2.

Table 2. Selected models for benchmarking

Index	Name	Technique	Architecture
1	FCN (Long et al., 2015)	Deconvolution	CNN
2	PSPNet (Zhao et al., 2017)	Pyramid pooling	CNN
3	DeepLab V3 (Chen et al., 2017)	Atrous spatial pyramid pooling	CNN
4	UPerNet (Xiao et al., 2018)	Unified perceptual parsing	CNN
5	DeepLab V3+ (Chen et al., 2018)	Atrous spatial pyramid pooling	CNN
6	FastFCN (Wu et al., 2019)	Joint pyramid upsampling	CNN
7	OCRNet (Yuan et al., 2020)	Object-contextual representation	CNN
8	PSANet (Zhao et al., 2018)	Pointwise spatial attention	Attention
9	ENCNet (Zhang et al., 2018)	Context encoding module	Attention
10	Non-Local Net (Wang et al., 2018)	Non-local operation	Attention
11	APCNet (He et al., 2019)	Adaptive context module	Attention
12	DANet (Fu et al., 2019)	Dual attention module	Attention
13	ISANet (Huang et al., 2019)	Interlaced sparse self-attention	Attention
14	ViT (Dosovitskiy et al., 2020)	Transformer	Transformer
15	BEiT (Bao et al., 2021)	Pretrained transformer	Transformer
16	SAN (Xu et al., 2023)	Pretrained multimodal	Transformer
17	Grounded SAM (Ren et al., 2024)	Vision foundational model	Transformer

3.3 Training

We established a unified model training platform for model training. In terms of hardware configuration, all models were trained and evaluated on a workstation equipped with an i9-13900KF central processing unit, an RTX 4090 graphics processing unit, and 128GB of memory. As for the software configuration, the operating system employed is Ubuntu 22.04, with the Python interpreter version being 3.8. The comprehensive benchmarking experiments were conducted using MMSegmentation 1.2.2 and PyTorch 2.0, with the entire platform constructed upon CUDA 11.8.

The training hyperparameters are set to the default configuration of the MMSegmentation framework, which emphasizes usability, consistency, and standardization, with proven performance comparable to other codebases (OpenMMLab, 2020). To ensure a fair comparison, hyperparameters remain consistent across all models. Specifically, we utilize the stochastic gradient descent optimizer (Amari, 1993) to optimize model parameters. The PolyLR (Mishra & Sarawadekar, 2019) serves as the learning rate adjustment policy, with $lr = max\{lr_0 \times (1 - i/T_i)^{power}, lr_{min}\}$, where the initial learning rate lr_0 is set to 10^{-2}, the polynomial power is set to 0.9, the total number of iterations T_i is set to 8×10^4, and the minimum learning rate at the end of scheduling is set to 10^{-4}. Batch size is determined based on the highest allowable size under the memory constraint. This study consistently uses a batch size of 4, guided by the memory demands of the BEiT model, to ensure experimental fairness.

The input image size for segmentation models is set to 512×512, adhering to the default configuration for model training. Consequently, a transformation is necessary to convert images from 1920×1080 to 512×512. This process involves randomly resizing the image, then flipping and cropping it to 512×512 dimensions. Additionally, some preprocessing techniques are adopted to enhance the model's generalization ability, including random adjustments to brightness, contrast, saturation, and color system (OpenMMLab, 2020).

3.4 Evaluation

3.4.1 Accuracy evaluation

To evaluate performance in terms of accuracy, we use the MIoU metric. Widely adopted in the research community, this comprehensive metric provides an overall measure of the segmentation accuracy of a model across all classes, making it easier for us to compare and benchmark different models. MIoU can be computed using Formula (1), where p_{ij} denotes the number of pixels belonging to the i_{th} category but predicted as the j_{th} category. k represents the total number of categories.

$$MIoU = \frac{1}{k}\sum_{i=1}^{k}\left(p_{ii}/\left(\sum_{j=1}^{k} p_{ij} + \sum_{j=1}^{k} p_{ji} - p_{ii}\right)\right) \quad (1)$$

3.4.2 Efficiency evaluation

Computation time is employed to assess efficiency performance. The efficiency of CV algorithms is a crucial factor to consider in practical applications. Although efficiency may be less critical in certain situations, such as smartphone-assisted household waste classification, stringent time performance is necessary in scenarios like automated waste sorting with robotics. This is particularly relevant when considering that a waste recycling facility can process thousands of tons of municipal solid waste per day, and sorting speed directly impacts overall throughput (Lu & Chen, 2022).

3.4.3 Robustness evaluation

The robustness evaluation aims to assess the stability of semantic segmentation methods' performance on various waste distributions. Lu, Yuan, et al. (2022) found that gross vehicle weight and vehicle type significantly affect waste distributions. Heavy inert construction waste is typically transported by dump trucks, while light construction waste is often hauled by skip container trucks due to their large-sized truck buckets. Skip container trucks can transport any type of construction waste. Regarding the influence of gross vehicle weight on waste distribution, heavier trucks are commonly used for transporting heavy construction waste. Lighter trucks, usually vans, are employed by small renovation companies for refurbishment or minor works.

