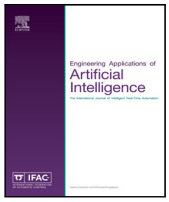
Engineering Applications of Artificial Intelligence 128 (2024) 107542

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



Hierarchical waste detection with weakly supervised segmentation in images from recycling plants

Dmitry Yudin [a,b,*], Nikita Zakharenko [c], Artem Smetanin [d,e], Roman Filonov [b], Margarita Kichik [b], Vladislav Kuznetsov [b], Dmitry Larichev [b], Evgeny Gudov [d], Semen Budennyy [a,c], Aleksandr Panov [a,b,*]

[a] Artificial Intelligence Research Institute (AIRI), 32 Kutuzovsky Ave., Moscow, 121170, Russia
[b] Moscow Institute of Physics and Technology, 9 Institutsky per., Dolgoprudny, 141701, Moscow Region, Russia
[c] Sber AI Lab, 32 Kutuzovsky Ave., Moscow, 121170, Russia
[d] Planetarium One, Naberezhnaya Obvodnogo kanala, 74c, Saint-Petersburg, 196084, Leningrad Oblast, Russia
[e] National Research University ITMO, Kronverkskiy 49, Saint-Petersburg, 197101, Leningrad Oblast, Russia

ARTICLE INFO

MSC:
C1250M
C5260B
C5530
C6260
C6261
C6264
E1840

Keywords:
Hierarchical detection
Waste recognition
Weakly supervised segmentation
Image processing
Recycling plant

ABSTRACT

Reducing environmental pollution with household waste and emissions from the computing clusters is an urgent technological problem. In our work, we explore both of these aspects: the deep learning application to improve the efficiency of waste recognition on recycling plant's conveyor, as well as carbon dioxide emission from the computing devices used in this process. To conduct research, we developed an unique open WaRP dataset that demonstrates the best diversity among similar industrial datasets and contains more than 10,000 images with 28 different types of recyclable goods (bottles, glasses, card boards, cans, detergents, and canisters). Objects can overlap, be in poor lighting conditions, or significantly distorted. On the WaRP dataset, we study training and evaluation of cutting-edge deep neural networks for detection, classification and segmentation tasks. Additionally, we developed a hierarchical neural network approach called H-YC with weakly supervised waste segmentation. It provided a notable increase in the detection quality and made it possible to segment images, learning only having class labels, not their masks. Both the suggested hierarchical approach and the WaRP dataset have shown great industrial application potential.

1. Introduction

The problem of garbage pollution reaches dangerous proportions (Hoornweg et al., 2013). It is predicted that by the end of the 21st century, the amount of garbage produced will reach 11 million tons per day. The main danger of garbage accumulation is a decrease of harmless organic waste and an increase of chemical active products in waste. Plastic garbage have radically changed the situation because it does not decompose. It can be recycled, but there is no adequate system for its storage. To solve the problem with garbage most effectively, it must be automatically sorted. For this purpose, robotic conveyor lines are commonly used. They are equipped with industrial manipulators and video cameras, capable of localizing the desired categories of waste and carrying out its capture and separation (Zhihong et al., 2017).

The development of such systems requires the creation of algorithms and software that reliably allow to recognize images by performing the detection of bounding boxes, classifying objects and segmenting them (Ni et al., 2021; Bashkirova et al., 2022; Demetriou et al., 2023). Accurate detection and segmentation are needed to determine the object location for the capture by the actuator, which is usually a pneumatic sucker (Koskinopoulou et al., 2021).

Such tasks are most effectively solved by deep neural networks (Chen and Xiong, 2020; Bobulski and Kubanek, 2021a; Terven and Cordova-Esparza, 2023). They have a significant limitation, which is the availability of a labeled suitable dataset for a specific task of classification, detection and/or image segmentation. Currently, there are no universal datasets for detecting and segmenting waste on a conveyor belt of recycling plant. This is because of a great variety

* Corresponding authors at: Moscow Institute of Physics and Technology, 9 Institutsky per., Dolgoprudny, 141701, Moscow Region, Russia.
E-mail addresses: yudin@airi.net (D. Yudin), nnzakharenko@sberbank.ru (N. Zakharenko), artem_smetanin@niuitmo.ru (A. Smetanin), filonov.rs@phystech.edu (R. Filonov), kichik.mg@phystech.edu (M. Kichik), vd.kuznetcov@yandex.ru (V. Kuznetsov), larichev.diu@phystech.edu (D. Larichev), evgenes@gmail.com (E. Gudov), budennyy@airi.net (S. Budennyy), panov@airi.net (A. Panov).

https://doi.org/10.1016/j.engappai.2023.107542
Received 17 July 2023; Received in revised form 21 October 2023; Accepted 15 November 2023
Available online 19 November 2023

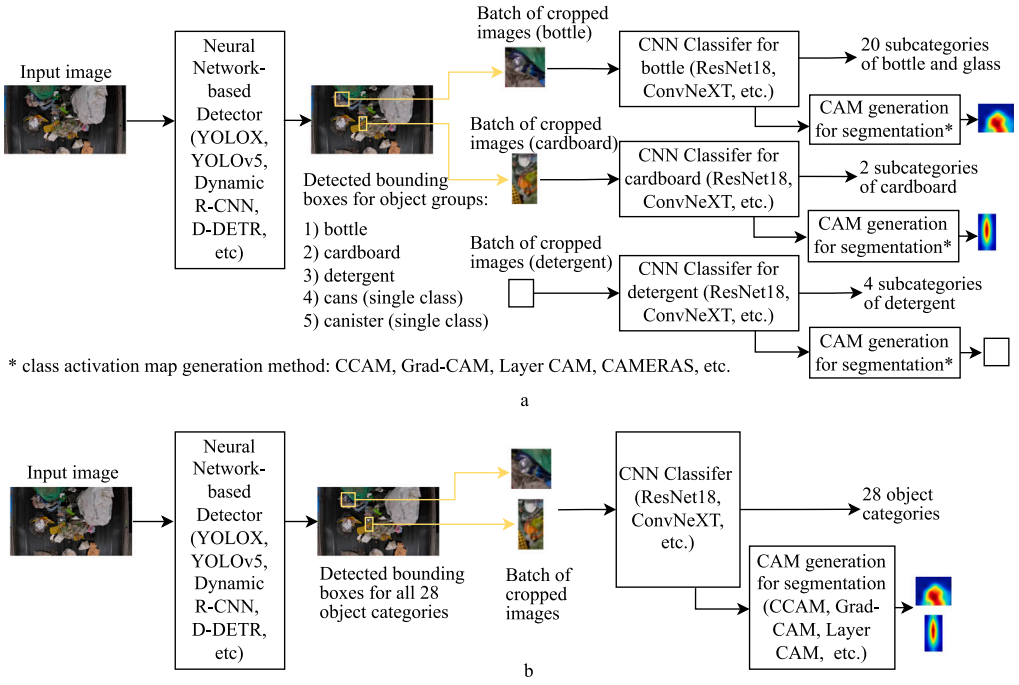


Fig. 1. Variants of the proposed hierarchical detector scheme: (a) H-YC(5) network, which detects five joint object groups and uses three additional CNN classifiers and class activation map generators for 'bottle', 'cardboard', and 'detergent' groups, (b) more simple H-YC(28) network, which detects bounding boxes for all 28 object categories and uses one classifier with class activation map generator. Such hierarchical model architectures and problem formulation for simultaneous detection and weakly supervised segmentation of described waste categories have not been considered in previous studies.

of processed objects, their possible overlap or deformations in camera images.

