Computer vision to recognize construction waste compositions: A novel boundary-aware Transformer (BAT) model

Zhiming Dong [a], Junjie Chen [a, *], Weisheng Lu [a]

[a] Department of Real Estate and Construction, The University of Hong Kong, Hong Kong, China

Abstract

Accurate recognition of construction waste (CW) compositions using computer vision (CV) is increasingly explored to enable its subsequent management, e.g., determining chargeable levy at disposal facilities, or waste segregation using robot arms. However, applicability of existing CV approaches for the recognition of CW mixtures is limited by their relatively low accuracy, characterized by a failure to distinguish boundaries among different waste materials. This paper aims to propose a novel boundary-aware Transformer (BAT) model for fine-grained composition recognition of CW mixtures. First, a preprocessing workflow is devised to separate the hard-to-recognize edges from the background. Second, a Transformer structure with a self-designed cascade decoder is developed to segment different waste materials from CW mixtures. Finally, a learning-enabled edge refinement scheme is used to finetune the ignored boundaries, further boosting the segmentation precision. Performance of the BAT model was evaluated on a benchmark dataset comprising nine types of materials in a cluttered and mixture state. It recorded a 5.48% improvement of MIoU (mean intersection over union) and 3.65% of MAcc (Mean Accuracy) against the baseline. The research contributes to the body of interdisciplinary knowledge by presenting a novel deep learning model for semantic segmentation in recognizing construction waste compositions. It can also expedite the applications of CV in construction waste management to achieve a circular economy.

Keywords: Construction and demolition waste; Waste composition; Artificial intelligence; Transformer; Material recognition; Semantic segmentation.

1. Introduction

[*] Corresponding author.

E-mail address: chenjj10@hku.hk (J. Chen).

Construction waste (CW), or construction and demolition (C&D) waste, accounts for a significant proportion in the total waste stream. As stated in a World Bank report (Hoornweg and Bhada-Tata, 2012), CW such as rubble, concrete and masonry is a major component that can represent as much as 40% of the total solid waste generated in some cities. In Hong Kong, while the construction sector contributes less than 5% of the annual gross domestic product (GDP) (Leung and Wong, 2004), CW it generated takes up one quarter of the waste that ends up in landfill (HKEPD, 2020). Faced with the mountainous CW, the importance of construction waste management (CWM) can never be overstated. Effective CWM relies on yardstick information of CW composition. For example, it is a common practice in countries such as the United Kingdom (Avery Weigh-Tronix, 2010) and Australia (NSWEPA, 2018) to levy different disposal fees according to the composition of CW dumps (Yuan et al., 2021a). In addition, when CW is segregated in recovery facilities, information on waste material types and composition is essential for sorting operation enabled by robots.

The use of computer vision (CV) in waste recognition is promising, as photographs are easy and cheap to collect, and suitable for the analyses of a great variety of waste materials. Relevant research has been ongoing for more than two decades (Faibish et al., 1997), trying to recognize waste materials from images and enabling various waste management applications, such as household waste classification (Srinilta and Kanharattanachai, 2019; Yang et al., 2021), bin level detection (Aziz et al., 2018; Hannan et al., 2016) and material segregation (Ku et al., 2020; Lukka et al., 2014). In early years, most of the research attentions were paid to the recognition of municipal solid waste (MSW) (Sauve and Van Acker, 2020). Recently, stimulated by the economic benefits and technological development, growing studies have been devoted to the applications of CV in the CWM. While some of these studies aimed to facilitate source separation on construction site (Lau Hiu Hoong et al., 2020; Wang et al., 2019a), others provided technologies to enable the processing of construction materials at centralized disposal facilities (Chen et al., 2021; Kujala et al., 2015; Lukka et al., 2014).

Despite the progress has been made, existing methods may encounter difficulties in transferring from laboratory environments to practical industrial practice, primarily for insufficient precision and granularity of their recognition results. First, existing research tends to focus on the task of image classification, which can only identify whether the waste item in a given image belongs to one of the several predetermined categories or not. Recognition with such low granularity might be suitable for assisting household residual classification, but is not sufficient for determining composition of CW mixtures, which usually comprise multiple types of materials in a highly cluttered state. Second, although there have been studies (Wang et al., 2019a) trying to distinguish and locate multiple instances of different waste materials by the use of object detection techniques, they are not oriented to practical engineering applications. Rather, such studies usually simplify the problem as one to recognize individual waste items appearing against a simple, unified background, ignoring the complexity of real-life context and the heterogeneous nature of CW (Awe et al., 2017; Nowakowski and Pamuła, 2020).

Realizing the above limitations, Lu et al. (2021) proposed an approach to recognizing CW composition in its original cluttered state by using semantic segmentation, a technique that can deliver fine-grained information such as types of waste materials, and their corresponding pixel areas in images. The research set a benchmark for subsequent studies, with a mean intersection over union (MIoU) of 0.56 in distinguishing 9 types of CW. However, there is room for further improvement. One notable limitation of the previous approach is its failure to precisely depict waste materials' boundaries, resulting in a relatively low MIoU. The deficiency in boundary detection can potentially be addressed by recent advancements in the CV community and the incorporation of boundary-aware processing techniques. For example, Transformer, a deep learning model primarily used for natural language processing (NLP), has been applied to undertake CV tasks, demonstrating superior performance than traditional convolutional neural network (CNN)-based structure (Dosovitskiy et al., 2020).

This paper aims to propose a boundary-aware semantic segmentation model based on the Transformer architecture for the robust CW composition recognition in fine granularity. We called the newly proposed framework "boundary aware Transformer (BAT)". It contributes to the problem of computer vision-enabled CW composition recognition, which allows the robust and fine-grained recognition of waste materials from cluttered CW mixtures. The novelty of the model lies in the integration of a preprocessing module that separately handles the micro inter-material edges, a Transformer-based waste segmentation structure with cascade decoding, and a model-agnostic boundary refinement scheme enabled by SegFix. This paper is organized as follows. Subsequent to this introductory section, Section 2 describes the status quo of CV in waste recognition. Section 3 illustrates the proposed boundary-aware model for CW composition recognition, and Section 4 delivers its implementation results. Section 5 concludes the paper with the main findings and potential future works.

2. Literature review

According to the differences of the used CV techniques, existing research on waste recognition can be divided into two streams. One is based on image classification, and the other is based on object detection or semantic segmentation. In this chapter, we first review the two streams of works in CV-enabled waste recognition in Sections 2.1 and 2.2, respectively, and then an introduction of attention mechanisms and Transformers is delineated in Section 2.3.

2.1. Waste recognition based on image classification

Waste recognition based on image classification aims to classify a given waste image into one of the predetermined categories. Previous research attention has been primarily paid to the classification of MSW, e.g., paper, plastic, organic, and metal. Traditionally, features of waste materials first need to be hand-engineered, and then input to machine learning models such as support vector machine (SVM) (Özkan et al., 2015; Paulraj et al., 2016; Wang et al., 2019b) and neural networks (Faibish et al., 1997; Ramli et al., 2010) for classifier training. Applicability of these traditional approaches is limited due to the extensive manual efforts for features handcrafting and relatively low robustness.

With the resurgence of deep learning (DL), CNN has become the predominant model in waste recognition. Based on a public dataset comprising six common waste types provided by (Yang and Thung, 2016), a series of research (Bircanoğlu et al., 2018; Huang et al., 2020; Mao et al., 2021; Meng and Chu, 2020; Zhang et al., 2021) has been carried out to recognize single waste objects appearing against a relatively simple background. Zhang et al. (2021) integrated a self-monitoring module into ResNet18 for recyclable waste classification, which can recognize the six waste types in TrashNet with an accuracy of 95.87%. Mao et al. (2021) employed a genetic optimization algorithm to finetune the hyperparameters of DenseNet, and achieved a 99.60% classification accuracy on TrashNet.