This loading pattern can be utilized to divide the evaluation subset into distinct groups. Vehicle types can be categorized into grab mount, dump, and skip container, while gross vehicle weight can be split into six types ranging from 9 to 30 tons. Table 3 illustrates the data split results. There are a total of nine groups; the 16-ton skip container group contains 19 images, and all other groups have 20 images. Consequently, the evaluation subset comprises 179 images. The pie chart in each group displays the data category distribution, revealing significant differences among categories. During the robustness evaluation, the MIoU of each group is initially calculated, and the standard deviation across all groups represents the robustness.

Table 3. Data category distribution of the evaluation subset

Group Index	Vehicle Type	Gross Vehicle Weight (ton)	Proportion of each category (%)						
			Rock	Gravel	Earth	Packaging	Wood	Others	Mixed
1	Skip container	16	4.9	19.1	7.8	13.9	42.3	5.9	6.0
2		24	33.8	20.8	3.8	27.5	10.9	0.3	3.0
3	Grab mount	16	11.9	10.0	0.2	7.3	47.1	19.2	4.3
4		24	5.2	13.5	5.4	20.5	21.9	13.2	20.3
5		30	8.5	3.6	8.8	8.7	33.4	3.0	34.0
6	Dump	9	0	1.9	0.4	14.4	61.6	21.6	0.1
7		13	0.5	8.8	0.3	10.0	56.5	16.0	7.9
8		14	4.7	7.2	1.1	3.6	56.8	22.6	3.9
9		16	2.2	9.8	0	7.6	57.4	21.6	1.4

Specifically, each model is firstly evaluated by MIoU on each group, and the average of all those MIoUs is recorded as $MIoU_{mean}$, while the standard deviation is recorded as $MIoU_{SD}$. The $MIoU_{SD}$ serves as an indicator of the method's robustness. The calculation formula is shown in Formula (2) and (3), where n is the number of loading patterns, n=9 in this case, $MIoU_i$ refers to the evaluation result on i-th group.

$$MIoU_{mean} = \frac{1}{n}\sum_{i=1}^{n} MIoU_i \tag{2}$$

$$MIoU_{SD} = \sqrt{\sum_{i=1}^{n}(MIoU_i - MIoU_{mean})^2/n} \tag{3}$$

4. Benchmarking experiments

In this Section, we analyze the benchmarking results of the 17 methods selected for comparison. It is worth noting that Grounded SAM is used in a zero-shot learning manner, meaning it is not trained on the training subset and is evaluated with the evaluation subset directly. All other methods are fully trained on the training subset and then evaluated using the evaluation subset. The accuracy, efficiency, and robustness performance aspects of all methods are discussed from Section 4.1 to Section 4.3.

4.1 Accuracy analysis

The overall accuracy evaluation results are presented in Table 4. The header of table contains the category name and corresponding color palette. The table content displays the evaluation results. Each row lists the detailed IoU value for each category of a specific method. Among all 17 methods, BEiT achieved the best performance with an MIoU value of 58.31%. BEiT performed well in most categories, including background, gravel, packaging, wood, others, grip, and truck. The vast number of parameters enables BEiT to learn rich, contextual representations from both local and global image features, therefore contributed to BEiT's overall superior results compared to other methods.

Table 4. Overall accuracy evaluation result.

Name	ID	Back-ground	Rock	Gravel	Earth	Pack-aging	Wood	Others	Mixed	Grip	Truck	MIoU
Palette												
FCN	1	93.27	32.04	41.56	35.57	55.57	69.72	39.58	31.89	90.39	77.32	56.691
PSPNet	2	93.09	27.03	39.58	34.29	56.76	69.76	39.18	33.31	90.25	77.5	56.075
DeepLabv3	3	93.28	34.53	41.61	32.81	56.37	70.51	39.28	33.86	90.4	77.35	57
UPerNet	4	93.33	36.13	39.58	36.32	57.11	69.93	38.58	34.21	90.36	76.87	57.242
DeepLabv3+	5	93.34	32.07	44.45	34.57	55.94	69.94	38.61	33.26	90.31	77.48	56.997
FastFCN	6	93.25	33.93	39.25	35.65	56.01	68.34	38.31	33.12	90.12	76.82	56.48
OCRNet	7	93.19	36.27	42.73	38.19	56.12	69.4	39.11	33.28	90.24	76.74	57.527
PSANet	8	93.3	37.85	42.64	36.4	56.47	70.28	40.5	36.34	90.16	77.23	58.117
ENCNet	9	93.27	35.43	41.72	33.58	55.98	69.68	38.03	35.11	90.24	77.23	57.027
Non-local Net	10	93.28	33.26	42.33	35.91	56	69.82	38.73	35.03	90.32	77.38	57.206
APCNet	11	93.32	36.36	41.63	33.75	55.35	70	38.57	32.55	90.29	77.24	56.906
DANet	12	93.25	41.2	37.59	39.31	56.41	69.82	37.28	35.49	89.87	76.62	57.684
ISANet	13	93.36	34.23	39.27	31.32	55.85	69.5	40.65	33.57	90.17	77.17	56.509
ViT	14	93.24	34.8	36.26	38.7	55.25	67.69	37.07	33.96	88.33	75.89	56.119
BEiT	**15**	**93.54**	**34.02**	**42.96**	**38.4**	**58.26**	**71.56**	**41.48**	**35.05**	**90.22**	**77.61**	**58.31**
SAN	16	92.9	32.9	42.1	36.5	53.2	67.8	36.2	35.2	87.3	72.7	55.68
Grounded SAM	17	51.58	5.36	21.07	0.73	2.58	3.85	0.89	0	1.54	14.48	10.208