The focus of this article is creation of a large custom dataset from real recycling plant and exploring the possibilities of modern deep learning approaches for waste recognition.

Contributions. In this article we propose a novel architecture of a hierarchical neural network called H-YC (see Fig. 1) that improves the quality of state-of-the-art object detection methods thanks to the developed joint learning algorithm with an additional classifier and the possibility of weakly supervised segmentation. A particular attention is paid to low response time models for their suitability to operate on the equipment of processing plants in real time mode.

To train the neural network, a new special open WaRP dataset was developed. This is the largest diverse dataset containing 28 object categories that can be found on the conveyor belt of recycling plants. It includes subcategories of bottles, glasses, card boards, cans, detergents, canisters that can overlap, be heavily deformed, or be in non-satisfactory lighting conditions.

The dataset and the implementation of hierarchical network modules are publicly available at https://github.com/AIRI-Institute/WaRP and on Kaggle platform.[1]

2. Related work

Waste classification. A number of papers consider waste recognition in images only as a classification problem. For example, Bircanoğlu et al. (2018) investigates ResNet50, MobileNet, Inception-ResnetV2, DenseNet121 and Xception models. They demonstrate acceptable quality on a dataset with 6 garbage categories images taken in good lighting conditions and without object overlap. Zhang et al. (2021b) proposed a simple ResNet-18-like convolutional model of waste classification. This demonstrated high quality recognition of cardboard, glass, metal, plastic and trash categories in good imaging conditions. Another example is the usage of EfficientNet classification models for waste image samples from ImageNet dataset (Malik et al., 2022).

There are methods for hierarchical two-stage waste classification based on popular feature extractors and neural network ensembles, for example, an accurate combined classification model based on the modified NASNetLarge encoder (Huang et al., 2020).

In the paper (Zhang et al., 2021a), the authors solve the problems of data imbalance, the same type of background and small image size using transfer learning with the DenseNet169 model. Using examples or prototypes of objects may improve the quality of the garbage classification on unbalanced samples, as well as classify new object classes that were not in the training dataset, as shown in the recent work (Han et al., 2023).

Mao et al. (2021) used the DenseNet121 model with image augmentation and a genetic algorithm to select hyperparameters. Binary classification of plastic waste using the Capsule neural networks allows marginally superior to simple convolutional neural networks under similarly good imaging conditions (Sreelakshmi et al., 2019).

There are hybrid approaches based on convolutional neural networks and multilayer perceptrons, which use information from extra sensors, in addition to the camera (Chu et al., 2018). This improves the quality, but is not always technically feasible in real practice. A good improvement in waste recognition was achieved by Ahmad et al. (2020), who used a hybrid classification model composed of models of different architectures. The disadvantage of this work is that the data used contained images with the uniform background. This is rarely seen in the industrial environment of recycling plants.

Vision transformer based on hybrid convolution neural network, proposed by Alrayes et al. (2023) showed an advantage in waste classification quality over some conventional convolutional models such as

[1] https://www.kaggle.com/datasets/parohod/warp-waste-recycling-plant-dataset.

Fig. 2. Parts of the developed WaRP dataset: WaRP-D — images with bounding box labeling for the detection task, WaRP-C — cropped images with class labels, WaRP-S — cropped images with labeling for weakly supervised segmentation.

ResNet50 and MobileNet on the TrashNet dataset (Thung and Yang, 2017). What has not been studied enough is the use of transformer architectures to classify waste on more diverse datasets and compare them with more modern convolutional models.

To classify a sequence of images with solid waste, Li and Chen (2023) propose a convolution neural network with Graph long short-term memory. It should be noted that there are practically no labeled datasets for training such models in the public domain.

Waste detection and segmentation. Some studies analyze the waste detection problem based on the single-stage YOLO neural network model family (Terven and Cordova-Esparza, 2023), for example YOLOV4 (Chen and Xiong, 2020), YOLOv5 (Yang et al., 2021), YOLOv6-v7 (Demetriou et al., 2023), YOLOv8 (Bawankule et al., 2023) models. They note the high speed of these methods, but a significant dependence of the recognition quality on the training dataset. It should be noted that there are also more complex approaches specialized for a specific task object detection with unstructured background (Tang et al., 2023a). Examples are classical image processing and YOLOv7 fusion algorithm (Zhou et al., 2022), 3D object detection using stereo vision with YOLOv3-v5 models (Tang et al., 2023b) or based on Depth image analysis and YOLOv5 model (Wu et al., 2022).

Simultaneous detection and segmentation of waste objects on the conveyor can be carried out using the Mask R-CNN (Koskinopoulou et al., 2021) two-stage model which is trained in a supervised manner and requires a large set of target objects labeled for the segmentation task. In general, two-stage RCNN-based waste detectors are usually significantly inferior in performance to single-stage models (SSD, YOLO) (Demetriou et al., 2023).

As for object segmentation, there are many supervised realtime segmentation models, for example, convolutional ones like DeepLabv3+ (Wu et al., 2023) or transformer ones like Segformer (Xie et al., 2021). However, they all assume well-labeled datasets with feature masks

It is worth studying the possibility of unsupervised or weakly supervised waste segmentation, which does not require the presence of segmentation masks in the dataset, but only information about belonging to one or another category of the whole image. This allows us to significantly save resources for labeling the dataset, and to quickly adapt recognition algorithms to a new domain (for example, associated with new camera installation locations, etc.)

In Bashkirova et al. (2022), the authors considered various neural network methods for deformable object segmentation in cluttered scenes. They conducted a study of fully-, semi-, and weakly-supervised learning for garbage segmentation, which demonstrated a significant superiority of methods implementing fully supervised approaches based on DeepLabv3+ and poor results for weakly-supervised methods (CAM, PuzzleCAM, EPS). There are another approaches based on class activation map generation for segmentation without direct supervision: Grad-CAM (Selvaraju et al., 2017), CAMERAS (Jalwana et al., 2021), Layer CAM (Jiang et al., 2021), contrastive learning based CAM (Xie et al., 2022), etc. Often, methods for constructing class activation maps are used not for explicit segmentation, but for visualizing and explaining the results of object classification, as for example, in the paper (Mao et al., 2021) authors used Grad-CAM to explain the results of hierarchical recycling waste classification.

In our article, we show that the use of weakly supervised segmentation methods as a part of the hierarchical detector allows us to achieve a sufficiently high quality of segmentation without ground truth masks labeling.

Datasets for waste recognition. A great number of datasets with various waste images have appeared recently. Some of them contain photographs of littered nature or urban infrastructure (UAVVaste Kraft et al., 2021), others are collected from photographs of various packaging items against a neutral background (TACO Proença and Simões, 2020, Trashnet Thung and Yang, 2017). The most popular modern datasets are listed in Table 1. Each of the mentioned datasets contains waste categories that we are interested in, such as plastic bottles or cans, but the environment of such objects in the photographs does not look like the one seen on a conveyor belt.

Among other datasets, the ZeroWaste Dataset (Bashkirova et al., 2022) stands out: this dataset contains photos of a transporter line at the paper recycling plant. There are such categories as metal, cardboard and plastic in the dataset markup. ZeroWaste Dataset is designed to solve the problem of paper waste segregation, while our recognition task includes different types of packaging for drinking and household fluids. To meet the specific challenges of recycling plants, this paper considers the creation of a new dataset, which is added to Table 1 and called WaRP (Fig. 2). It is described in detail in the next Section.