Compared with MSW recognition, only a limited number of works have focused on using image classification techniques for CW recognition (Brisola et al., 2010; Chen et al., 2021; Lau Hiu Hoong et al., 2020; Xiao et al., 2020). Xiao et al. (2020) integrated handcrafted features such as colors and gray level co-occurrence matrix and CNN-extracted features with the extreme learning machine (ELM) for the classification of five typical CW categories, i.e., wood, brick, rubber, rock, and concrete. Lau Hiu Hoong et al. (2020) proposed a method based on CNN which can determine composition of recycled aggregates in near real time. Chen et al. (2021) proposed a hybrid approach to integrating visual features extracted by a DenseNet-169 and physical features such as weight and waste depth for unattended gauging of inert content (e.g., rock, gravel, earth and sand) proportion in CW mixtures.

Despite the high performance attained by the aforementioned research, image classification can only reveal if an image contains a certain material category, but fails to provide information of finer granularity regarding the location, geometry and boundaries of waste materials. Such fine-grained information is essential to enable various applications in industrial practice, e.g., composition measuring and waste segregation with robotics. This is especially the case when multiple targets appear simultaneously in real-life context, which is the common settings in practice.

2.2. Waste recognition based on object detection/semantic segmentation

In recent years, more and more researchers have realized the limitations of image classification and turned to investigate the applications of object detection or semantic segmentation in the waste management industry (Anjum and Umar, 2018; Liang and Gu, 2021; Panwar et al., 2020; Wang et al., 2019a). In the field of computer vision, object detection is a task that aims to locate objects of different types in images with bounding boxes, while semantic segmentation goes further in granularity by distinguishing pixel areas corresponding to different semantic classes (Bhola et al., 2018; Mansouri, 2019). Previous research has investigated the applicability of various CNN architectures such as R-CNN (Ku et al., 2020), Faster R-CNN (Awe et al., 2017; Nowakowski and Pamuła, 2020), and Mask R-CNN (Panwar et al., 2020; Proença and Simões, 2020) in detecting or segmenting MSW in contexts. Liang and Gu (2021) proposed a multi-task learning architecture based on CNN to simultaneously classify and locate household and residential wastes. To enable such research, corresponding datasets with multiple waste items in real-life background were collected or even made publicly available (Liang and Gu, 2021; Proença and Simões, 2020).

Similar research efforts have been made in construction waste management. Lukka et al. (2014) and Kujala et al. (2015) incorporated computer vision as a core module of a robotic system called ZenRobotics Recycler, which can detect, locate, and classify construction wastes on conveyor belts for automatic segregation. Ku et al. (2020) devised a grasp detection approach based on R-CNN for the processing of construction and demolition wastes. In (Wang et al., 2019a), CW detection models were trained based on the Faster R-CNN and Mask R-CNN architecture, which can enable robots to recycle nails, screws, and residual pipes and cables on construction site. It is observed that most of previous research mainly focused on detecting separate CW objects in a relatively well-control condition. While such research is helpful for waste segregation in semi-structured environments such as recovery facilities, it fails to work in scenarios where heterogenous materials are randomly mixed up, e.g., truck-loaded CW.

To address the issue, Lu et al. (2021) proposed an approach based on DeepLabv3+ to recognizing composited material components from cluttered CW mixtures, which demonstrated the feasibility of semantic segmentation in distinguishing highly unstructured materials in mixtures states. However, its precision is still not sufficiently high for practical applications in CWM, primarily because the deficiency in boundary detection. To enable fine-grained composition recognition for CW mixtures, a boundary-aware semantic segmentation model is required that can depict edges among different waste materials. Such boundary-aware precise waste segmentation can potentially be achieved by Transformer, a DL framework that is gaining momentum in the field of computer vision.

2.3. Attention mechanism and Transformers

Transformer is proposed first for NLP (Vaswani et al., 2017). It is a deep learning model different from CNN and recurrent neural network (RNN), and has achieved remarkable performance in a number of NLP tasks such as machine translation (Takase and Kiyono, 2021) and language modelling (Brown et al., 2020). A transformer encoder is mainly consisted by self-attention layers for feature extraction, and Feed Forward Neural Networks (FFN) for spatial transformation.

The Self-attention layer serves as the primary feature extractor, which creates three tensors: query tensor (Q), key tensor (K) and value tensor (V) to consider the internal correlation of the input tensor, and calculate the embedded features. The attention mechanism can be represented as Eq. (1):

$$\mathrm{Attention}(Q, K, V) = \mathrm{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{1}$$

The Q should dot with K firstly, which indicates the score of correlation between each element. Division and Softmax normalization operation are used to keep the gradient stable (d is the dimension of Q and K). The softmaxed tensor is finally multiplied with V to calculate the weighted output.

In recent years, Transformer is widely used in many computer vision tasks. ViT clips images into flatten patches sequence, which is used as the input of Transformer model (Dosovitskiy et al., 2020). DETR is a Transformer-based end-to-end object detection network, which has advantages of anchor-free and NMS-free. The method significantly outperforms competitive baselines (Carion et al., 2020). Image GPT directly reshapes two-dimensional images into one-dimensional as model input, which are used for training an image generation model in unsupervised way, thus Transformer is used in pixel prediction task (Chen et al., 2020); SegFormer combined a hierarchical Transformer encoder and a lightweight decoder, and has achieved a considerable performance in image segmentation task (Xie et al., 2021). However, little research, if any, has applied advanced Transformer models in CW-related visual recognition tasks.

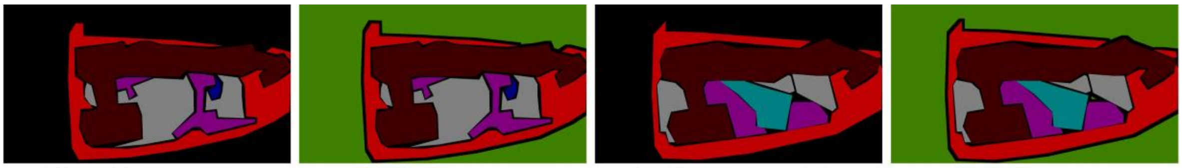
In our work, a Transformer-based image segmentation framework is proposed to tackle the challenging CW composition recognition task. The proposed framework uses a typical encoder-decoder structure. The encoder uses the self-attention mechanism, where the query, key and value tensors are generated with the same embedding. The decoder, on the other hand, uses the cross-attention mechanism, where the query tensor, and key and value tensors are generated by different embeddings.

3. The proposed boundary-aware transformer model

This research proposed a boundary-aware Transformer framework for fine-grained recognition of construction waste based on semantic segmentation. The framework includes three mutually interconnected steps: First, a dataset of mixed construction wastes is preprocessed to clarify waste boundary pixels from the background; Second, a Transformer-based model, which comprises a self-attention encoder module and a cascade decoder, is trained on the dataset for CW semantic segmentation; Finally, the segmentation results provided by the Transformer model are improved by a deep learning-based boundary refinement scheme.

3.1 Preprocessing the waste annotations

This study is based on a dataset collected and prepared by Lu et al. (2021), which includes 5,366 photos of highly cluttered CW mixtures. The dataset comprises seven types of CW (i.e., rock, gravel, earth, packaging, wood, other non-inert, and mixed) and two types of relevant objects (i.e., grip and truck). Annotating such a large CW dataset is challenging as different waste materials are usually intertwined with each other, and the boundaries wherein can be vague. As a result, the annotators tend to leave the ambiguous boundaries between different waste categories as an unlabeled background, which is imprecise and can undermine the performance of the segmentation model. To overcome the adverse impact of mislabeled boundary, a morphology-based preprocessing method is used to distinguish pixels of "background" from the "ignore" category. Erosion operation is implemented to process the background category, which can remove pixels at the edge of waste objects. After processing, the pixels between different categories are removed from the background. As the ground-truth labels of those pixels are unknown, they are treated as the "ignore" category in the training process. This means that during the training process, the predicted probability distribution of those pixels has no influence on loss calculation and gradient backward broadcast.