For the other results, OCRNet achieved the best performance with an MIoU value of 57.53% when comparing with other models with CNN architecture, such results mainly contributed by the good performance in the rock and earth categories. PSANet achieved the best performance across all attention-based methods, with an overall MIoU result of 58.12%. The strong performances con categories packaging, wood, and mixed contribute positively to PSANet's overall good performance.

The MIoU of Grounded SAM is 10.21%, a significant performance gap compared with other fully supervised methods. Despite the overall accuracy performance being subpar, there are noteworthy aspects that showcase the considerable promise of zero-shot learning methods, such as Grounded SAM, where the model needs no training phase and can be directly utilized for C&D waste sorting. For the output of Grounded SAM, some good samples prove it can pinpoint basic objects and discern their contours. However, failure examples also reveal the instability of Grounded SAM's performance, as it struggles with scenarios involving multiple waste types or catastrophic errors in object category identification.

In real-time waste sorting applications, accuracy is the most crucial factor for successful operations. Highly accurate models ensure that materials are properly categorized, preventing contamination and costly errors in waste management. While efficiency and robustness are important, accuracy remains the cornerstone of effective real-world waste sorting systems. Engineering managers should prioritize accuracy, even if it requires sacrificing some efficiency. When trade-offs are necessary between accuracy, efficiency, and robustness, they must be carefully weighed.

4.2 Efficiency analysis

Computing time is utilized to assess efficiency performance. For the overall evaluation subset, the average computing time of all images is recorded and employed for evaluation. The evaluation result is depicted in Fig. 1. Fig. 1 is a combination chart includes a bin chart and a line chart. The MIoU value can be obtained from the primary axis on the left, and it is displayed in blue bar chart as a percentage. The calculation time value is obtained from the secondary axis on the right, and it is displayed in orange line chart as milliseconds.