3. WaRP dataset

Waste recycling plants need to automatically select and sort recyclable items on the conveyor. In our case, these objects should fall into several main categories: plastic and glass bottles, card boards, detergents, canisters and cans. For the first three categories, it is desirable to know what color they are and what they are used for, since recycling technologies differ. There are no open datasets containing all the required object categories for such an application. Therefore, there was a need to develop our own dataset in order to train and test methods for detecting, classifying and segmenting waste on it.

Our dataset named WaRP (abbreviation of Waste Recycling Plant) consists of manually labeled pictures of an industrial conveyor. We selected 28 recyclable waste categories. Objects in the dataset are divided into the following groups (see Table 2): plastic bottles of 17 categories (class name with the bottle-prefix), glass bottles of three types (the glass-prefix), card boards of two categories, detergents of four categories, canisters and cans. The -full postfix means that the bottle is filled with air, i.e. not flat. This is important for the correct work of the manipulator on the conveyor. Examples of instances of each category of the WaRP Dataset are presented in Fig. 3. An important difference from other datasets is that objects can overlap, be heavily deformed, or be in poor lighting conditions.

It should be noted that the collected and manually labeled dataset is unbalanced; for example, there are significantly more objects of bottles than canisters. This is due to the fact that household waste on the conveyor belt of a recycling plant has a natural uneven distribution due to the different frequency of use of various objects.

The dataset has three parts (see Fig. 2): WaRP-D, WaRP-C, and WaRP-S. The first two parts are intended for training and objective

Table 1
Modern datasets for waste recognition images.

Dataset	Categories (sub-categories)	Images	Task	Description
Trashnet (Thung and Yang, 2017)	6	2527	Classification	Contains 501 annotation per category glass and 482 per plastic
Glassense-Vision (Sosa-García and Odone, 2017)	7 (136)	2000	Classification	144 pictures with bottles and 158 pictures with cans inside
Open litter map (GeoTech Innovations, 2020)	11 (187)	>100,000	Multilabel classification	This is a website that collects a dataset from images with garbage from all over the world
Spotgarbage (Mittal et al., 2016)	3	~2400	Classification	Pictures with garbage on the streets (scraped from Bing search)
Waste Class. data v2 (Sapal6, 2019)	3	~27,500	Classification	Organic, recyclable and non-recyclable categories, pictures scraped from google search
Waste_pictures (Wang, 2019)	34	~24,000	Classification	There are 209 cans, 201 glass bottles and 160 plastic bottles in this dataset, all of them were scrapped from Google
Wade-ai (Haamer, 2020)	1	>1500	Instance-segmentation	Outdoor images with different garbage
TACO (Proença and Simões, 2020)	28 (60)	1500	Segmentation	Dataset contains ~420 annotations per bottle category and ~230 annotations per can category
Sushi Restaurant (Cen, 2020)	16	500	Classification	Dataset contains 61 annotations per plastic cup category, 37 per plastic utensil category and other items
Litter (Mikołajczyk, 2021)	24	~14,000	Detection	This is a website with limited access to datasets
Drinking Waste Class (Serezhkin, 2020)	4	9640	Detection	Dataset contains images with cans (~1000), glass bottles (~1200), plastic bottles (~2500)
MJU-Waste v1.0 (Wang et al., 2020)	1	2475	Segmentation	Contains photographs of people holding different types of garbage in their hands (one category — garbage)
Google Open Images (Kuznetsova et al., 2020)	3	14,226	Detection, Instance-seg.	Outdoor and indoor images (bottles, plastic bags, tin cans)
UAVVaste (Kraft et al., 2021)	1	772	Segmentation	Dataset contains 772 pictures from drone camera with different rubbish
WaDaBa (Bobulski and Kubanek, 2021b)	8	4000	Classification	All images contain objects made of different type of plastic
Domestic Trash (Datacluster-labs, 2021)	10	>9000	Classification, Detection	Waste in the wild, paid license, 250 images for free
NWNU-TRASH (Zhang et al., 2021a)	5	20,000	Classification	Images with heterogeneous background
ZeroWaste Dataset (Bashkirova et al., 2022)	4	12,125	Detection, Segmentation	Conveyor images, contain cardboard, metal and plastic objects
WRTD (ZotBins, 2023)	61	3010	Detection, Classification, Segmentation	Dataset consists of garbage images in an urban environment
WaRP (ours), 2023	5 (28)	2974 (10,406)	Detection, Classification, Segmentation	Images from the conveyor of recycling plant with categories of bottles, cardboards, detergents, canisters and cans

Fig. 3. Example labeled images (for classes of 'bottle', 'cans', 'cardboard', 'canister', 'detergent') in the WaRP dataset.

quality assessment of detection (WaRP-D) and classification (WaRP-C) tasks, and the third WaRP-S is for validation of weakly supervised segmentation methods. The full statistics of our dataset parts are given in Table 2.

The main dataset part WaRP-D contains 2452 images in the training sample and 522 images in the validation sample. The images have full HD resolution of 1920 × 1080 pixels.

WaRP-C is cut-out image areas from the WaRP-D set with class labels. This part includes 8823 images for training and 1583 for testing. The images range in size from 40 to 703 pixels wide and 35 to 668 pixels high. The dataset is unbalanced because of the real conditions of an industrial enterprise. The rarest class is the bottle-oil-full (air-filled plastic sunflower oil bottles) category, which includes only 32 crops. The most common category is bottle-transp (transparent bottles), with 1667 clipped images.

WaRP-S contains a total of 112 images ranging in size from 100 × 96 pixels to 412 × 510 pixels, each category has 4 images with significantly deformed recyclable objects.

Table 2
WaRP dataset statistics.

Category	WaRP-D-Train	WaRP-D-Test	WaRP-C-Train	WaRP-C-Test	WaRP-C-Total	WaRP-S-Test
Bottle-blue	535	87	634	106	740	4
Bottle-green	403	65	466	75	541	4
Bottle-dark	451	80	533	96	629	4
Bottle-milk	324	54	347	60	407	4
Bottle-transp	947	164	1432	235	1667	4
Bottle-multicolor	125	28	127	31	158	4
Bottle-yogurt	261	41	277	42	319	4
Bottle-blue-full	263	40	285	45	330	4
Bottle-transp-full	457	79	528	93	621	4
Bottle-dark-full	173	31	185	36	221	4
Bottle-green-full	229	33	238	35	273	4
Bottle-multicolor-full	105	20	107	22	129	4
Bottle-milk-full	110	21	110	21	131	4
Bottle-oil	254	46	276	48	324	4
Bottle-oil-full	23	8	24	8	32	4
Bottle-blue5l	345	60	413	75	488	4
Bottle-blue5l-full	87	23	89	24	113	4
Glass-transp	165	34	177	37	214	4
Glass-dark	132	24	136	25	161	4
Glass-green	131	23	135	25	160	4
Juice-cardboard	251	63	260	71	331	4
Milk-cardboard	358	85	390	96	486	4
Detergent-white	300	42	319	44	363	4
Detergent-color	277	43	296	44	340	4
Detergent-transparent	245	39	262	42	304	4
Detergent-box	66	17	66	17	83	4
Canister	144	28	149	30	179	4
Cans	495	88	562	100	662	4
Total	**2452**	**522**	**8823**	**1583**	**10 406**	**112**

4. Neural network for hierarchical waste detection with weakly supervised segmentation

On complex datasets containing images of objects with overlaps and deformations, state-of-the-art detection methods usually work imperfectly and generate false positives and miss objects. Such datasets include the proposed WaRP dataset.