(a) #1 original annotation (b) #1 after process (c) #2 original annotation (d) #2 after process

Fig. 1. Schematic diagram showing how original annotations are preprocessed.

The preprocessing workflow is illustrated in Fig. 1, where (a) and (c) are the original ground truth while (b) and (d) are the corresponding processed labels. In Fig. 1 (a) and (c), black pixels refer to the "background" category, but there are also some pixels in object boundary are mislabeled as background. In (b) and (d), the morphology operation is used to distinguish boundary pixels from the background, where the green pixels represent the background, and black pixels represent the "ignore" category.

3.2 Transformer-based semantic segmentation

In this research, a Transformer-based semantic segmentation framework is proposed to explore the potential of Transformer in construction waste composition recognition. The mix transformer encoder (MiT) in Segformer (Xie et al., 2021) is integrated to apply the global attention mechanism in the proposed framework. A decoder based on multilayer perceptron (MLP) and object-contextual representation (OCR) is proposed and integrated into the proposed Transformer-based semantic segmentation framework. A cross-attention module is used in the decoder to predict the semantic category of each pixel more precisely based on the enhanced feature representation.

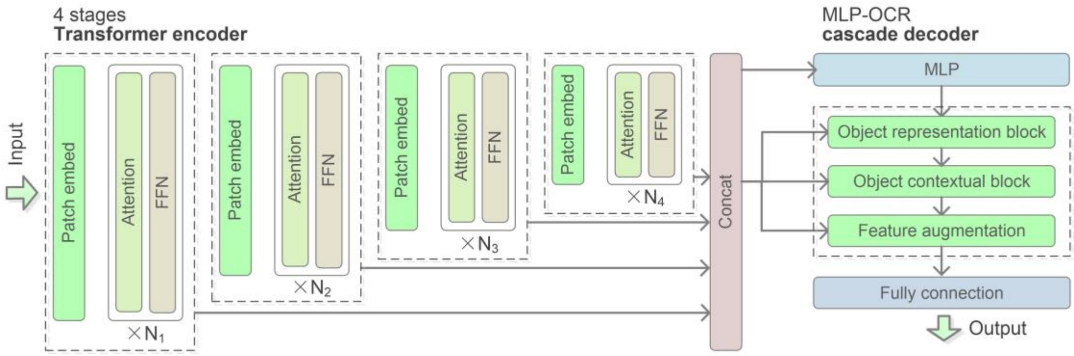


Fig. 2. The proposed Transformer-based framework for construction waste semantic segmentation. FFN and MLP stand for feed-forward network and multilayer perceptron, respectively.

Fig. 2 shows the architecture of the proposed Transformer-based semantic segmentation framework. The encoder contains 4 stages. In each stage, feature tensors are first embedded to token, then they are sent as the input of the Transformer encoder, which includes N_1, N_2, N_3 and N_4 stacked encoder block respectively. Each encoder block has a self-attention module, followed by a feed-forward network (FFN). The decoder has a cascade structure, where the embedded features output by each stage of the encoder are first upsampled and concatenated together in the Concat layer (the pink rectangle in Fig. 2), then processed by MLP layer (identified by blue rectangle in Fig. 2), and finally handled by OCR module (identified by green rectangles in decoder part of Fig. 2) to better consider representation of corresponding object class.

3.2.1 Hierarchical Transformer encoder

The input image tensor of size $(\mathrm{B} \times \mathrm{H} \times \mathrm{W} \times \mathrm{C})$ should be embedded to vector sequence with size $(\mathrm{B} \times \mathrm{N} \times \mathrm{C}_{embed})$, then it can be used as theinput of the Transformer block in each stage. B is the batch size, and H and W represent the height and width of the image, respectively. Similar to Transformer structures used in NLP, N can be seen as the length of a sequence, and C_{embed} is the dimension of embedding. While most of existing vision Transformer models (Dosovitskiy et al., 2020) crop and reshape the input image tensor to a sequence of flattened token embedding to handle 2D images in Transformer, the proposed Transformer framework uses a different approach introduced by MiT (Xie et al., 2021). To be more specific, an overlapped embedding scheme is used to consider the continuity of adjacent patches better. 2D convolution is used to project the overlapped patches to embedding, and then the embedded features are flattened and normalized to generate the embedded token.

We used an efficient self-attention module in MiT as the main feature extractor instead of CNN. There is generally an overall self-attention map (Fu et al., 2019) with size $(\mathrm{B} \times \mathrm{N} \times \mathrm{N})$ in the self-attention module, where N is the sequence length and $\mathrm{N} = \mathrm{H} \times \mathrm{W}$. The calculation process of overall self-attention map is compute-intensive and requires large storage resources, easily becoming a network bottleneck. Therefore, reduction ratio R is introduced to reduce the size of overall self-attention map to $(\mathrm{B} \times \mathrm{N}\,\mathrm{N}/R^2)$.

$$\mathrm{Attention}(\mathrm{Q},\mathrm{K},\mathrm{V}) = \mathrm{Softmax}\left(\frac{QK^T}{\sqrt{C}}\right)V \tag{2}$$

In Eq. (2), Q, K and V refer to query tensor, key tensor and value tensor respectively, they are calculated from input feature map. The shape of Q is $(\mathrm{B} \times \mathrm{N} \times \mathrm{C})$, and the shape of K, V is $(\mathrm{B} \times \mathrm{N}/R^2 \times \mathrm{C})$, where C is the channel length of input tensor.

Each self-attention module is connected to a feed-forward network (FFN), which consists of a convolutional layer, two fully connected layers, and an activation function. The FFN module can introduce more non-linear spatial transformations into the model, thereby enhancing the model's performance. FFN is widely used in various Transformer-based models (Vaswani et al., 2017).

3.2.2 MLP-OCR Cascade decoder

A lightweight multilayer perceptron(MLP) used in Segformer (Xie et al., 2021) is selected as the first stage decoder in the proposed method. Eq. (3) to Eq. (6) illustrates the calculation process of MLP. $Linear$ is the fully connection layer (FC), we can see that MLP is implemented by $\mathrm{FC} \rightarrow \mathrm{Upsample} \rightarrow \mathrm{FC} \rightarrow \mathrm{FC}$, Where F_i refers to embedded feature from i-th stage, the channel size is transformed from C_i to C.

$$\hat{F}_i = Linear(C_i, C)(F_i), \forall i \tag{3}$$

$$\hat{F}_i = Upsample\left(\frac{H}{4} \times \frac{W}{4}\right)(\hat{F}_i), \forall i \tag{4}$$

$$\mathrm{F} = \mathrm{Linear}(4\mathrm{C},\mathrm{C})\left(Concat(\hat{F}_i)\right), \forall i \tag{5}$$

$$\mathrm{M} = \mathrm{Linear}(C, N_{cls})(F) \tag{6}$$

An Object-Contextual Representation module (Yuan et al., 2021b) is used in the decoder to enhance its ability to predict the semantic category and feature representation of each pixel. The OCR comprises three parts: object representation block, object contextual block and feature augmentation.

The object representation block multiplies pixel representation (extracted from backbone network) and categories probability map to obtain a context matrix that characterizes the similarity between object features and each category. The formula is shown as Eq. (7):

$$f_k = \sum_{i \in \mathcal{L}} \widetilde{m}_{ki} x_i \tag{7}$$

Where $\mathcal{L}$ refers to pixels in an image, x_i represents the feature of i-th pixel, $\widetilde{m}_{ki}$ refers to the probability of i-th pixel belong to k-th category.