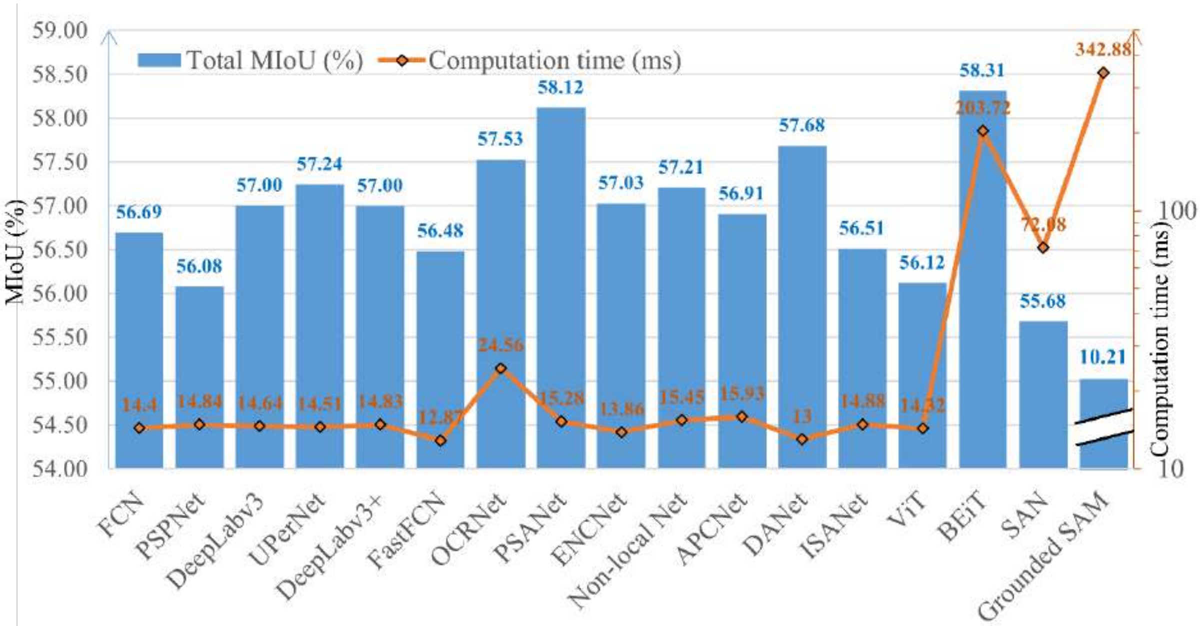


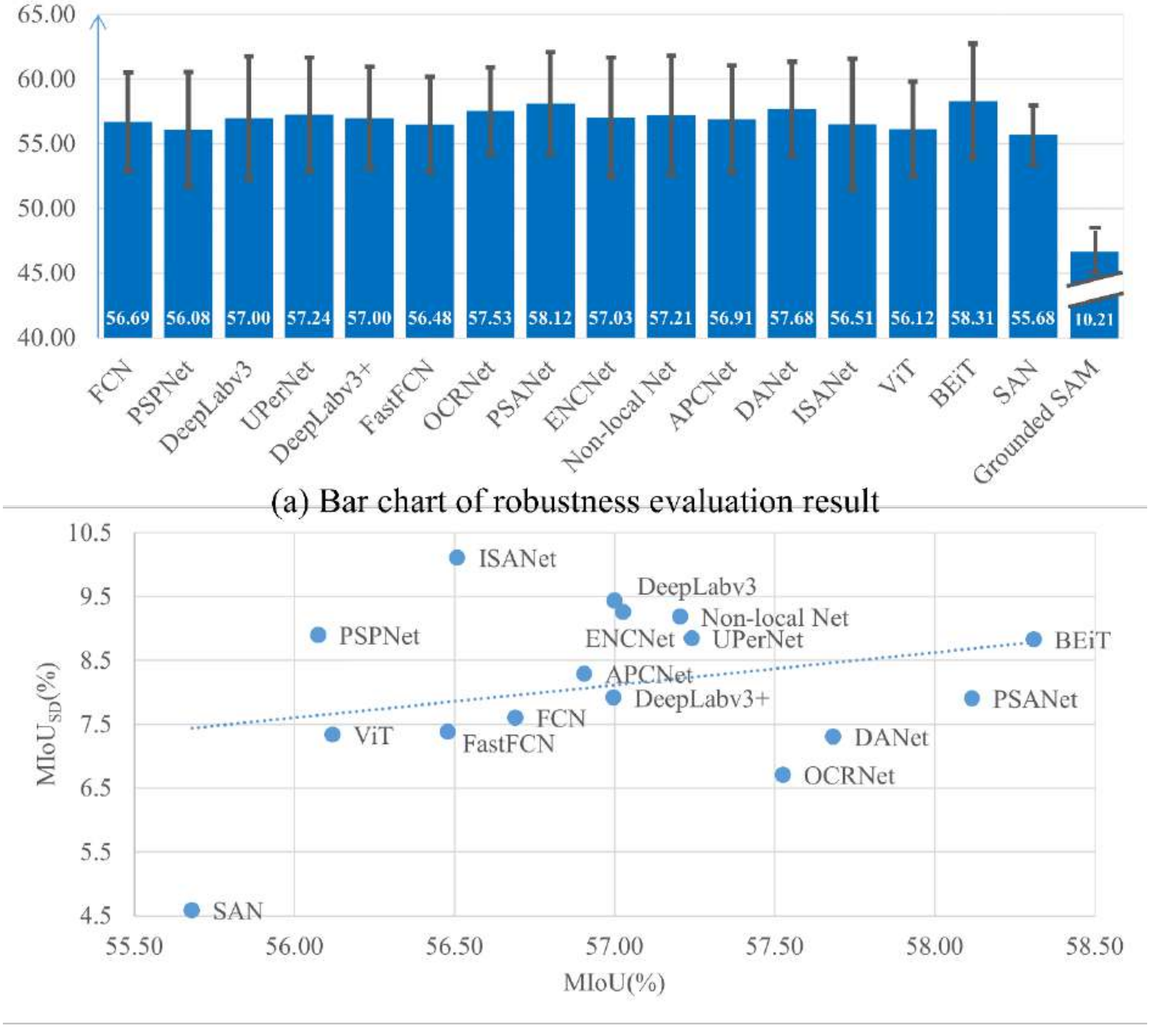
Fig. 1. Efficiency evaluation result. The blue bin chart represents the MIoU performance of each model, while the orange points on the line chart indicate the corresponding computing time.

Fig. 1 illustrates that both the CNN-based and attention-based methods consistently achieved a computing time below 20ms for an input image size of 1920×1080. The FastFCN method demonstrated the fastest computation time, requiring only 12.87ms to calculate the segmentation result for a single image. Such fast inference primarily due to its Joint Pyramid Upsampling module to minimizes the computational complexity and memory usage. Among all the CNN-based and attention-based methods, OCRNet exhibited the longest computation time, which amounted to 24.56ms to allow its Object-Contextual Representation module to capture detailed relationships between pixels and their surrounding context. However, it also achieved the highest accuracy performance among all the CNN-based and attention-based methods. In terms of efficiency performance, Grounded SAM's computing time was 342.88ms, ten times that of CNN-based and attention-based methods due to the computational complexity of integrating visual segmentation with language grounding. However, such latency remains acceptable for practical C&D waste sorting tasks (Lu & Chen, 2022).

Efficiency performance holds significant practical importance. Depending on the speed of sorting hardware, such as conveyors and robotic arms, real-world applications exhibit varying sensitivities to latency. In high-speed sorting lines where waste moves rapidly, models with low latency should be prioritized. Lower latency typically allows a CV model to process more frames per second, accelerating the waste sorting process and enhancing the efficiency and throughput of the entire system. Conversely, if the system is less sensitive to latency, models with higher accuracy can be considered to reduce errors in automated waste sorting and improve overall performance.