It is promising to improve the quality of pre-trained detection neural network with the additional classification and segmentation modules. On the one hand, this does not require intervention in the architecture of the detector, and on the other hand, it can clarify the assignment of certain labels to the found bounding boxes. Adding the ability to semantic segmentation of objects without resource-intensive supervised learning is also beneficial.

This article proposes to explore two main variants of the hierarchical classifier scheme, which are shown in Fig. 1.

The first option (Fig. 1,a) involves the neural network-based detection of object bounding boxes belonging not to all 28 categories of the WaRP-D dataset, but to 5 “supercategories”: bottle (including glass), card board, detergent, cans and canisters. The first three categories include 20, 2 and 4 subcategories, respectively, and for them it is proposed to train independent classifiers. Their feature maps can be used to generate class activation maps and further segmentation without additional model training and supervision.

The second option (Fig. 1,b) involves the detection by the neural network of objects belonging to 28 categories at once, and further refinement of the found classes using an additional classifier. Its class activation maps can also be used for weakly supervised segmentation. The second option is closer to the industrial application of neural networks, when the modularity of the solution is important.

The detector is separately trained with a supervision on the WaRP-D dataset. As basic models, we investigate fast one-stage models YOLOv3 (Redmon and Farhadi, 2018), YOLOv5 (Yang et al., 2021), YOLOX (Ge et al., 2021), CenterNet (Zhou et al., 2019), two-stage approaches Faster R-CNN (Ren et al., 2015), Dynamic R-CNN (Zhang et al., 2020), Sparse R-CNN (Sun et al., 2021), transformer architectures D-DETR (Zhu et al., 2021), TOOD (Feng et al., 2021).

As basic classification models, it is proposed to study both architectures that have become classical (ResNet (He et al., 2016), DenseNet (Huang et al., 2019), MobileNet (Howard et al., 2017), EfficientNet (Tan and Le, 2019), ResNeXT (Xie et al., 2017)), and more modern neural networks: ConvNeXT (Liu et al., 2022), Vision Transformer (Dosovitskiy et al., 2020), Data-Efficient Image Transformers (Touvron et al., 2021), Swin Transformer (Liu et al., 2021), ReXNet (Han et al., 2021), RepVGG (Ding et al., 2021).

The article also explores 2 training cases for the proposed hierarchical detector. In the first case, the base detector and the classifier learn independently of each other, the base detector learns on images with ground truth (GT) markup from WaRP-D, and the classifier learns on crops from these images included in the WaRP-C sample. In the second case, the classifier is trained on the crops obtained by predicting the WaRP-D training dataset by the basic detector, while the class labels for the crops are assigned based on the intersection over union with the boxes from the original GT-labeling.

As for weakly supervised segmentation, we explore the possibilities of popular methods for constructing class activation maps based on Grad-CAM (Selvaraju et al., 2017) and its modification mGrad-CAM (Kuznetsov and Yudin, 2022) that does not use average pooling, Layer CAM methods (Jiang et al., 2021) and CAMERAS (Jalwana et al., 2021), as well as a new unsupervised approach CCAM (Xie et al., 2022) using contrastive learning. To move from class activation maps to segmentation masks, we use the algorithm proposed by the authors of the current article in Kuznetsov and Yudin (2022).

5. Experimental results and discussion

Waste detection. We performed experiments with different state-of-the-art neural networks on WaRP-D dataset (see Table 3). Each image was annotated with bounding boxes. There was a significant overfitting problem, while training our YOLO models, solved by using an efficient

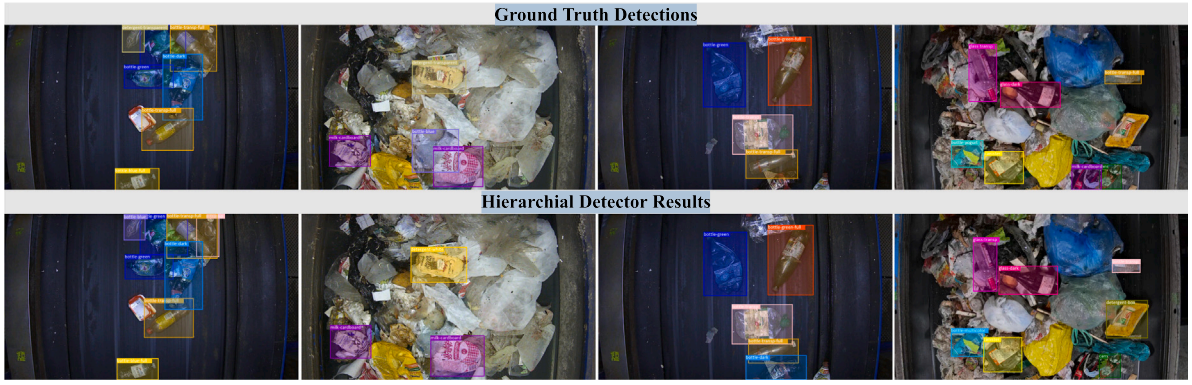


Fig. 4. Examples of Hierarchical detector work on test sample of the WaRP-D dataset.

Table 3
Detection quality on 28 categories of WaRP-D dataset for state-of-the-art detectors. In brackets we show detection metrics for 5 joined categories.

Detector	mAP_{50} (bottle), %	mAP_{50} (cardboard), %	mAP_{50} (detergent), %	AP_{50} (canister), %	AP_{50} (cans), %	mAP_{50}, %	$mAP_{50:95}$, %	FPS (bs = 1)
YOLOV3	44.8 (75.0)	12.9 (37.6)	21.1 (43.4)	27.3 (44.4)	20.6 (49.8)	37.6 (50.0)	26.0 (32.5)	52.9
YOLOV5-x	62.6 (79.6)	41.1 (46.4)	38.6 (48.8)	32.0 (56.8)	58.1 (55.7)	56.4 (57.5)	**46.6** (43.8)	**66.6**
YOLOX-m	63.5 (80.9)	39.7 (**54.6**)	45.0 (48.6)	**54.0** (46.6)	59.3 (**62.9**)	**58.6** (**58.3**)	45.7 (**44.5**)	**64.9**
YOLOX-l	51.8 (80.9)	27.9 (52.3)	27.7 (47.7)	28.1 (44.8)	45.2 (59.5)	45.6 (57.0)	34.6 (43.5)	52.4
D-DETR	60.1 (**83.0**)	**41.3** (48.7)	44.3 (50.9)	43.5 (**57.0**)	55.3 (54.7)	55.7 (**58.9**)	40.3 (42.8)	13.9
Dynamic-RCNN	61.6 (77.2)	35.8 (40.9)	44.9 (**51.2**)	39.7 (50.3)	55.2 (55.2)	56.4 (55.0)	38.3 (40.4)	33.8
Faster-RCNN	41.6 (75.1)	24.3 (39.5)	31.0 (47.6)	24.5 (36.3)	33.4 (47.6)	56.4 (48.0)	38.0 (31.8)	35.3
TOOD	**65.8** (78.5)	34.5 (41.9)	**47.2** (51.0)	52.7 (46.2)	**61.5** (57.5)	**60.2** (55.0)	46.5 (41.4)	28.9
CenterNet	56.0 (76.2)	24.4 (36.3)	36.4 (37.7)	30.5 (38.3)	56.2 (52.1)	50.1 (48.1)	37.6 (33.0)	9.1
ATSS	62.6 (79.0)	32.3 (41.6)	41.9 (48.9)	52.9 (51.5)	51.5 (48.4)	56.7 (53.9)	43.0 (40.6)	38.5
Sparse-RCNN	50.9 (75.0)	24.8 (37.3)	30.2 (40.3)	32.2 (35.7)	45.0 (51.9)	45.2 (48.1)	32.0 (33.0)	27.3

set of augmentations. The highest impact obtained within mosaic augmentation (Ghiasi et al., 2020). MixUp was set to 50%, 90 degrees rotation and resize to 448 × 832 (height/width). Keeping mosaic until about the middle of the process and then turning it off gave huge leap in metrics. It becomes easier for the model to perceive images, a consequence of this mAP_{50} (mean average precision for boxes with intersection over union more than 50%) instantly increase by ∼15%.