The object contextual block utilizes a cross-attention module similar to (Wang et al., 2018), which calculates a relation matrix between pixel representation and context matrix generated from object region block, then weight the original pixel representation, the formula is shown as Eq. (2), which is similar to the self-attention module used in MiT, but the calculation of this cross-attention is different: while the query, key and value tensors in MiT are generated with the same embedding (thus is called self-attention), the query tensor in OCR is generated from the context matrix, and the key and value sensors are generated from image features (thus is called cross-attention).

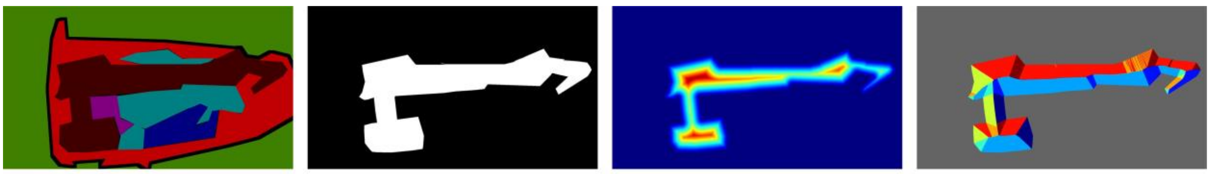
The last step concatenates the outputs of object contextual block and original embedding to get the augmented representation:

$$z_i = g([x_i^T \ y_i^T]^T) \tag{8}$$

In Eq. (8), $g(\cdot)$ refers to a non-linear transform, x_i and y_i refer to embedding generated from encoder and object contextual block. Splicing OCR with the feature representation of the deepest input of the network as the context information enhanced feature representation which is called feature argumentation in OCR, the semantic category of each pixel can be predicted based on the enhanced feature representation more precisely.

3.3. Boundary refinement

SegFix is used to refine the prediction results, focusing particularly on edge pixels at waste boundaries (Yuan et al., 2020). SegFix is a deep learning-based image segmentation post-processing scheme compatible with different models for segmentation refinement. SegFix uses a fine-designed object direction map as ground truth for model training to obtain an offset map. HRNet (Sun et al., 2019) is used in the proposed method as the backbone of the SegFix. Two branches are designed to learn the offset from the boundary, i.e., a boundary branch and a direction branch. The boundary branch learns a probability map $B_{boundry}$ with size $(H \times W \times 1)$, where H and W are the height and width of an image, respectively, and each element in $B_{boundry}$ refers to the probability of a pixel belong to designated boundary The direction branch learns a direction map $B_{direction}$ with size $(H \times W \times 2)$, of which an element b_{ij} represents the direction of the pixel p_{ij} away from the edge. The value of b_{ij} is discretized, and equals on the following: $(1,0)$, $(-1,0)$, $(0,1)$, $(1,1)$, $(-1,1)$, $(1,-1)$, $(-1,-1)$ and $(0,-1)$.



(a) ground truth (b) binary map (c) distance map (d) direction map

Fig. 3. The general procedure on how Segifx predicts edge and direction maps.

Fig. 3 illustrates the label generation procedure of SegFix. Distance map (c) and direction map

(d) are used as the supervision of boundary branch and direction branch accordingly. The binary map (b) of a single category are extracted firstly, then a distance transform implemented by SciPy (Virtanen et al., 2020) is used to calculate the distance map (c), lastly, Sobel filter (Sobel, 2014) is used to calculate the direction map (d).

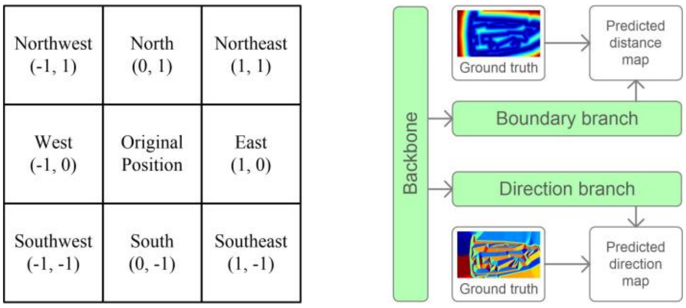


Fig. 4. (a) Corresponding relationship between direction and offset values; (b) Structure of the SegFix framework.

Fig. 4 (a) Illustrates the corresponding relationship between direction and offset value, where eight directions are encoded into a vector for the convenience of proceeding. Fig. 4 (b) is the framework of SegFix, which first uses a segmentation backbone to get the embedded feature map of input image, then sends the embedded feature to two different branch to predict the distance map and direction map respectively, and finally process the two maps to generate an offset map for inference. The binary cross-entropy loss is used in boundary loss and direction loss. HRNet (Sun et al., 2019) is used as the backbone.

The predicted offset map is used to refine the segmentation map generated by the previous Transformer framework. For each element s_{ij} in segmentation map, offset it with stride d along direction b_{ij} predicted from SegFix to $s_{i'j'}$, and sample the new category in position $s_{i'j'}$ as the refined segmentation map. SegFix can use edge information to refine the segmentation map, thereby improving the proposed method's ability to process object edge.

4. Implementation and results

4.1 Dataset, implementation details and baseline

The dataset in this research is collected from waste disposal facilities in Hong Kong. There are 5,366 images in this dataset, each with a manually-annotated segmentation label. The dataset is randomly split to train set, validation set and test set according to the ratio of 7:1.5:1.5.

Experiments are conducted in a computing server with Ubuntu 18.04 system and NVIDIA A100-SXM4-40GB GPU, and a Python-based deep learning framework PyTorch is used in the implementation of deep learning network architecture. Several data augmentation schemes are used in image segmentation, including random crop and flip, and normalization. For the training of the Transformer-based segmentation model, AdamW (Loshchilov and Hutter, 2017) is used as optimizer, the cross-entropy loss is used as loss function, and the max iteration is set as 160,000. To train the SegFix, HRNet-18 (Sun et al., 2019) is used as the backbone, and binary cross-entropy loss is used as boundary loss and direction loss. The used training strategy is stochastic gradient descent (SGD), and the learning rate and max iteration are set as 0.004 and 80,000, respectively.

MIoU and MAcc is used as evaluation metrics. MIoU is a widely used evaluation metrics in semantic segmentation tasks, which is defined as the mean intersection over union (IoU) of all categories in the dataset:

$$\mathrm{MIoU} = \frac{1}{k}\sum_{i=1}^{k}\left(p_{ii}/\left(\sum_{j=1}^{k}p_{ij} + \sum_{j=1}^{k}p_{ji} - p_{ii}\right)\right) \tag{9}$$

Where, p_{ij} indicates the number of pixels for which the ground truth belongs to the i-th category, and for which the predicted value belongs to the j-th category. k is the total number of categories.

MAcc refers to the mean accuracy, which is the average of segmentation accuracy across all categories:

$$\mathrm{MAcc} = \frac{1}{k}\sum_{i=1}^{k}\left(p_{ii}/\sum_{j=1}^{k}p_{ij}\right) \tag{10}$$

Where, p_{ij} is the number of pixels that belong to i-th category in the ground truth, and also be predicted as j-th category, k is the number of category in the dataset.

A highly optimized DeepLab V3+ proposed in (Lu et al., 2021) is used as the baseline. There are nine categories and background in the dataset, and the IoU and Acc of each category are shown in Table 1. The MIoU and MAcc is 56.2% and 69.19% accordingly.

Table 1. Performance of baseline.

	background	rock	gravel	earth	packaging	wood	others	mixed	grip	truck	overall
MIoU	97.1%	38.2%	37.3%	37.5%	52%	66.2%	35%	38.6%	87.7%	72.9%	56.2%
MAcc	98%	48%	53%	51%	71%	84%	45%	60%	95%	87%	69.19%

4.2 Ablation experiments

A group of experiments is designed to analyze influences of the four different modules in the proposed framework, i.e., preprocessing, encoder, decoder and post-processing. Whether the respective module is applied or what options are used in the modules will have an influence on the final performance. In this section, such effects are comprehensively investigated by comparing the MIoU and MAcc metrics.