4.3 Robustness analysis

A robust model ensures that the system can reliably handle real-world waste management tasks, especially C&D waste with the characteristic of bulky and heterogeneous. Therefore, as detailed in Section 3.4.3, the deviation on different data distribution patterns $MIoU_{SD}$ is used to evaluate the robustness. There are totally nine group of loading pattern, which is the combination of different vehicle types and gross vehicle weight, are used to calculate the deviation.



(a) Bar chart of robustness evaluation result

(b) Scatter chart with trendline of the relationship between MIoU and $MIoU_{SD}$

Fig. 2. Robustness evaluation result. For sub-figure (a), blue bar refers to the MIoU performance of each model, while the error bar refers to the robustness metric $MIoU_{SD}$ calculated by formula (3). For sub-figure (b), horizonal axis is the accuracy evaluation result represented by MIoU, vertical axis refers to the robustness evaluation result represented by $MIoU_{SD}$. The blue dotted line is the trendline of the relationship between MIoU and $MIoU_{SD}$.

The robustness results are presented in Fig. 2. In Fig. 2 (a), the bars represent the MIoU values on the total evaluation dataset, while the error bars indicate the corresponding $MIoU_{SD}$ values. A smaller $MIoU_{SD}$ value indicates better robustness performance. SAN exhibits the smallest $MIoU_{SD}$ of 4.58%, making it the most robust method. ISANet demonstrates the largest $MIoU_{SD}$ value of 10.11%, indicating the lowest level of robustness. Despite BEiT's outstanding accuracy performance, its $MIoU_{SD}$ value stands at 8.83%. Concerning robustness performance, Grounded SAM's $MIoU_{SD}$ is 1.49%, the smallest among all methods. However, such a low $MIoU_{SD}$ does not guarantee robustness as its MIoU is also low, indicating that Grounded SAM's fluctuation range must remain minimal, resulting in a low $MIoU_{SD}$ value.

Fig. 2(b) illustrates a correlation between robustness and accuracy. Grounded SAM is excluded from this analysis due to its significantly different accuracy performance. A linear trendline is applied to fit the correlation, which is calculated using the least squares method. The results show that MIoU and $MIoU_{SD}$ are positively correlated, indicating that higher accuracy generally corresponds to lower robustness. This is likely because models highly optimized for accuracy on a specific dataset tend to overfit to details or noise within that data (Tsipras et al., 2018). The reliability of this trend is measured by the R-squared value of 0.0774, suggesting the correlation is not absolute, and that the complexity of the model architecture significantly impacts performance evaluation. There are several noteworthy deviations from the linear relationship between accuracy and robustness. For example, OCRNet exhibits the second-best robustness after SAN, with an $MIoU_{SD}$ of 6.7%, yet its MIoU is 1.85% higher than SAN. PSANet, with the second-highest accuracy at 58.12%, demonstrates more stable performance in robustness evaluation, with an $MIoU_{SD}$ of 7.9%.

This robustness analysis highlights an important consideration when evaluating accuracy: a model must not only perform well under optimal conditions but also consistently handle diverse real-world scenarios. In real-world waste sorting, when waste distribution is relatively simple and stable over time and space, users can prioritize model accuracy. However, when waste is highly heterogeneous or bulky, both accuracy and robustness must be considered, and a balance between the two should be struck. This ensures that the CV model can function effectively in ideal conditions while providing reliable information for the automatic sorting of construction waste.

5. Discussion

This research presents a benchmark study that compares the performance of various CV models in the task of automated construction waste sorting. Compared with model development studies, this work offers a broader and more objective comparison by evaluating CV models with different architectures under a unified standard. The key advantage of this approach is its ability to ensure consistent evaluation across models, enabling more informed decisions when selecting models for specific applications, thereby enhancing practical deployment. Compared with general benchmark studies on public datasets, this research focuses on the specialized task of construction waste sorting, which presents unique challenges due to domain gaps. Waste objects are highly cluttered and heterogeneous, making this task significantly different from general vision tasks. The strength of this study lies in providing benchmark results on a construction waste sorting dataset, offering domain-specific insights crucial for advancing waste management automation. Based on the comparison results, valuable insights regarding model selection, upcoming challenges, and future research recommendations are detailed in Sections 5.1 to 5.3.

5.1 Suggestions for model selection

The evaluation results of all fully supervised methods from this research indicate that in terms of accuracy, the MIoU varies from 55.68% to 58.31%, with BEiT achieving the best performance. In terms of efficiency, the computation time ranges from 12.87ms to 203.72ms, and FastFCN exhibits the best performance. Regarding robustness, the standard deviation of

MIoU across different data distribution patterns varies between 4.48% and 10.11%, with SAN demonstrating the best performance.

While the assessment outcome is clear-cut, potential users and researchers should weigh the tradeoffs among these three aspects for successful implementation. For instance, even though BEiT boasts the best performance, it requires 204ms to process a single image, over 10 times longer than that required by CNN-based and attention-based methods, rendering it potentially unsuitable for time-sensitive C&D waste sorting situations. Consequently, we can offer recommendations for model selection based on these evaluation findings.