We trained YOLOV5 with SGD+Nesterov setting the initial learning rate to 10^{-2}, weight decay $5e{-}4$, initial momentum 0.937. Linear scheduler was used, warmup for 3 epochs. Training was performed on Tesla V100 32 GB.

Fig. 4 contains several more examples of waste detection on the test sample of the proposed WaRP-D dataset. Two images at the bottom line of the figure illustrate working with errors.

Referring to Table 3, we see that the best mAP_{50} results on 28 classes of WaRP-D dataset is obtained by TOOD, which is slightly inferior to the YOLOV5 model in $mAP_{50:95}$. The most problematic classes, processing the lowest metrics are bottle-oil-full, juice-cardboard and bottle-multicolor. We can see that two-stage detectors architectures behave unpredictably, some of them get quite high accuracy on the same category and others show very poor performance, unlike one-stage models, which show themselves well on each class. The YOLOV5 model shows the best inference speed (FPS) and corresponding detection quality. The transformer detector D-DETR loses a lot in speed compared to the YOLO models although it outperforms other models in terms of detection metric mAP_{50} of 5 categories (see Table 3). It should be noted that the fast YOLOX-m model also shows consistently high quality indicators, and is able to recognize objects of 5 categories with the best quality in terms of the $mAP_{50:95}$ metric.

The main criterion for choosing a detector model for our hierarchical approach is the trade-off between its speed and a high quality metric mAP_{50} for all 28 object categories, which indicates correct coarse bounding box detection. We have the two fastest models: YOLOV5-x (66.6 FPS) and YOLOX-m (64.9 FPS), as well as the two most accurate models: TOOD (60.2% mAP) and YOLOX-m (58.6% mAP). Thus, for further experiments in the hierarchical model, we select the YOLOX-m compromise detector.

Classification. As a part of the experiment, several types of classifiers were trained on the WaRP-C dataset. Architecture types and training results are shown in Table 4. Classification model performance is demonstrated in Table 5 We used model implementations from the time deep learning library (Wightman, 2019).

Table 4
Classifier module quality on WaRP-C dataset.

Classifier	Mean recall, %			Recall, %		Accuracy, %
	Bottle	Cardboard	Detergent	Canister	Cans	
ViT (bs8)	49.1	54.4	60.8	46.7	70.4	51.5
DEiT (bs8)	64.6	70.1	59.9	90.0	84.7	68.0
RepVGG (bs8)	55.0	70.0	51.9	83.3	92.9	59.1
SWiN (bs8)	58.8	66.3	42.7	43.3	93.9	64.8
ResNet18 (bs8)	63.8	74.2	53.4	59.0	91.6	67.6
ResNet18 (bs32)	67.6	78.9	65.15	60.0	93.0	74.2
MobileNetv3 (bs8)	70.7	79.3	69.9	66.7	94.9	72.8
MobileNetv3 (bs32)	75.7	82.2	70.3	73.3	92.9	77.4
DenseNet121 (bs8)	75.5	79.6	71.0	90.0	96.9	78.3
DenseNet121 (bs32)	76.8	84.1	72.9	63.3	82.7	76.6
RexNet (bs8)	74.8	78.3	72.4	90.0	94.9	76.8
RexNet (bs32)	79.5	84.0	74.5	80.0	93.9	80.1
EfficientNet-B5 (bs8)	78.8	83.2	67.2	96.7	95.9	79.8
EfficientNet-B5 (bs32)	**79.6**	**83.3**	**76.3**	**86.7**	95.9	**81.9**
ResNeXT (bs8)	78.8	80.9	72.7	86.7	93.9	79.0
ResNeXT (bs32)	76.9	80.9	77.4	83.3	95.9	79.5
ConvNeXT(28) (bs8)	73.7	84.1	69.1	73.3	95.9	78.8
ConvNeXT(28) (bs32)	77.0	83.0	75.0	**86.7**	**98.0**	**81.8**
ConvNeXT(20) (bottle)	75.7	–	–	–	–	75.4
ConvNeXT(2) (cardboard)	–	90.5	–	–	–	90.1
ConvNeXT(4) (detergent)	–	–	91.7	–	–	92.4

Table 5
Classifier module performance on WaRP-C dataset.

Classifier	FPS (bs = 1)	Model name	Params	Layers
ViT	100	ViT_small_resnet50d_s16_224	57M	277
DEiT	120	DEiT_tiny_patch16_224	6M	188
RepVGG	98	RepVGG_a2	28M	306
SWiN	77	SWiN_tiny	28M	217
ResNet18	**230**	Resnet18	12M	71
MobileNetv3	107	MobileNetv3_large_100	5M	195
DenseNet121	42	DenseNet121	8M	433
RexNet	74	ReXNet_100	5M	313
EfficientNet-B5	38	EfficientNet-B5	30M	551
ResNeXT	82	ResNeXT50_32 × 4d	25M	177
ConvNeXT	48	ConvNeXT_tiny	28M	202

For improving the model quality, image augmentation methods were applied. The following augmentation approaches were used: resizing the image with adding paddings for preserving the original image sides ratios; adding a partially covering mask (for helping CAM method to localize as many pixels of the object as possible), 20% of image is closed; random shifts and turns with 80% probability, 0.2 shift limit, 0.2 scale limit, rotate limit of 90 degrees; random changes in brightness and contrast with 50% probability, 0.1 brightness limit; random color changes for each RGB channel with 50% probability, color shift limit of 15; random vertical and horizontal flips with 50% probability.

Each model used pre-trained weights, which were further tuned during the experiments. Cross Entropy Loss was chosen as error function. The models were trained for 40 epochs with an initial learning rate of 0.001, which decreased during the training process if the quality metric on the validation data was not improving over several epochs.

The training and the test datasets had similar unbalanced distribution, so balancing methods like equivalent inter-class sampling and Weighted Cross Entropy Loss did not significantly improve results compared to the conventional training.

From Table 4 with the obtained quality metrics on the WaRP-C dataset, we can see that the highest quality scores are achieved by the ConvNeXt and EfficientNet-B5 models, while the ResNet-18 model is the fastest one. ConvNeXt is also significantly faster than EfficientNet-B5. So, ConvNeXt-tiny is most promising for use as a part of the hierarchical detector. ResNet-18 can also be used if we need the best possible detector speed.

To select a classifier in our hierarchical model, we must provide a trade-off between its accuracy and performance. Therefore, we chose the ConvNeXT classifier architecture, which provided an Accuracy of 81.8% on the WaRP-C dataset and a performance of 48FPS (see Table 5). Its performance is significantly higher than the 38FPS of the competing model EfficientNet-B5, which is better in quality by only an insignificant 0.1%.