Table 2 illustrates the experiment result, where different methods are distinguished by different index. For preprocessing, method #1 to method #3 use the original dataset for training, and method #4 to method #6 use the preprocessed dataset. For the encoder, method #1 to method #3 use MiT-B0, MiT-B2 and MiT-B5 respectively to explore the influence of different encoders.

For decoder, method #1 to method #4 use MLP as decoder, whereas method #5 and method #6 use the proposed MLP-OCR as decoder. Method #6 has applied SegFix post-processing, while the others have not

Table 2. Results of ablation experiments.

Method	Preprocessing	Encoder	Decoder	Post-processing	MIoU	MAcc
#1		MiT-B0	MLP		53.36%	67.92%
#2		MiT-B2	MLP		56.02%	70.38%
#3		MiT-B5	MLP		56.9%	70.26%
#4	✓	MiT-B5	MLP		60.58%	72.04%
#5	✓	MiT-B5	MLP-OCR		61.45%	72.64%
#6	✓	MiT-B5	MLP-OCR	SegFix	61.68%	72.84%

4.2.1 Influence of preprocessing

A similar network structure is used in this section to compare the influence of preprocessing procedure. The network use MiT-B5 as encoder and MLP as decoder, method #3 and method #4 is trained on the original dataset and the preprocessed dataset accordingly, the hyperparameter is set as the same, and they can all converge under this set of parameters. Evaluation results are shown in Table 2, line 3 and line 4 shows the evaluation metrics, the MIoU is 56.9% in method #3, and it has an improvement of 3.68%, in method #4, which is 60.58%. The MAcc of method #3 and method #4 is 70.26% and 72.04% accordingly, it is shown an improvement of 1.78%. In this comparison experiment we can see that the preprocessing procedure can improve the performance of Transformer network.

4.2.2 Influence of different MiT variants (encoders)

The MiT encoder has several different variants: MiT-B0 to MiT-B5. They follows the same structure but uses different parameters such as the number of Transformer blocks in each stage. Among the variants, MiT-B0 is the most lightweight whereas MiT-B5 has the largest number of parameters. Therefore MiT-B5 tends to perform better in segmentation accuracy while MiT-B0 has greater inference speed. Table 3 illustrates the parameter used in the different variants of MiT:

Table 3. Model parameters of different MiT encoder variants.

MiT encoder	Stage #1	Stage #2	Stage #3	Stage #4	Num. of Params
MiT-B0	2/32	2/64	2/160	2/256	3.4M
MiT-B1	2/64	2/128	2/320	2/512	13.1M
MiT-B2	3/64	4/128	6/320	3/512	24.2M
MiT-B2	3/64	4/128	18/320	3/512	44.0M
MiT-B4	3/64	8/128	27/320	3/512	60.8M
MiT-B5	3/64	6/128	40/320	3/512	81.4M

[*] The fractions from column 2 to column 5 represent “stack number/length”

The size of MiT encoder is mainly influenced by two parameters: the stack numbers of Transformer in each stage and the vector length of embedded patch in each stage. The larger the parameters, the larger the size of the model and the more parameters. Table 3 shows details of the two parameters, and the number of parameters of each model. MiT-B0, MiT-B2 and MiT-B5 are selected for the comparison of performance. Table 2 lists the evaluation results of methods using different encoder. Method #1, #2, and #3 used MiT-B0 , MiT-B2, and MiT-B5 as encoders, respectively. Other parameters, including training hyper parameters and model configuration parameters, of the three methods are kept the same to allow direct performance comparison.

As shown in Table 2, MIoU and MAcc of the methods changed with the variation of the model size. The more parameters of the model, the better the performance. TheMIoU of the three methods are 53.36%, 56.02% and 56.9% respectively, and the MAcc are 67.92%, 70.38% and 70.26% respectively. The results indicate that models with a larger number of parameters tend to have the better performance. Therefore, the MiT-B5 encoder is selected as the encoder of the proposed TransFormer-based framework.

4.2.3 Influence of different decoders

Two different decoder structures, i.e., the MLP decoder and the proposed MLP-OCR decoder, are used in the ablation experiments respectively. In the experiment, method #4 uses MiT-B5 as its encoder and MLP as its decoder, whereas method #5 uses MiT-B5 and MLP-OCR as its encoder and decoder, respectively. In SegFormer, MLP is the default decoder, which has a lightweight structure to avoid the side influence of hand-crafted components. In our method, a MLP-OCR structure is proposed and used as the decoder in the TransFormer-based framework, so as to improve the feature representation ability.

The resulted performance is shown in row 4 and row 5 in Table 2. The MIoU of method #4 and method #5 are 60.58% and 61.45%, respectively; the MAcc, on the other hand, are, respectively, 72.04% and 72.64%. As the result shows, compared with the simple MLP decoder, MLP-OCR decoder can lead to higher segmentation precision.

4.2.4 Influence of SegFix

SegFix is used as a post-processing scheme to refine the predicted label. In this section, the effectiveness of SegFix is evaluated by comparing the SegFix refinement result (method #6) with the original results predicted by method #5. The evaluation metrics is shown in row 5 and row 6 in table 2. We can find that, after refinement, the MIoU and MAcc are improved by 0.23% and 0.20%, respectively. To visualize the refinement details, three patches are clipped from the test set and shown in Fig. 5. SegFix can learn an offset map from the original images, and there are eight different directions in the offset map, indicating how the predicted labels should be refined. In Fig. 5, the arrows represent the directions of offset, and offset distance is set to 2 pixels. For example, an arrow point to the right side in Fig. 5 means that using the current pixel as source position, shift 2 pixels' distance, and use the category in the new position to refine the category in the source position. Fig. 5 uses different colors of arrows to distinguish the actual effects exerted by SegFix: the yellow arrow indicates the corresponding position was originally assigned a wrong label but rectified by SegFix; the pink arrow, on the other hand, indicates the position has a correct label initially, but was changed to a wrong label by SegFix. And the blue one indicates those pixels that have not been changed. It is observed that SegFix can effectively refine the boundary pixels and improve the segmentation performance.

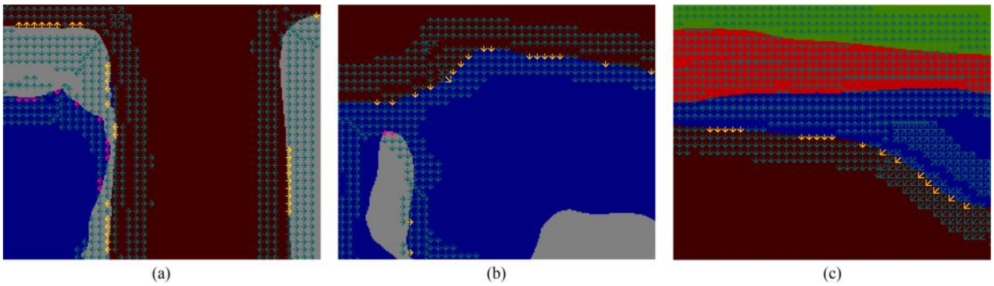


Fig. 5. Refinement by SegFix.

Fig. 6 show the difference of boundary detection ability of the proposed segmentation framework and the SegFix post processing method. Three samples are selected for illustration. In Fig. 7, (a), (b) and (c) are the ground truth of three selected samples, (d), (e) and (f) are the corresponding stacked predicted results of the proposed segmentation framework and the SegFix post processing method. In (d), (e) and (f), white pixels refer to the boundary prediction result of SegFix, while other colors refer to the original predicted categories without applying SegFix.