1) *General performance*: The balanced mode necessitates a compromise between accuracy, efficiency, and robustness. This widely preferred mode can be considered the default choice. PSANet is the most fitting model for this mode, as its MIoU is 58.12%, only 0.19% lower than the best performance. It requires 15.28ms for computation, which is a mere 2.41ms longer than the best performance. The standard deviation is 7.91%, which is 3.43% higher than the best performance.
2) *Accuracy-prioritized*: For users who need extremely accurate performance and can tolerate longer computing times, BEiT is the most suitable model. It has the highest MIoU of 58.31% and performs well in both overall categories and waste category evaluations.
3) *Efficiency-prioritized*: For users who require a CV method for real-time C&D waste sorting tasks or have limited hardware capabilities, FastFCN meets their needs. It requires only 12.87ms for computation and achieves a MIoU of 57.53% with a 7.38% standard deviation.
4) *Robustness-prioritized*: For users who prioritize prediction stability in C&D waste sorting tasks with a wide variety of waste types, mixture patterns, or distributions, SAN is the most suitable option due to its low standard deviation of 4.59%.
5) *Annotation-free*: For users who lack the resources to annotate custom C&D waste sorting data or only require basic semantic understanding without any workload, Grounded SAM is the ideal choice. Based on general visual understanding capabilities, Grounded SAM can provide semantic segmentation results for C&D waste sorting in a zero-shot manner.

5.2 Challenges ahead

The CV-based automated waste sorting system can foster a transition from manual sorting to automated separation. Nonetheless, this benchmarking study identifies certain drawbacks that could present obstacles in translating current academic endeavors into practical applications.

Firstly, the low quality of C&D waste sorting data hinders the accuracy of CV methods. C&D waste sorting datasets are often captured under complex real-life conditions where waste objects lack clear boundaries and fixed shapes. These objects are stacked together, resulting in construction waste that is bulky, heterogeneous, cluttered, and heavily overlapping (Lu & Chen, 2022). Even manual identification of construction waste features from images can be challenging due to these factors (Dong et al., 2022). This presents a challenge for the accuracy performance of CV methods applied to C&D waste sorting.

The second challenge arises from the inherent data imbalance present in construction waste data. Due to the varying demands for different types of materials during construction processes, there are significant disparities in the waste quantity distribution. In cases of imbalanced data, the model tends to be biased towards the majority class, which has more samples to learn from (Dong et al., 2020). This imprecision in C&D waste sorting can result in inaccurate component estimation. More critically, the majority class often consists of auxiliary categories such as background, whereas the minority class typically includes waste types that researchers are more interested in (Lu & Chen, 2022). Although accurately identifying auxiliary categories can assist the model in locating and recognizing waste, the data imbalance negatively affects the recognition results in a way that cannot be overlooked.

The final challenge is the time consumption of transformer-based models. These models, containing billions of parameters, offer benefits such as improved performance and reduced annotation workload. However, they also present challenges in computational requirements and memory usage. Evaluation results indicate that CNN and attention-based models typically take about 15ms for a single image, whereas transformer-based models require around 200ms for individual tasks. Notably, while experimental results are conducted on an RTX 4090 GPU, actual computing speeds may vary depending on the hardware, as only a few platforms possess such powerful GPUs. This discrepancy may result in significant latency for real-world C&D waste sorting computing tasks.

5.3 Future research recommendations

Considering the current challenges and limitations, future research may explore the following avenues. First, researchers can address low-quality data through data augmentation, cleaning, or implementing more robust architectures (Lu & Chen, 2022). Second, data imbalance can be tackled using resampling techniques, such as over-sampling for minority categories or under-sampling for majority categories (Khushi et al., 2021). Researchers are also exploring modified loss functions to address imbalance, including weighted cross-entropy loss or focal loss (Dong et al., 2020). Lastly, the time consumption issue of transformer-based vision foundation models can be managed by deploying more powerful computing platforms or utilizing cloud services to offload local computing power (Ni et al., 2024). Although transformer-based methods demonstrate high performance and potential in C&D waste sorting, users must consider all factors and select the most suitable methods for their specific tasks. These efforts ultimately contribute to the practical significance of construction waste management in engineering projects, by reducing environmental impacts, promoting recycling, curbing illegal dumping, and fostering sustainable economic growth.

6. Conclusions

CV offers significant potential for facilitating the transition from manual sorting to automated separation. This study aimed to benchmark existing CV models that have been developed for automating C&D waste sorting, categorizing them into three groups: CNN-based, attention-based, and transformer-based. We evaluated all models using a unified standard and compare them in terms of accuracy, efficiency, and robustness. The contributions of this research are:

1) A benchmarking pipeline designed for performance evaluation. The research methodology for deriving benchmarks consists of dataset preparation, model selection, training, and evaluation. Each step was carefully designed to establish a universal and standard pipeline for automated CV-based C&D waste sorting. All hyperparameters are well controlled for fair comparison. Based on the benchmarking pipeline, seventeen existing common CV models were evaluated and compared across three different measurement dimensions, i.e., accuracy, efficiency, and robustness.
2) Upon analysing the evaluation results, this research offers practical suggestions based on different real-world usage requirements and outlines the challenges faced by current CV methods for C&D waste sorting, proposing future directions to address each challenge. The study holds both theoretical significance and practical value.

