



Skip-YOLO: Domestic Garbage Detection Using Deep Learning Method in Complex Multi-scenes

Zhao Lun¹ · Yunlong Pan^{1,2} · Sen Wang² · Zeshan Abbas¹ · Md Shafiqul Islam³ · Sufeng Yin⁴

Received: 5 May 2023 / Accepted: 11 August 2023
© The Author(s) 2023

Abstract

It is of great significance to identify all types of domestic garbage quickly and intelligently to improve people's quality of life. Based on the visual analysis of feature map changes in different neural networks, a Skip-YOLO model is proposed for real-life garbage detection, targeting the problem of recognizing garbage with similar features. First, the receptive field of the model is enlarged through the large-size convolution kernel which enhanced the shallow information of images. Second, the high-dimensional features of the garbage maps are extracted by dense convolutional blocks. The sensitivity of similar features in the same type of garbage increases by strengthening the sharing of shallow low semantics and deep high semantics information. Finally, multiscale high-dimensional feature maps are integrated and routed to the YOLO layer for predicting garbage type and location. The overall detection accuracy is increased by 22.5% and the average recall rate is increased by 18.6% comparing the experimental results with the YOLOv3 analysis. In qualitative comparison, it successfully detects domestic garbage in complex multi-scenes. In addition, this approach alleviates the overfitting problem of deep residual blocks. The application case of waste sorting production line is used to further highlight the model generalization performance of the method.

Keywords Image procession · Garbage detection · Feature mappings · Dense convolution block · YOLOv3

1 Introduction

The intelligent recycling of municipal solid waste can not only bring economic benefits, but also has research value in the fields of social research and natural science [1, 2]. The recycling system of domestic garbage can be simply divided into four stages: garbage collection, transportation, transfer, and recycling. A reasonable waste management process can produce social and economic benefits. For example, a survey

of rural areas in developing countries calls for increased density of garbage collection points, which can reduce environmental degradation from the source [3]. It can reduce investment costs and improve recycling efficiency by optimizing the placement of garbage bins in urban communities [4]. The daintiness and perceptibility of the trash can help the correct collection and classification of waste [5]. Converting the collection and transportation of municipal solid waste (MSW) into an integer program can provide the best solution for waste collection and transportation [6]. Multi-level solid waste collection with operating stations and transportation system with reliability framework can be used to manage the uncertainty of multi-level SWM systems [7]. It will help solid waste recycling by improving the harmless treatment method of domestic waste or extracting valuable organic matter from organic waste [8, 9]. It is important for domestic waste to process intelligently, which can further coordinate urban development and environmental protection [10]. The simulation of various policies on MSW management from a dynamic and complex perspective found that economic policies have a great impact on the future management of municipal solid waste [11]. Sensor-based Internet of Things

✉ Sen Wang
wangsen0401@126.com

¹ Institute of Intelligent Manufacturing Technology, Shenzhen Polytechnic, Shenzhen 518055, China

² Faculty of Mechanical and Electrical Engineering, Kunming University of Science and Technology, Kunming 650500, China

³ Department of Mechanical Engineering, Faculty of Engineering, Blekinge Institute of Technology, 37179 Karlskrona, Sweden

⁴ School of Mechanical Engineering, Guangdong Songshan Polytechnic, Shaoguan 512026, China

(IoT) can improve the generation, collection, transportation, and final disposal of food waste [12]. Non-autoregression (NAR) model can be used to predict the future generation of MSW [13]. In general, many scholars are searching for efficient waste disposal systems, which require efficient waste sorting methods.

With the rapid development of computer vision technology, deep learning methods have gradually attracted the attention of some researchers. For example, intelligent classification of glass and metal in garbage bags by training convolutional neural network (CNN) [14]. Optimize the fully connected layer of CNN through genetic algorithm (GA) can improve the performance of waste detection [15]. However, the full connection layer needs a lot of parametric optimizations. Therefore, it is difficult to realize fast and real-time recognition. Based on ResNet, an intelligent waste classification system can achieve accurate waste classification [16, 17]. However, it cannot identify the remains and locate the object. It can realize real-time monitoring of digested and non-digested waste to combine with IoT and CNN intelligent waste management system architecture. Therefore, it needs a lot of hardware support [18]. In addition, deep learning has been successfully applied in separation and classification of waste electrical and electronic equipment (WEEE) batteries [19], E-waste collection [20], construction solid waste classification [21], and automatic detection of waste in water [22]. The efficient and accurate detection of domestic waste will assist the intelligent development of waste treatment. The garbage collection robot that uses neural networks to identify garbage and the surface cleaning robot based on the YOLOv3 algorithm can both replace manual garbage collection [23–25]. However, the number of garbage detected is limited and the background information is single. It is different from real life.