From Fig. 6, we can see that SegFix can better grasp the boundary information in images. This is because SegFix use direction map and distance map as supervision condition, which includes richer edge information compared with normal segmentation ground truth. The results demonstrate SegFix is effective in refining the prediction results generated by image segmentation model.

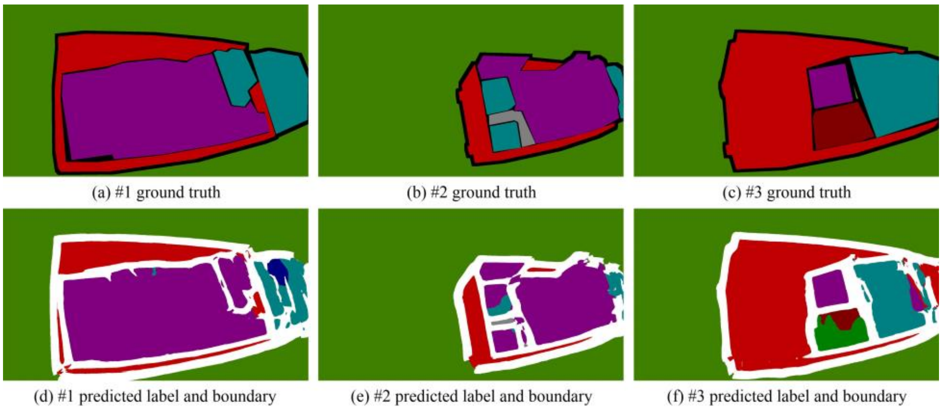


Fig. 6. Examples showing the effects of Segfix: The second row is the predicted results with SegFix applied, and the first row is the corresponding ground truth.

4.3 Performance comparison

Several classical CNN-based models were trained on the same dataset to compare their results with our the BAT framework. The trained CNN models include FCN (Long et al., 2015), DANet (Fu et al., 2019), DeepLab V3+ (Chen et al., 2018) and HRNet (Sun et al., 2019). FCN is a representative deep learning work applied in image segmentation. It is an end-to-end image segmentation method that allows the network to make pixel-level predictions. ResNet-50 is used as the backbone network of the FCN. DANet is a typical network which combined CNN architecture with attention module. It proposed two attention modules to further improve the feature representation of segmentation network. The DeepLab series have the advantages of fast and high performance, and thus are widely used in various datasets. In (Lu et al., 2021), a DeepLab V3+ model was trained and calibrated via orthogonal experiments for CW segmentation on the same dataset; thus, it will be considered as the baseline in this study. HRNet maintains high-resolution representations by connecting high-resolution to low-resolution convolutions in parallel, which has achieved state-of-the-art performance in several tasks. In the comparison, a variant HRNet-48 is used for comparison. Same training schedule is used in training process: SGD is used as optimizer; max training iteration is set to 80,000. MIoU and MAcc is used as evaluation metrics, the evaluation results are shown in Table 4.

Table 4. Performance of different semantic segmentation methods.

Method	MIoU	MAcc
FCN	45.72%	56.58%
DANet	49.66%	62.17%
DeepLabV3+ (baseline)	56.2%	69.19%
HRNet	52.05%	64.4%
Ours (BAT)	61.68%	72.84%

4.4 Discussion

4.4.1 Comparison with baseline

As listed in Table 4, the BAT surpassed the baseline obtained by DeepLab v3+ in terms of both MIoU and MAcc. The level of improvement reaches 9.8% and 5.3%, respectively. Some examples are selected to intuitively illustrate the improvement. Five examples are shown in Fig. 7, where column (a) to (d) are the original image, the ground truth, the segmentation result of the baseline method, and results provided by our BAT method. In (b), green refers to the background category, and black refers to the ignore category. Suppose a pixel belongs to ignore category in the ground truth. In that case, its predicted value will be not used to calculate loss and evaluation matric, thus it can be predicted as any other categories according to their embedded feature and contextual information, and the predicted result (c) and (d) will not include the ignore category. It is observed that while the baseline method performed poorly in distinguishing the object boundary, and the proposed BAT method has successfully recognized the minor details and edges among the waste materials. For example, in (1), the left area contains several categories, which are packed in a small area. The baseline method failed to effectively process this area, with many pixels at the boundary and corners misclassified as background. As a comparison, the proposed BAT method can distinguish them better. In addition, the proposed method has a more robust performance on recognizing the waste categories in images. In (2), the center area belongs to the "rock" category, which has been correctly identified by the proposed method, but mislabeled as the "earth" category by the baseline method.

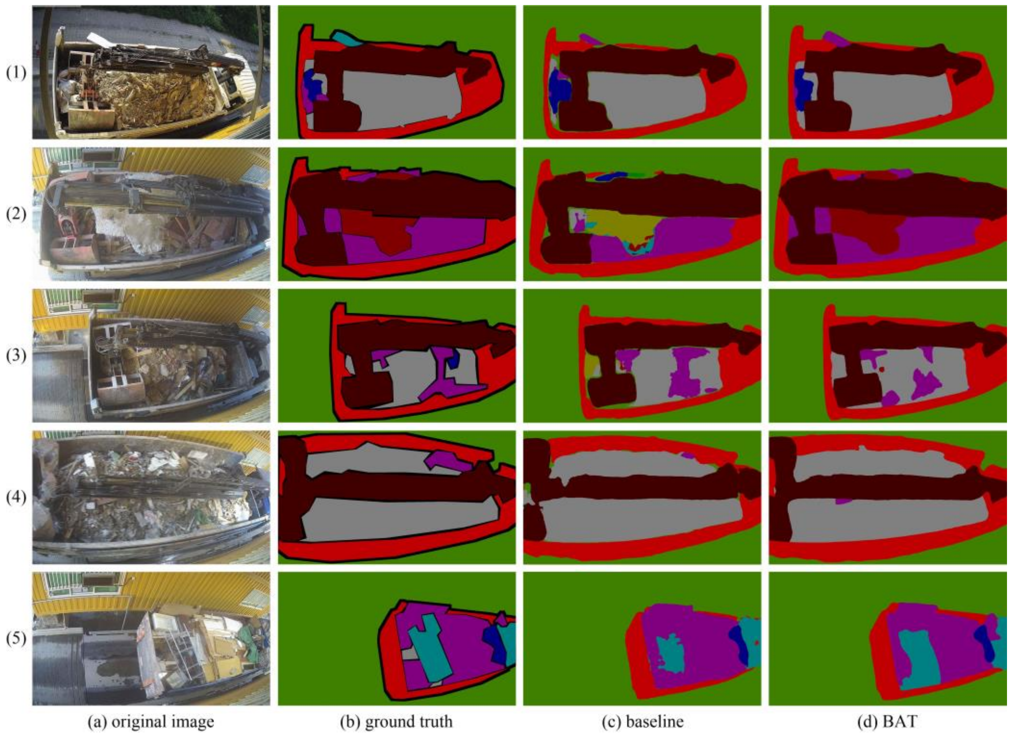


Fig. 7. Examples of segmentation results.

4.4.2 Analysis of erroneous cases

Some erroneous cases are examined in this section. As detailed in section 4.3 the MIoU of the proposed method is 61.68%. MIoUPrediction results of three selected samples and their corresponding ground truth are shown in Fig. 8, of which overall IoU and the IoU for each categorie are listed in Table 5. As shown in Fig. 8, (d), (e) and (f) are the predicted label, (a), (b) and (c) are the corresponding ground truth. Those examples are represented as #1, #2 and #3. While the overall MIoU for #3 (67.16%) exceeds the average value of 61.68%, those for MIoU#1 and #2 (56.51% and 50.88%, respectively) are below the bar. We can see that the proposed method has a better performance for some majority categories such as grip or truck, for which the proposed BAT method can predict their shapes and boundaries more accurately. For some minority categories, the corresponding pixel areas have not been predicted well. For example, pixels belong to the "wood" category only take up 0.83% in the entire image of #3, which is a minority category. In #3, the IoU of the "wood" category is only 12.76%, and since the MIoU is defined as the average of IoU over all categories, the low IoU of several categories (e.g., the "wood" category in image #3) can significantly undermine the final result For some categories with fewer pixels, if no pixels are predicted to be in this category, the IoU is 0, which will have a greater impact on MIoU. For example, packaging category in #1, the pixel ratio is 0.28%, and the category IoU is 0. The category imbalance problem degrades the model performance. Although this research has tryed several techniques (e.g.,weighted cross-entropy loss (Aurelio et al., 2019), focal loss (Lin et al., 2017) or over-sampling) to deal with the problem, further research is still required to better handle its negative effects.