Theoretically, it allows researchers to identify the strengths and weaknesses of various CV approaches, laying the groundwork for refining existing models and developing more innovative, robust solutions for waste sorting. From an engineering perspective, it assists practitioners and system developers in selecting the most suitable CV models for real-world applications, improving automated sorting performance and enhancing waste management outcomes. Ultimately, this research fosters more efficient recycling and reduces the environmental impact of construction and demolition waste.

This pipeline not only ensures that benchmarks are derived in a systematic and rigorous manner, but also facilitates a fair comparison and understanding of the performance of various models and techniques within the C&D waste sorting domain.
The benchmarking results reveal that BEiT achieved the highest accuracy performance, with a MIoU of 58.31%. FastFCN exhibited the shortest computing time, clocking in at 12.87ms during efficiency evaluation. SAN demonstrated the smallest standard deviation of 4.41% in terms of robustness evaluation. PSANet emerges as the recommended model due to its excellent balance across the three aspects, while Grounded SAM offers a straightforward "load and go" approach to alleviate the annotation workload.

By benchmarking these models against a unified standard, this study identified three key challenges: low data quality, data imbalance, and the time consumption of transformer-based models. For each challenge, it proposed future directions to address low-quality data, mitigate data imbalance, and manage the computational demands of transformer-based models. This study facilitates informed decision-making by researchers and practitioners in the selection and development of CV models. Furthermore, it instills confidence in the adoption of CV-based automated C&D waste sorting for normal waste management workflows in the construction industry.

References

Amari, S.-i. (1993). Backpropagation and stochastic gradient descent method. *Neurocomputing*, *5*(4-5), 185-196.

Bao, H., Dong, L., Piao, S., & Wei, F. (2021). Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*.

Chen, J., Lu, W., & Xue, F. (2021). "Looking beneath the surface": A visual-physical feature hybrid approach for unattended gauging of construction waste composition. *Journal of environmental management*, *286*, 112233.

Chen, J., Lu, W., Yuan, L., Wu, Y., & Xue, F. (2022). Estimating construction waste truck payload volume using monocular vision. *Resources, conservation and recycling*, *177*, 106013.

Chen, L.-C., Papandreou, G., Schroff, F., & Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018, September 8-14). Encoder-decoder with atrous separable convolution for semantic image segmentation. Proceedings of the European conference on computer vision (ECCV), Munich, Germany.

Csurka, G., Volpi, R., & Chidlovskii, B. (2022). Semantic image segmentation: Two decades of research. *Foundations and Trends® in Computer Graphics and Vision*, *14*(1-2), 1-162.

Davis, P., Aziz, F., Newaz, M. T., Sher, W., & Simon, L. (2021). The classification of construction waste material using a deep convolutional neural network. *Automation in construction*, *122*, 103481.

Demetriou, D., Mavromatidis, P., Robert, P. M., Papadopoulos, H., Petrou, M. F., & Nicolaides, D. (2023). Real-time construction demolition waste detection using state-of-the-art deep learning methods; single–stage vs two-stage detectors. *Waste Management*, *167*, 194-203.

Dong, Z., Chen, J., & Lu, W. (2022). Computer vision to recognize construction waste compositions: A novel boundary-aware transformer (BAT) model. *Journal of Environmental Management*, *305*, 114405.

Dong, Z., Wang, J., Cui, B., Wang, D., & Wang, X. (2020). Patch-based weakly supervised semantic segmentation network for crack detection. *Construction and Building Materials*, *258*, 120291.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., & Gelly, S. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

European Commission, E. C. (2023). *Database-Waste-Eurostat-European Commission*. Retrieved 1 December 2023 from https://ec.europa.eu/eurostat/web/waste/data/database

Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., & Lu, H. (2019). Dual attention network for scene segmentation. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, Long Beach, CA, United States.

Gov.AU. (2021). *National Waste Report 2020*. Retrieved 1 December 2023 from https://shorturl.at/orEKS

He, J., Deng, Z., Zhou, L., Wang, Y., & Qiao, Y. (2019). Adaptive pyramid context network for semantic segmentation. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, Long Beach, CA, United States.

HKEPD. (2011). *Construction Waste Statistics*. Retrieved 2 from https://www-epd-gov-hk.eproxy.lib.hku.hk/epd/misc/cdm/trip.htm

Hoornweg, D., & Bhada-Tata, P. (2012). What a waste: a global review of solid waste management.

Huang, L., Yuan, Y., Guo, J., Zhang, C., Chen, X., & Wang, J. (2019). Interlaced sparse self-attention for semantic segmentation. arXiv 2019. *arXiv preprint arXiv:1907.12273*.

Jiang, Y., Huang, Y., Liu, J., Li, D., Li, S., Nie, W., & Chung, I.-H. (2022). Automatic volume calculation and mapping of construction and demolition debris using drones, deep learning, and GIS. *Drones*, *6*(10), 279.

Khushi, M., Shaukat, K., Alam, T. M., Hameed, I. A., Uddin, S., Luo, S., Yang, X., & Reyes, M. C. (2021). A comparative performance analysis of data resampling methods on imbalance medical data. *IEEE access*, *9*, 109960-109975.