Quality of hierarchical waste detection. The quality indicators of various options for implementing a hierarchical approach to waste detection were analyzed. The results are shown in Table 6. We named as H-YC our hierarchical neural networks which include YOLOX-m detection module and ConvNeXt classification module. Table shows that for hierarchical model H-YC(5) independent training of the YOLOX-m(5) detector on 5 detected classes and three ConvNeXt-tiny models for bottles (20 categories), detergents (4 categories) and cardboards (2 categories) does not improve the mAP_{50} metric (the first scheme of the approach demonstrated in Fig. 1,a). So, training of three independent classifiers leads to a significant deterioration in the quality metrics, and such an implementation of the hierarchical detector is inappropriate.

In the same time we have improvement of the mAP_{50} and $mAP_{50..95}$ metrics for option shown in Fig. 1,b for hierarchical network H-YC(28) with YOLOX-m(28) detector and ConvNeXt-tiny(28) classifier trained on the all 28 categories.

Weakly supervised waste segmentation. After classification, CAM (Class Activation Map) methods are applied, which allow one to build classifier attention maps for a given image according to a certain class. These maps are subsequently converted into binary segmentation maps based on the algorithm described in Kuznetsov and Yudin (2022). To assess the quality of the methods, a standard semantic segmentation metric – mIoU (mean Intersection over Union) – was used. The fastest ResNet-18 model was chosen as the base model for the verification of using these methods. The obtained quality scores are listed in Table 7. The visualization of the generated class activation maps and binary object masks is shown in Fig. 5.

The best quality is shown by the unsupervised approach CCAM(5) based on contrastive learning, which is trained for 5 different "supercategories". CCAM(28) trained on the combined 28 categories is slightly inferior to it. Among the rest of the methods, the best approach is CAMERAS, which uses classifier directly trained on 28 classes.

Cropped images as input of segmentation method may suffer from overlapping objects, poor lighting conditions, or significantly distorted. We are primarily interested in the rough segmentation mask, which we obtain automatically in weakly-supervised mode, having a trained classification model on the selected data set. Therefore, if the data used

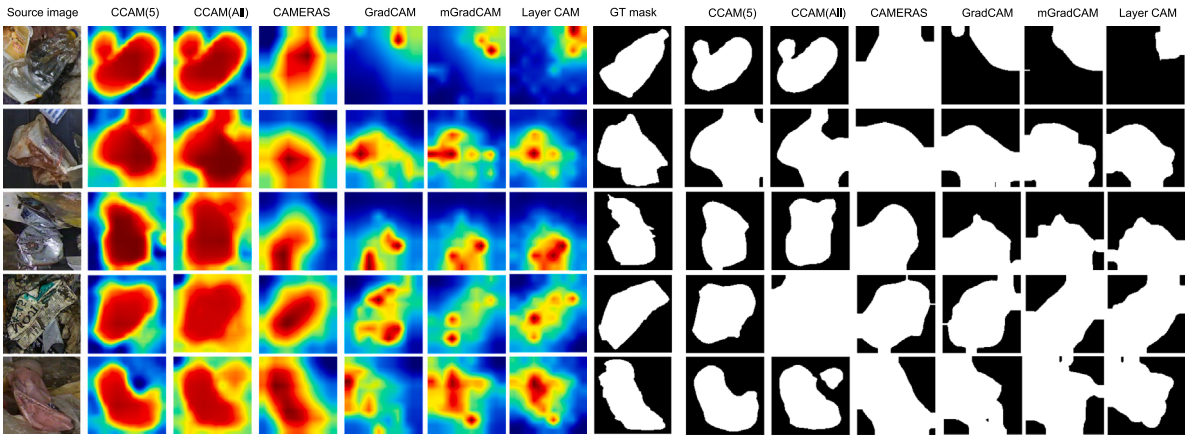


Fig. 5. The results of various weakly supervised waste segmentation approaches. For each image, the generated class activation maps and binarized masks based on them are shown.

Table 6
Hierarchical detector metrics.

Per category AP, %	YOLOX-m(28)	H-YC(5)	H-YC(28)
Bottle-blue-full	**66.8**	51.6	**66.8**
Bottle-transp-full	**70.8**	64.4	70.6
Bottle-dark-full	**85.5**	77.7	79.9
Bottle-green-full	85.3	83.7	**85.6**
Bottle-multicolor-full	77.4	60.9	**78.5**
Bottle-blue5l-full	**84.5**	67.3	**84.5**
Bottle-milk-full	88.6	81.3	**93.9**
Bottle-blue5l	64.2	51.9	**64.5**
Bottle-blue	**59.3**	48.5	58.3
Bottle-green	**74.5**	63.7	74.1
Bottle-dark	73.1	72	**74.1**
Bottle-milk	**46.6**	41.8	46.4
Bottle-transp	**54.6**	40.4	53
Bottle-multicolor	**36.0**	30.4	31.8
Bottle-yogurt	37.9	31	**40.5**
Bottle-oil-full	44.5	35.1	**58.9**
Bottle-oil	22.2	**35.1**	23.6
Glass-transp	53.4	54.7	**55.3**
Glass-dark	**74.7**	73.7	**74.7**
Glass-green	69.2	58.8	**72.3**
Juice-cardboard	35	**40.0**	35.9
Milk-cardboard	**44.5**	38.7	44.2
Cans	59.3	52.9	**61.2**
Canister	54	42.6	**55.1**
Detergent-color	43	34.8	**43.1**
Detergent-transparent	**37.0**	34.4	36.8
Detergent-box	53.3	**68.0**	59.4
Detergent-white	46.7	**47.6**	46.6
mAP_{50}	58.6	52.7	**59.6**
$mAP_{50:95}$	45.7	40.4	**46.7**

to train the base classifier contains object overlaps, various lighting augmentations, or distortions, then acceptable segmentation quality can be expected. However, if there is no such diversity in the training set, then this is a significant limitation of the proposed approach.

6. Results of model integration

After the model integration, useful fractions of garbage detecting experiments were carried out. During the experiment, a small fragment of video from the camera was recorded. Then, the video was viewed by an expert and the number of correctly and incorrectly recognized objects, as well as the number of missed objects, were counted. In different days, 3 experiments were made in total. The duration of a single measurement was 2 min. This period is equivalent to 1830 pictures. As a result, total amount of the analyzed information is 5490 pictures.

For statistics calculation we grouped 28 classes into 4 more general classes: bottle, cardboard, detergent and cans. Canisters were not detected during the experiment. It is important to note that the analyzed data is quite different from the training dataset because camera located at the end of the pipeline. Thus, the problem made by this camera have diverse background, lighting conditions, angle and composition of moving objects. Despite this fact, the model shows good results of detection and classification. F1-score was 63% for detergent recognition, 73% for cardboard, 79% for bottle and 81% for cans. This indicates high generalization ability of the detection model.

7. Conclusion

In the study we proved the problem of waste recognition on the conveyor of recycling plants to be successfully tackled with various architectures of deep neural networks, even being integrated into in-plant exploitation processes.

At the same time, it was noted there were no suitable open datasets containing the required categories of recyclable waste. The created specialized WaRP dataset is a unique and diverse tool that allows to train and test neural network methods for detection (WaRP-D set), classification (WaRP-C set) and segmentation (WaRP-S set) of recyclable waste in non-satisfactory lighting conditions, overlapping and deformations.