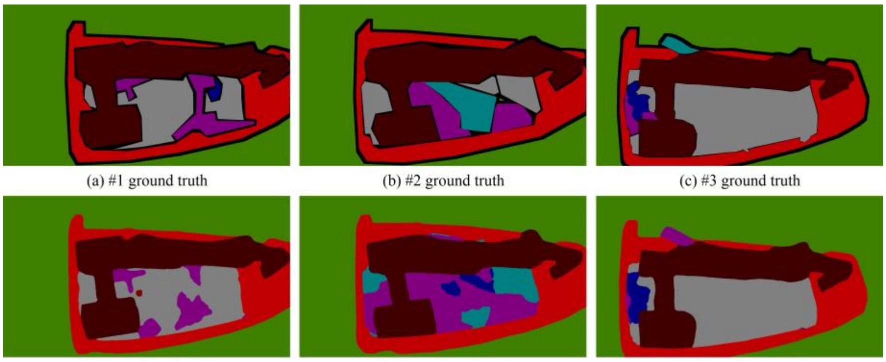


Fig. 8. Examples showing unsatisfied prediction results.

Table 5. The IoU of each category. The "/ " means that no pixels fall into this category in ground truth.

	background	rock	gravel	earth	packaging	wood	others	mixed	grip	truck	total
Palette											
#1	99.79	0	/	/	0	39.86	0	67.1	97.28	91.52	56.51
#2	99.48	/	/	/	0	53.78	11.07	0	96.38	95.46	50.88
#3	99.99	/	/	/	67.44	12.76	0	96.29	97.94	95.67	67.16

5. Conclusions

Precise composition information is a prerequisite of effective construction waste management. Semantic segmentation, a computer vision subtask, has been used to automatically recognize material composition of construction waste mixtures from images.

However, the performance of previous research methods is not sufficient for practical engineering applications. This study proposed a boundary-aware Transformer (BAT) framework for fine-grained composition recognition of construction waste mixture. The model first applies morphology operation to distinguish the background and boundary; a Transformer-based semantic segmentation method is proposed to segment construction waste; finally, a deep learning-based boundary refinement scheme is used to refineboundaries of the segmentation results. Comprehensive ablation experiments were implemented to investigate the effects of different modules of the BAT model. It was found that all of the proposed modules have contributed positively to the improvements in performance. The optimal performance of our framework was compared with that of other state-of-the-art segmentation models. The MIoUof the proposed method is 61.68%, which is 9.8% higher than the baseline. The results demonstrate the effectiveness of the BAT model in improving the performance of construction waste image segmentation.

In future research, the problem of category imbalance should be further researched for better performance. The proportion of each category can be balanced through some technical solutions. For example, re-collecting data to narrow the gap between the majority category and the minority category. In addition, it might be viable to crop the images to patches, from which patches of the rare categories can be over-sampled to balance the dataset. Improving the image quality by updating the camera also has potential to improve the performance, since images with higher resolution can distinguish the category boundaries better, and more details of CW can be preserved in the images.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This research is jointly supported by the Strategic Public Policy Research (SPPR) Funding Scheme (Project No.: S2018.A8.010.18S) and the Environment and Conservation Fund (ECF) (Project No.: ECF 2019-111) of the Government of the Hong Kong Special Administrative Region.

References

Anjum, M., Umar, M.S., 2018. Garbage Localization Based on Weakly Supervised Learning in Deep Convolutional Neural Network, Proceedings - IEEE 2018 International Conference on Advances in Computing, Communication Control and Networking, ICACCCN 2018, pp. 1108-1113.

Aurelio, Y.S., de Almeida, G.M., de Castro, C.L., Braga, A.P., 2019. Learning from Imbalanced Data Sets with Weighted Cross-Entropy Function. Neural processing letters 50, 1937-1949.

Avery Weigh-Tronix, 2010. Driver Operated Weighbridge System at a Wrg Waste Transfer Station. https://www.youtube.com/watch?v=Ukz_t7wDZzk (Accessed August 30 2021).

Awe, O., Mengistu, R., Sreedhar, V., 2017. Smart Trash Net: Waste Localization and Classification, arXiv preprint.

Aziz, F., Arof, H., Mokhtar, N., Shah, N.M., Khairuddin, A.S.M., Hanafi, E., Talip, M.S.A., 2018. Waste Level Detection and Hmm Based Collection Scheduling of Multiple Bins. PLoS ONE 13. 10.1371/journal.pone.0202092.

Bhola, R., Krishna, N.H., Ramesh, K.N., Senthilnath, J., Anand, G., 2018. Detection of the Power Lines in Uav Remote Sensed Images Using Spectral-Spatial Methods. Journal of Environmental Management 206, 1233-1242. https://doi.org/10.1016/j.jenvman.2017.09.036.

Bircanoğlu, C., Atay, M., Beşer, F., Genç, Ö., Kızrak, M.A., 2018. Recyclenet: Intelligent Waste Sorting Using Deep Neural Networks, 2018 Innovations in Intelligent Systems and Applications (INISTA), pp. 1-7.

Brisola, D.F., Cunha, B.M., Gomes, O., Lima, P., Paciornik, S., 2010. Automatic Classification of Particles from Construction and Demolition Waste through Digital Image Analysis, 65th ABM International Congress, 18th IFHTSE Congress and 1st TMS/ABM International Materials Congress 2010, pp. 3046-3052.

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., 2020. Language Models Are Few-Shot Learners. arXiv preprint arXiv:2005.14165.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., 2020. End-to-End Object Detection with Transformers, European Conference on Computer Vision. Springer, pp. 213-229.

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. https://doi.org/10.1016/j.jenvman.2021.112233.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation, Proceedings of the European conference on computer vision (ECCV), pp. 801-818.

Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., Sutskever, I., 2020. Generative Pretraining from Pixels, International Conference on Machine Learning. PMLR, pp. 1691-1703.

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.

Faibish, S., Bacakoglu, H., Goldenberg, A.A., 1997. An Eye-Hand System for Automated Paper Recycling, Proceedings of International Conference on Robotics and Automation, pp. 9-14 vol.11.

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, pp. 3146-3154.

Hannan, M.A., Arebey, M., Begum, R.A., Basri, H., Al Mamun, M.A., 2016. Content-Based Image Retrieval System for Solid Waste Bin Level Detection and Performance Evaluation, Waste Management, pp. 10-19.

HKEPD, 2020. Hong Kong Waste Treatment and Disposal Statistics. https://www.epd.gov.hk/epd/english/environmentinhk/waste/data/stat_treat.html (Accessed April 7 2021).

Hoornweg, D., Bhada-Tata, P., 2012. What a Waste: A Global Review of Solid Waste Management. World Bank, Washington, DC.

Huang, G.-L., He, J., Xu, Z., Huang, G., 2020. A Combination Model Based on Transfer Learning for Waste Classification. Concurrency and Computation: Practice and Experience 32, e5751. https://doi.org/10.1002/cpe.5751.

Ku, Y., Yang, J., Fang, H., Xiao, W., Zhuang, J., 2020. Deep Learning of Grasping Detection for a Robot Used in Sorting Construction and Demolition Waste. Journal of Material Cycles and Waste Management. 10.1007/s10163-020-01098-z.