There is a challenge to object detection performance, when the domestic waste is diverse and existence scene is uncertain. For example, disposable chopsticks and banana peels are non-recyclable garbage, but their attributes are greatly different. The problem of target occlusion can be solved by designing high-performance filters to extract high-quality feature maps [26, 27]. Based on the deep learning framework, the method of multi-level feature fusion performs well in target detail feature extraction [28, 29]. The skipping connection based on residual convolution module has significant advantages in image super-resolution reconstruction. In response to this problem, this paper proposes a method that combines the YOLOv3 with densely connected convolutional blocks [30, 31]. This method can be applied to domestic garbage detection in the multiple natural scenes or different quantity distributions. First, we analyzed different types of neural networks through the visualization of feature mappings during the process of network training. Second, a feature extractor is constructed using densely connected

convolutional blocks to obtain high-dimensional feature mappings. Finally, multi-scale high-dimensional feature mappings are merged and three different YOLO layers are engaged to predict various types of domestic waste. In addition, we provide an application case in the experimental part as a reference for intelligent waste management.

The main contributions of this paper are as follows:

- (1) This work propose a Skip-YOLO model for intelligent detection of domestic garbage based on the visual analysis of different convolution neural networks targeting the problem that domestic garbage with similar characteristics are difficult to recognize in complex real-life scene.
- (2) The dense convolution module is used to improve the backbone network, which enhance the quality of high-dimensional feature mapping, and effectively alleviate the over-fitting of the deep network.
- (3) The domestic garbage data set is established based on a variety of indoor and outdoor life scenes. Through a large number of comparative experiments, the accuracy and reliability of the proposed method in the process of identifying domestic garbage are proved.