The proposed hierarchical approach to waste detection made it possible to improve the quality of the basic pre-trained models, and also to carry out additional weakly supervised object segmentation with acceptable accuracy. Such a solution is practically useful, since for industrial applications it is necessary to constantly expand and re-label the existing dataset in order to provide the best recognition quality for new domains (for different conveyors and plants). For weakly supervised segmentation in the formulation considered, it is sufficient to simply label objects in images as bounding boxes with the object categories. Moreover, for a CCAM algorithm based on contrastive learning, categorization information is not necessary. This labeling is much easier, faster and cheaper to implement.

The experiment with the developed approach at the waste processing complex RT Invest Recycle confirmed its applicability at the conveyor site after the manual waste sorting. The neural network detector of recyclable objects (bottles, card boards, detergents, cans) passed by people showed acceptable precision and recall of recognition. This indicates its superiority over manual conveyor monitoring, which is monotonous and harmful to human health.

Table 7
Weakly supervised segmentation quality, %.

Method	Approach architecture	$mIoU_{bottle}$	$mIoU_{cardb}$	IoU_{cans}	$IoU_{canister}$	$mIoU_{detergent}$	$mIoU_{all}$
CCAM(5)	H-YC(5)	**64.78**	**71.73**	**63.01**	**65.28**	**69.31**	**65.88**
CCAM(28)	H-YC(28)	62.48	66.54	69.18	69.11	65.25	63.64
CAMERAS	H-YC(28)	55.63	59.40	57.83	61.00	60.51	56.87
GradCAM	H-YC(28)	55.01	60.39	38.20	59.31	58.22	55.41
LayerCAM	H-YC(28)	60.19	63.71	44.76	66.77	60.30	60.14
mGradCAM	H-YC(28)	52.87	59.40	39.73	60.29	55.31	53.48

Promising topics for further development of the study are the integration of few shot learning methods for working with rare categories of objects and the issue of quality improving of waste detection not from single images, but from a video sequence.

CRediT authorship contribution statement

Dmitry Yudin: Investigation, Writing – review & editing. **Nikita Zakharenko:** Investigation, Software. **Artem Smetanin:** Writing – original draft, Investigation, Software. **Roman Filonov:** Data curation. **Margarita Kichik:** Investigation, Software. **Vladislav Kuznetsov:** Investigation, Software. **Dmitry Larichev:** Software, Validation. **Evgeny Gudov:** Supervision. **Semen Budennyy:** Supervision. **Aleksandr Panov:** Methodology, Supervision, Writing – review & editing.

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.

Data availability

Data will be made available on request.

Acknowledgments

For providing exclusive data for research, assistance in annotation, prompt consultation on the specifics of the technological process of sorting waste and assistance in integrating the model into production, the authors express their gratitude to waste processing complex RT Invest Recycle, and its director Evgeny Komarov and Planetarium One company and its employees: Natalia Kashirina, Vladislav Makarovsky, Evgeny Yakovlev, Konstantin Roslyakov, Mikhail Shimusyuk, Valeria Kuznetsova, Yankovskiy Nikita.

References

Ahmad, K., Khan, K., Al-Fuqaha, A., 2020. Intelligent fusion of deep features for improved waste classification. IEEE Access 8, 96495–96504.

Alrayes, F.S., Asiri, M.M., Maashi, M.S., Nour, M.K., Rizwanullah, M., Osman, A.E., Drar, S., Zamani, A.S., 2023. Waste classification using vision transformer based on multilayer hybrid convolution neural network. Urban Clim. 49, 101483.

Bashkirova, D., Abdelfattah, M., Zhu, Z., Akl, J., Alladkani, F., Hu, P., Ablavsky, V., Calli, B., Bargal, S.A., Saenko, K., 2022. ZeroWaste dataset: Towards deformable object segmentation in cluttered scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21147–21157.

Bawankule, R., Gaikwad, V., Kulkarni, I., Kulkarni, S., Jadhav, A., Ranjan, N., 2023. Visual detection of waste using YOLOv8. In: 2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS). IEEE, pp. 869–873.

Bircanoğlu, C., Atay, M., Beşer, F., Genç, Ö., Kızrak, M.A., 2018. RecycleNet: Intelligent waste sorting using deep neural networks. In: 2018 Innovations in Intelligent Systems and Applications (INISTA). IEEE, pp. 1–7.

Bobulski, J., Kubanek, M., 2021a. Deep learning for plastic waste classification system. Appl. Comput. Intell. Soft Comput. 2021.

Bobulski, J., Kubanek, M., 2021b. Deep learning for plastic waste classification system. Appl. Comput. Intell. Soft Comput. http://dx.doi.org/10.1155/2021/6626948, APET waste classification method and Plastic Waste DataBase WaDaBa.

Cen, A., 2020. Waste images from sushi restaurant. https://www.kaggle.com/datasets/arthurcen/waste-images-from-sushi-restaurant. Accessed: 2022-06-20.

Chen, Q., Xiong, Q., 2020. Garbage classification detection based on improved YOLOV4. J. Comput. Commun. 08, 285–294. http://dx.doi.org/10.4236/jcc.2020.812023.

Chu, Y., Huang, C., Xie, X., Tan, B., Kamal, S., Xiong, X., 2018. Multilayer hybrid deep-learning method for waste classification and recycling. Comput. Intell. Neurosci. 2018.

Datacluster-labs, 2021. Domestic trash dataset. https://github.com/datacluster-labs/Domestic-Trash-Dataset. Accessed: 2022-06-20.

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 Manage. 167, 194–203.

Ding, X., Zhang, X., Ma, N., Han, J., Ding, G., Sun, J., 2021. Repvgg: Making vgg-style convnets great again. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13733–13742.

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

Feng, C., Zhong, Y., Gao, Y., Scott, M.R., Huang, W., 2021. Tood: Task-aligned one-stage object detection. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE Computer Society, pp. 3490–3499.

Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J., 2021. Yolox: Exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430.

GeoTech Innovations, L., 2020. Open litter map. https://openlittermap.com/. Accessed: 2022-06-20.

Ghiasi, G., Cui, Y., Srinivas, A., Qian, R., Lin, T.-Y., Cubuk, E.D., Le, Q.V., Zoph, B., 2020. Simple copy-paste is a strong data augmentation method for instance segmentation. arXiv preprint arXiv:2012.07177.

Haamer, K., 2020. WADE dataset. https://github.com/letsdoitworld/wade-ai. Accessed: 2022-06-20.

Han, H., Fan, X., Li, F., 2023. Prototype enhancement-based incremental evolution learning for urban garbage classification. IEEE Trans. Artif. Intell..

Han, D., Yun, S., Heo, B., Yoo, Y., 2021. Rethinking channel dimensions for efficient model design. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 732–741.

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

Hoornweg, D., Bhada-Tata, P., Kennedy, C., 2013. Environment: Waste production must peak this century. Nature.

Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H., 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.

Huang, G.-L., He, J., Xu, Z., Huang, G., 2020. A combination model based on transfer learning for waste classification. Concurr. Comput.: Pract. Exper. 32 (19), e5751.

Huang, G., Liu, Z., Pleiss, G., Van Der Maaten, L., Weinberger, K., 2019. Convolutional networks with dense connectivity. IEEE Trans. Pattern Anal. Mach. Intell..

Jalwana, M.A., Akhtar, N., Bennamoun, M., Mian, A., 2021. CAMERAS: Enhanced resolution and sanity preserving class activation mapping for image saliency. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16327–16336.

Jiang, P.-T., Zhang, C.-B., Hou, Q., Cheng, M.-M., Wei, Y., 2021. Layercam: Exploring hierarchical class activation maps for localization. IEEE Trans. Image Process. 30, 5875–5888.