Kujala, J.V., Lukka, T.J., Holopainen, H., 2015. Picking a Conveyor Clean by an Autonomously Learning Robot, arXiv preprint arXiv:1511.07608.

Lau Hiu Hoong, J.D., Lux, J., Mahieux, P.-Y., Turcry, P., Aït-Mokhtar, A., 2020. Determination of the Composition of Recycled Aggregates Using a Deep Learning-Based Image Analysis. AUTOMAT CONSTR 116, 103204. https://doi.org/10.1016/j.autcon.2020.103204.

Leung, C., Wong, S.K., 2004. The Construction and Related Industries in a Changing Socio-Economic Environment: The Case of Hong Kong.

Liang, S., Gu, Y., 2021. A Deep Convolutional Neural Network to Simultaneously Localize and Recognize Waste Types in Images. Waste Management 126, 247-257. https://doi.org/10.1016/j.wasman.2021.03.017.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal Loss for Dense Object Detection, Proceedings of the IEEE international conference on computer vision, pp. 2980-2988.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully Convolutional Networks for Semantic Segmentation, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3431-3440.

Loshchilov, I., Hutter, F., 2017. Decoupled Weight Decay Regularization. arXiv preprint arXiv:1711.05101.

Lu, W., Chen, J., Xue, F., 2021. Using Computer Vision to Recognize Composition of Construction Waste Mixtures: A Semantic Segmentation Approach. Resources, Conservation & Recycling Manuscript submitted for publication (under 2nd review).

Lukka, T.J., Tossavainen, T., Kujala, J.V., Raiko, T., 2014. Zenrobotics Recycler–Robotic Sorting Using Machine Learning, Proceedings of the International Conference on Sensor-Based Sorting (SBS), pp. 1-8.

Mansouri, I., 2019. Computer Vision Part 6: Semantic Segmentation, Classification on the Pixel Level. https://medium.com/analytics-vidhya/computer-vision-part-6-semantic-segmentation-classification-on-the-pixel-level-ee9f5d59c1c8 (Accessed April, 7 2021).

Mao, W., Chen, W., Wang, C., Lin, Y., 2021. Recycling Waste Classification Using Optimized Convolutional Neural Network. Resources, Conservation and Recycling 164, 105132. https://doi.org/10.1016/j.resconrec.2020.105132.

Meng, S., Chu, W., 2020. A Study of Garbage Classification with Convolutional Neural Networks, 2020 Indo – Taiwan 2nd International Conference on Computing, Analytics and Networks (Indo-Taiwan ICAN), pp. 152-157.

Nowakowski, P., Pamuła, T., 2020. Application of Deep Learning Object Classifier to Improve E-Waste Collection Planning. Waste Management 109, 1-9. https://doi.org/10.1016/j.wasman.2020.04.041.

NSWEPA, 2018. Waste Levy Guidelines. https://www.epa.nsw.gov.au/-/media/epa/corporate-site/resources/wasteregulation/181272-waste-levy-guidelines.pdf (Accessed August 19 2021).

Özkan, K., Ergin, S., Işık, Ş., Işıklı, İ., 2015. A New Classification Scheme of Plastic Wastes Based Upon Recycling Labels, Waste Management, pp. 29-35.

Panwar, H., Gupta, P.K., Siddiqui, M.K., Morales-Menendez, R., Bhardwaj, P., Sharma, S., Sarker, I.H., 2020. Aquavision: Automating the Detection of Waste in Water Bodies Using Deep Transfer Learning. Case Studies in Chemical and Environmental Engineering 2, 100026. https://doi.org/10.1016/j.cscee.2020.100026.

Paulraj, S.G., Hait, S., Thakur, A., Asme, 2016. Automated Municipal Solid Waste Sorting for Recycling Using a Mobile Manipulator, Proceedings of the Asme International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, 2016.

Proença, P.F., Simões, P., 2020. Taco: Trash Annotations in Context for Litter Detection. arXiv preprint arXiv:2003.06975.

Ramli, S., Mustafa, M.M., Wahab, D.A., Hussain, A., 2010. Plastic Bottle Shape Classification Using Partial

Erosion-Based Approach, 2010 6th International Colloquium on Signal Processing & its Applications, pp. 1-4.

Sauve, G., Van Acker, K., 2020. The Environmental Impacts of Municipal Solid Waste Landfills in Europe: A Life Cycle Assessment of Proper Reference Cases to Support Decision Making. Journal of Environmental Management 261, 110216. https://doi.org/10.1016/j.jenvman.2020.110216.

Sobel, I., 2014. History and Definition of the Sobel Operator. Retrieved from the World Wide Web 1505.

Srinilta, C., Kanharattanachai, S., 2019. Municipal Solid Waste Segregation with Cnn, 2019 5th International Conference on Engineering, Applied Sciences and Technology (ICEAST), pp. 1-4.

Sun, K., Zhao, Y., Jiang, B., Cheng, T., Xiao, B., Liu, D., Mu, Y., Wang, X., Liu, W., Wang, J., 2019. High-Resolution Representations for Labeling Pixels and Regions. arXiv preprint arXiv:1904.04514.

Takase, S., Kiyono, S., 2021. Lessons on Parameter Sharing across Layers in Transformers. arXiv preprint arXiv:2104.06022.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention Is All You Need, Advances in neural information processing systems, pp. 5998-6008.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., 2020. Scipy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature methods 17, 261-272.

Wang, X., Girshick, R., Gupta, A., He, K., 2018. Non-Local Neural Networks, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7794-7803.

Wang, Z., Li, H., Zhang, X., 2019a. Construction Waste Recycling Robot for Nails and Screws: Computer Vision Technology and Neural Network Approach. AUTOMAT CONSTR 97, 220-228. https://doi.org/10.1016/j.autcon.2018.11.009.

Wang, Z., Peng, B., Huang, Y., Sun, G., 2019b. Classification for Plastic Bottles Recycling Based on Image Recognition, Waste Management, pp. 170-181.

Xiao, W., Yang, J., Fang, H., Zhuang, J., Ku, Y., 2020. Classifying Construction and Demolition Waste by Combining Spatial and Spectral Features, Proceedings of the Institution of Civil Engineers - Waste and Resource Management. ICE Publishing, pp. 79-90.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P., 2021. Segformer: Simple and Efficient Design for Semantic Segmentation with Transformers. arXiv preprint arXiv:2105.15203.

Yang, J., Zeng, Z., Wang, K., Zou, H., Xie, L., 2021. Garbagenet: A Unified Learning Framework for Robust Garbage Classification. IEEE Transactions on Artificial Intelligence, 1-1. 10.1109/TAI.2021.3081055.

Yang, M., Thung, G., 2016. Classification of Trash for Recyclability Status, CS229 Project Report.

Yuan, L., Lu, W., Xue, F., 2021a. Estimation of Construction Waste Composition Based on Bulk Density: A Big Data-Probability (Bd-P) Model. Journal of Environmental Management 292, 112822. https://doi.org/10.1016/j.jenvman.2021.112822.

Yuan, Y., Chen, X., Chen, X., Wang, J., 2021b. Segmentation Transformer: Object-Contextual Representations for Semantic Segmentation, European Conference on Computer Vision (ECCV).

Yuan, Y., Xie, J., Chen, X., Wang, J., 2020. Segfix: Model-Agnostic Boundary Refinement for Segmentation, European Conference on Computer Vision. Springer, pp. 489-506.

Zhang, Q., Zhang, X., Mu, X., Wang, Z., Tian, R., Wang, X., Liu, X., 2021. Recyclable Waste Image Recognition Based on Deep Learning. Resources, Conservation and Recycling 171, 105636. https://doi.org/10.1016/j.resconrec.2021.105636.