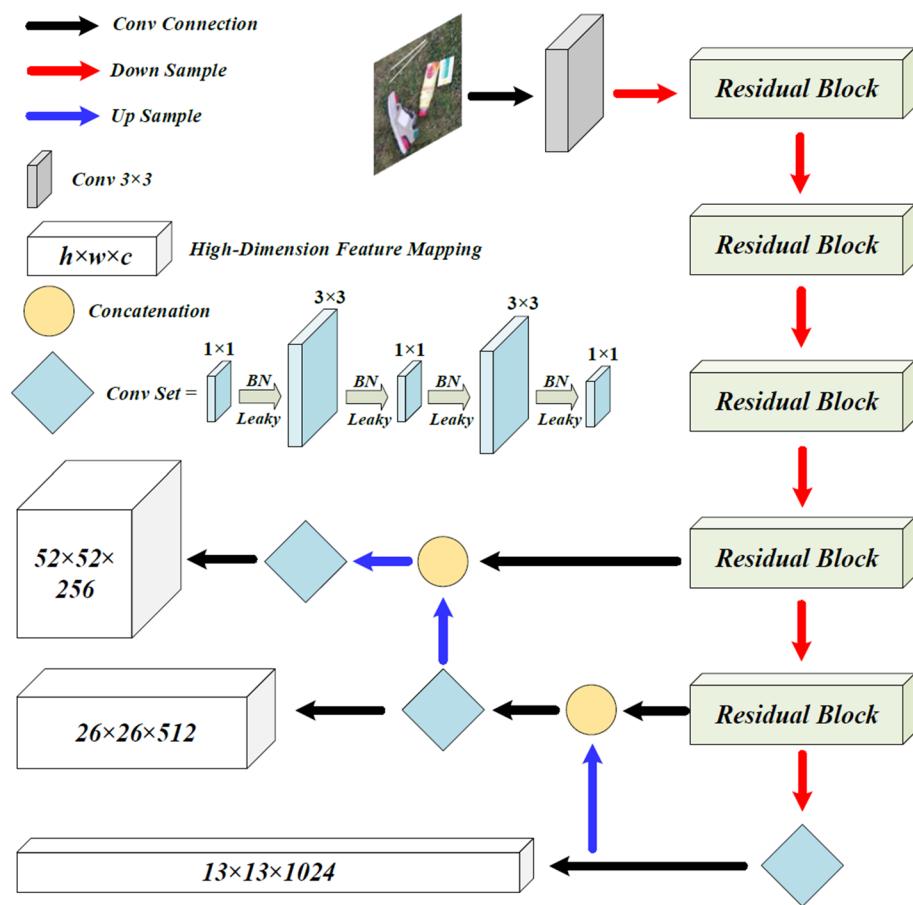
2 Deep Learning Methods

The different types of debris are the objects that need to be detected and usually need to train the current model before making predictions. The deep learning model is composed of four parts: Shallow network, Backbone network, Neck and Head. Among these, the Shallow network and the Backbone network are mainly responsible for extracting semantic information such as the shape, color and location from the input feature and converging them into high-dimensional feature mappings. The Neck can optimize the extracted high-dimensional feature mappings, which helps the Head to decode higher quality feature.

2.1 Object Detection Algorithm

YOLOv3 is a one-stage anchor-based object detection, which is mainly composed of the Darknet architecture, three convolution sets and three YOLO layers. The YOLOv3 algorithm is widely used in the fields of construction [32], agriculture [33] and transportation [34] etc. As shown in Fig. 1, Darknet is mainly composed of five residual blocks and several convolutional layers which are connected to the residual blocks. Convolution set alternately uses $1 \times 1 \times c$ and $3 \times 3 \times 2c$ convolutional layers ($1 \times 1 \times c$ means the size of filter kernel is 1×1 and dimension are c) to effectively extract and merge the mapping information. Among them, $1 \times 1 \times c$ convolutional layers can effectively compress the

Fig. 1 Structure of YOLOv3 algorithm



feature information which expanded by the previous convolution layer. $3 \times 3 \times 2c$ convolutional layers can expand the feature information and reduce the model calculation parameters without changing the scale of the input feature. The feature mapping will enter corresponding residual block to achieve multi-scale feature extraction after each down sample. Finally, three groups of high-dimension feature mappings with the scale of 52×52 , 26×26 and 13×13 , respectively, will be the output. YOLOv3 draws on the idea of multi-scale feature fusion in FPN [35]. It is fused with the corresponding feature convolutions of two different scales through up-sampling with 13×13 feature mappings. After multi-scale feature fusion, three YOLO layers are used for prediction and regression at the same time.

2.2 Related Convolution Block

The residual block is composed of multiple residual units for image feature extraction as the black dashed box is shown in Fig. 2. Each residual unit is composed of a 1×1 convolution kernel with k channels and a 3×3 convolution kernel. The input feature and the 3×3 convolution kernel are connected by residuals, which can continuously overlay input features of

the same dimension. The calculation of residual connection is as follows:

$$X_n = (Y_n + X_{n-1}) \quad (1)$$

where X_n represents the input feature of the n th layer, Y_n indicates the output feature of the $(n-1)$ th layer. Therefore, each residual unit will be affected by the output from the previous residual unit layer.

In contrast, the dense block is similar to an enhanced version of the residual block. Therefore, the each dense block consists of several dense units as the red dashed box is shown in Fig. 2. Therefore, each unit contains of a 1×1 convolution kernel with k channels and a 3×3 convolution kernel with $4k$ channels. The number of channels input for each dense block is k_0 . After n times of convolution stacking, the feature mapping with $k_0 + (n-1)k$ channels is final output. Among them, the non-linear function y_n needs to be obtained by the operation of batch normalization (BN) [36], ReLu [37] activation function and 3×3 convolutional layer in turn. The calculation between densely convolutional blocks is as follows:

$$X_n = Y_n([X_0, X_1, \dots, X_{n-1}]) \quad (2)$$

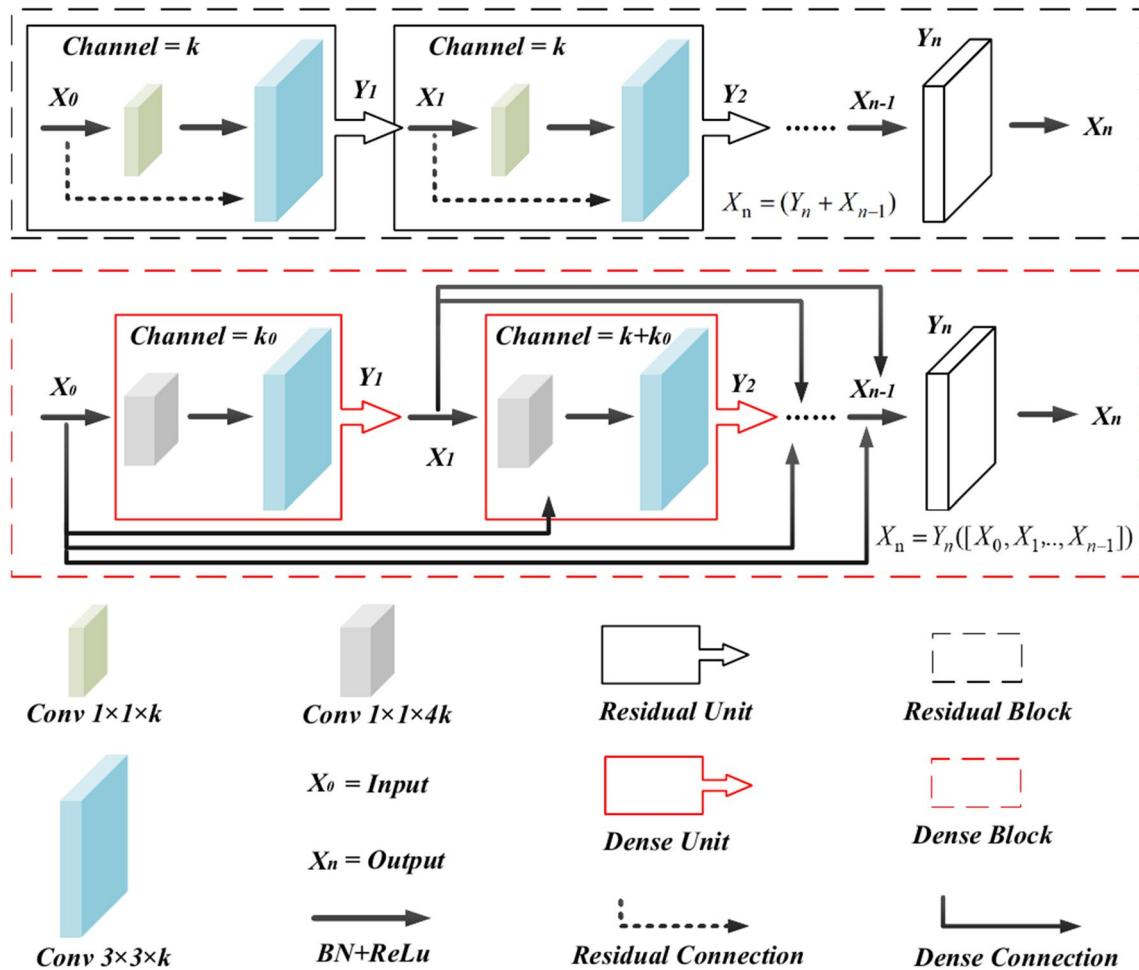


Fig. 2 Illustration of different convolution blocks

where $[X_0, X_1, \dots, X_{n-1}]$ represents the input mapping from layer 0 to layer $(n-1)$. More, X_n represents the output of n th layer. Y_n denotes the non-linear function of the output. Therefore, the output of each dense block layer is related to all previous input layers.

3 Methodology

3.1 Overview

The main work of this paper is shown in Fig. 3 which explains the entire features of the model. In the stage 1, indoor and outdoor garbage images have obtained in different scenarios, and divide all images into two types: single-class and multi-class. Then, all images are resized to \$416\times 416\$, and randomly allocate training data set, validation data set and test data set. In the stage 2, we analyzed the parameter transfer forms of three classic convolutional neural networks. The same type of any domestic garbage has

quite different characteristics. This makes garbage detection more complicated. Therefore, how to get the most important pixels between similar characteristic is the key of garbage detection. This paper will find an advantageous solution from the perspective of feature mapping. In the stage 3, we combined the analysis results of feature mapping to improve the backbone network of the YOLOv3. We conducted two different tests: one is based on the test data set, and the other is an application case that simulates a waste sorting production line. To test whether the proposed model has sufficient generalization ability, we replaced some untrained garbage in the production line application case.