Koskinopoulou, M., Raptopoulos, F., Papadopoulos, G., Mavrakis, N., Maniadakis, M., 2021. Robotic waste sorting technology: Toward a vision-based categorization system for the industrial robotic separation of recyclable waste. IEEE Robot. Autom. Mag. 28 (2), 50–60.

Kraft, M., Piechocki, M., Ptak, B., Walas, K., 2021. Autonomous, onboard vision-based trash and litter detection in low altitude aerial images collected by an unmanned aerial vehicle. Remote Sens. 13 (5), http://dx.doi.org/10.3390/rs13050965, URL: https://www.mdpi.com/2072-4292/13/5/965.

Kuznetsov, V.I., Yudin, D.A., 2022. Neural networks for classification and unsupervised segmentation of visibility artifacts on monocular camera image. Opt. Mem. Neural Netw. (Inf. Opt.).

Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al., 2020. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. Int. J. Comput. Vis. 128 (7), 1956–1981.

Li, N., Chen, Y., 2023. Municipal solid waste classification and real-time detection using deep learning methods. Urban Clim. 49, 101462.

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

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., Xie, S., 2022. A ConvNet for the 2020s. arXiv preprint arXiv:2201.03545.

Malik, M., Sharma, S., Uddin, M., Chen, C.-L., Wu, C.-M., Soni, P., Chaudhary, S., 2022. Waste classification for sustainable development using image recognition with deep learning neural network models. Sustainability 14 (12), 7222.

Mao, W.-L., Chen, W.-C., Wang, C.-T., Lin, Y.-H., 2021. Recycling waste classification using optimized convolutional neural network. Resour. Conserv. Recy. 164, 105132.

Mikołajczyk, A., 2021. Litter dataset. https://github.com/AgaMiko/waste-datasets-review. Accessed: 2022-06-20.

Mittal, G., Yagnik, K.B., Garg, M., Krishnan, N.C., 2016. Garbage in images (GINI) dataset. https://github.com/spotgarbage/spotgarbage-GINI. Accessed: 2022-06-20.

Ni, D., Xiao, Z., Lim, M.K., 2021. Machine learning in recycling business: an investigation of its practicality, benefits and future trends. Soft Comput. 25 (12), 7907–7927.

Proença, P.F., Simões, P., 2020. TACO: Trash annotations in context for litter detection. arXiv preprint arXiv:2003.06975.

Redmon, J., Farhadi, A., 2018. YOLOv3: An incremental improvement. ArXiv abs/1804.02767.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. 39.

Sapal6, 2019. Waste classification data v2. https://www.kaggle.com/datasets/sapal6/waste-classification-data-v2. Accessed: 2022-06-20.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 618–626.

Serezhkin, A., 2020. Drinking waste classification. https://www.kaggle.com/datasets/arkadiyhacks/drinking-waste-classification. Accessed: 2022-06-20.

Sosa-García, J., Odone, F., 2017. "Hands on" visual recognition for visually impaired users. ACM Trans. Access. Comput. (TACCESS) 10 (3), 1–30.

Sreelakshmi, K., Akarsh, S., Vinayakumar, R., Soman, K., 2019. Capsule neural networks and visualization for segregation of plastic and non-plastic wastes. In: 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS). IEEE, pp. 631–636.

Sun, P., Zhang, R., Jiang, Y., Kong, T., Xu, C., Zhan, W., Tomizuka, M., Li, L., Yuan, Z., Wang, C., et al., 2021. Sparse r-cnn: End-to-end object detection with learnable proposals. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14454–14463.

Tan, M., Le, Q., 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning. PMLR, pp. 6105–6114.

Tang, Y., Qiu, J., Zhang, Y., Wu, D., Cao, Y., Zhao, K., Zhu, L., 2023a. Optimization strategies of fruit detection to overcome the challenge of unstructured background in field orchard environment: A review. Precis. Agric. 1–37.

Tang, Y., Zhou, H., Wang, H., Zhang, Y., 2023b. Fruit detection and positioning technology for a Camellia oleifera C. Abel orchard based on improved YOLOv4-tiny model and binocular stereo vision. Expert Syst. Appl. 211, 118573.

Terven, J., Cordova-Esparza, D., 2023. A comprehensive review of YOLO: From YOLOv1 to YOLOv8 and beyond. arXiv preprint arXiv:2304.00501.

Thung, G., Yang, M., 2017. Trashnet dataset. https://github.com/garythung/trashnet. Accessed: 2022-06-20.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H., 2021. Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning. PMLR, pp. 10347–10357.

Wang, Z., 2019. Waste_pictures. https://www.kaggle.com/datasets/wangziang/waste-pictures. Accessed: 2023-06-20.

Wang, T., Cai, Y., Liang, L., Ye, D., 2020. A multi-level approach to waste object segmentation. Sensors.

Wightman, R., 2019. PyTorch image models. https://github.com/rwightman/pytorch-image-models.

Wu, F., Duan, J., Ai, P., Chen, Z., Yang, Z., Zou, X., 2022. Rachis detection and three-dimensional localization of cut off point for vision-based banana robot. Comput. Electron. Agric. 198, 107079.

Wu, F., Yang, Z., Mo, X., Wu, Z., Tang, W., Duan, J., Zou, X., 2023. Detection and counting of banana bunches by integrating deep learning and classic image-processing algorithms. Comput. Electron. Agric. 209, 107827.

Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K., 2017. Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1492–1500.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P., 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. Adv. Neural Inf. Process. Syst. 34, 12077–12090.

Xie, J., Xiang, J., Chen, J., Hou, X., Zhao, X., Shen, L., 2022. C2AM: Contrastive learning of class-agnostic activation map for weakly supervised object localization and semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 989–998.

Yang, G., Jin, J., Lei, Q., Wang, Y., Zhou, J., Sun, Z., Li, X., Wang, W., 2021. Garbage classification system with YOLOV5 based on image recognition. In: 2021 IEEE 6th International Conference on Signal and Image Processing (ICSIP). IEEE, pp. 11–18.

Zhang, H., Chang, H., Ma, B., Wang, N., Chen, X., 2020. Dynamic R-CNN: Towards high quality object detection via dynamic training. In: European Conference on Computer Vision. Springer, pp. 260–275.

Zhang, Q., Yang, Q., Zhang, X., Bao, Q., Su, J., Liu, X., 2021a. Waste image classification based on transfer learning and convolutional neural network. Waste Manage. 135, 150–157.

Zhang, Q., Zhang, X., Mu, X., Wang, Z., Tian, R., Wang, X., Liu, X., 2021b. Recyclable waste image recognition based on deep learning. Resour. Conserv. Recy. 171, 105636.

Zhihong, C., Hebin, Z., Yanbo, W., Binyan, L., Yu, L., 2017. A vision-based robotic grasping system using deep learning for garbage sorting. In: 2017 36th Chinese Control Conference (CCC). IEEE, pp. 11223–11226.

Zhou, Y., Tang, Y., Zou, X., Wu, M., Tang, W., Meng, F., Zhang, Y., Kang, H., 2022. Adaptive active positioning of Camellia oleifera fruit picking points: Classical image processing and YOLOv7 fusion algorithm. Appl. Sci. 12 (24), 12959.

Zhou, X., Wang, D., Krähenbühl, P., 2019. Objects as points. arXiv preprint arXiv:1904.07850.

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

ZotBins, 2023. Waste recognition with TACO dataset. https://universe.roboflow.com/zotbins/waste-recognition-with-taco. visited on 2023-10-20.