Liu, J., Li, Y., & Wang, Z. (2023). The potential for carbon reduction in construction waste sorting: a dynamic simulation. *Energy*, *275*, 127477.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition, Boston, Massachusetts, USA.

Lu, W. (2019). Big data analytics to identify illegal construction waste dumping: A Hong Kong study. *Resources, conservation and recycling*, *141*, 264-272.

Lu, W., & Chen, J. (2022). Computer vision for solid waste sorting: A critical review of academic research. *Waste Management*, *142*, 29-43. https://doi.org/10.1016/j.wasman.2022.02.009

Lu, W., Chen, J., & Xue, F. (2022). Using computer vision to recognize composition of construction waste mixtures: A semantic segmentation approach. *Resources, conservation and recycling*, *178*, 106022.

Lu, W., Yuan, L., & Lee, W. M. (2022). Understanding loading patterns of construction waste hauling trucks: triangulation between big quantitative and informative qualitative data. *Environmental Science and Pollution Research*, *29*(33), 50867-50880.

Lu, W., Yuan, L., & Xue, F. (2021). Investigating the bulk density of construction waste: A big data-driven approach. *Resources, conservation and recycling*, *169*, 105480.

Mishra, P., & Sarawadekar, K. (2019). Polynomial learning rate policy with warm restart for deep neural network. TENCON 2019-2019 IEEE Region 10 Conference (TENCON), Kochi, India.

Na, S., Heo, S., Han, S., Shin, Y., & Lee, M. (2022). Development of an artificial intelligence model to recognise construction waste by applying image data augmentation and transfer learning. *Buildings*, *12*(2), 175.

Nežerka, V., Zbíral, T., & Trejbal, J. (2024). Machine-learning-assisted classification of construction and demolition waste fragments using computer vision: Convolution versus extraction of selected features. *Expert Systems with Applications*, *238*, 121568.

Ni, C., Wu, J., Wang, H., Lu, W., & Zhang, C. (2024). Enhancing Cloud-Based Large Language Model Processing with Elasticsearch and Transformer Models. *arXiv preprint arXiv:2403.00807*.

OpenMMLab. (2020). *MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark*. Retrieved 13rd May from https://github.com/open-mmlab/mmsegmentation

Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., & Yan, F. (2024). Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*.

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., & Madry, A. (2018). Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*.

USEPA. (2023). *Sustainable Management of Construction and Demolition Materials*. Retrieved 1 December 2023 from https://shorturl.at/hitV9

Wang, X., Girshick, R., Gupta, A., & He, K. (2018). Non-local neural networks. Proceedings of the IEEE conference on computer vision and pattern recognition, Salt Lake City, UT, USA.

Wu, H., Zhang, J., Huang, K., Liang, K., & Yu, Y. (2019). Fastfcn: Rethinking dilated convolution in the backbone for semantic segmentation. *arXiv preprint arXiv:1903.11816*.

Xiao, T., Liu, Y., Zhou, B., Jiang, Y., & Sun, J. (2018). Unified perceptual parsing for scene understanding. Proceedings of the European conference on computer vision (ECCV), Munich, Germany.

Xu, M., Zhang, Z., Wei, F., Hu, H., & Bai, X. (2023). Side adapter network for open-vocabulary semantic segmentation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada.

Yong, Q., Wu, H., Wang, J., Chen, R., Yu, B., Zuo, J., & Du, L. (2023). Automatic identification of illegal construction and demolition waste landfills: A computer vision approach. *Waste Management*, *172*, 267-277.

Yuan, L., Lu, W., & Xue, F. (2021). Estimation of construction waste composition based on bulk density: A big data-probability (BD-P) model. *Journal of environmental management*, *292*, 112822.

Yuan, L., Yang, B., Lu, W., & Peng, Z. (2024). Carbon footprint accounting across the construction waste lifecycle: A critical review of research. *Environmental Impact Assessment Review*, *107*, 107551.

Yuan, Y., Chen, X., & Wang, J. (2020, August 23–28, 2020). Object-contextual representations for semantic segmentation. Computer Vision–ECCV 2020: 16th European Conference, Proceedings, Part VI 16, Glasgow, UK.

Zhang, H., Dana, K., Shi, J., Zhang, Z., Wang, X., Tyagi, A., & Agrawal, A. (2018). Context encoding for semantic segmentation. Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA.

Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid scene parsing network. Proceedings of the IEEE conference on computer vision and pattern recognition, Honolulu, HI, USA.

Zhao, H., Zhang, Y., Liu, S., Shi, J., Loy, C. C., Lin, D., & Jia, J. (2018). Psanet: Point-wise spatial attention network for scene parsing. Proceedings of the European conference on computer vision (ECCV), Munich, Germany.

Zhao, Q., Gao, W., Su, Y., Wang, T., & Wang, J. (2023). How can C&D waste recycling do a carbon emission contribution for construction industry in Japan city? *Energy and Buildings*, *298*, 113538. https://doi.org/https://doi.org/10.1016/j.enbuild.2023.113538

Zhou, Q., Liu, H., Qiu, Y., & Zheng, W. (2022). Object detection for construction waste based on an improved YOLOv5 model. *Sustainability*, *15*(1), 681. https://doi.org/https://doi.org/10.3390/su15010681