3.2 Analysis of Feature Mapping

In the training of the deep learning model, the shallow network has rich feature information, such as edge contour, brightness, and color etc. However, the lack of sufficient receptive fields results in the limitation of shallow feature extraction. The deep network can not only express the global

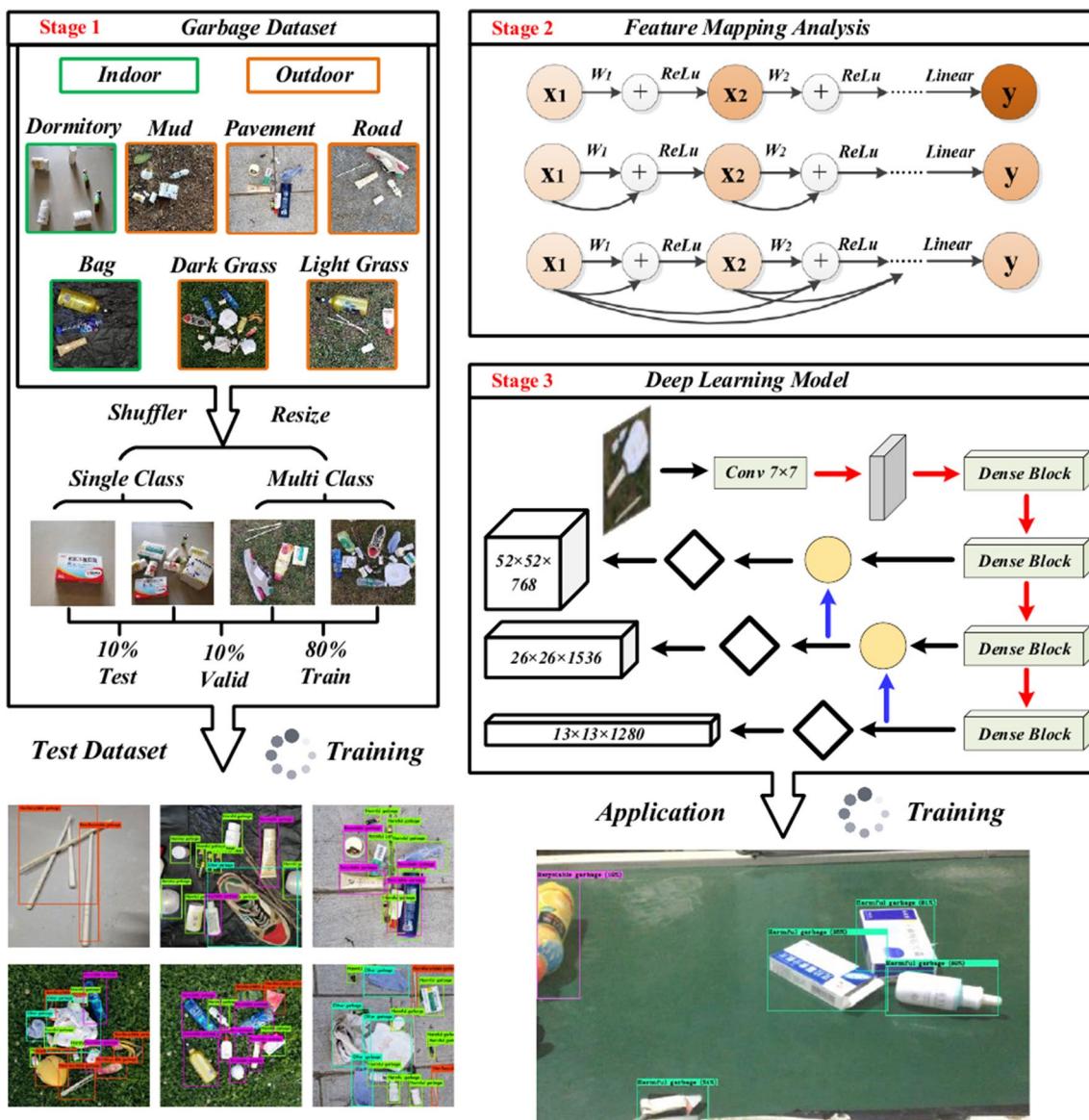


Fig. 3 Illustration of different convolution blocks. The illustration of main research works throughout this study. In the stage 1, a domestic waste data set is set, which contain simple and complex garbage object. In the stage 2, high-quality feature mappings have more important pixels. However, the quality of feature mapping is affected

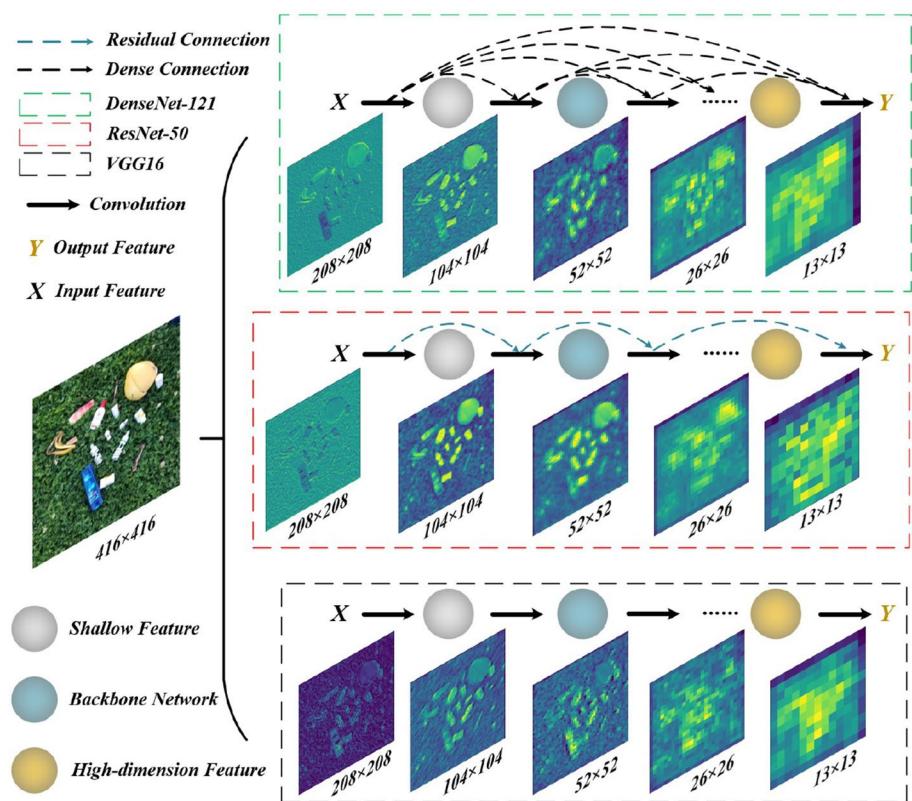
by the characteristics of the object itself. Therefore, we design an improved model for the feature extraction of domestic waste in stage 3. Finally, we evaluate the performance of this approach through the test data set and apply it to real scenarios

features of each object in the image, but also recognize the detailed information inside the object. With the deepening from the network depth, it is easy to produce network degradation. Since, the useful features will gradually become saturated. For example, only simple features such as rough outlines, colors, backgrounds, and shadows can be obtained when the shallow network detects expired drugs. In addition, the deep network focuses on semantic and detailed features, such as graphical information on the packaging.

The parameter transfer process of different networks is shown in Fig. 4. As shown by the black dashed box in

Fig. 4, linear transmission can reduce the impact of data fluctuations on the output with the plane neural network (such as VGG16 [38]) learning parameters. However, the continuous increase of network depth will also lead to the gradual saturation of useful features and gradually cause network degradation. Therefore, some key pixels are missing in the high-dimensional feature mapping. As shown by the red dashed box in Fig. 4, the output of the lower layer will be impacted by the input of the upper layer, which can generate more features mapping in the residual network. Therefore, the residual network is more sensitive to data

Fig. 4 Parameter transfers process of different networks. The characteristics of input image for visualization are quite different. The visualization image in the figure comes from a representative channel image in the feature mappings



fluctuations. This ability to use the features of the previous layer for identifying mapping solves the problem of network degradation. However, the data description of the residual network is prone to overfitting in the deep network, which will eventually affect the detection accuracy. The densely connected networks have been shown in the green dashed box in Fig. 4. The parameters of the upper layer can jump to the next layer at will, so that each layer of the densely connected network contains all the previous layer information when learning the parameters. Densely connected networks and residual networks use the features of the previous layer for mapping learning, but each layer of dense connected networks only learns less features. Therefore, it can be more flexible to choose the effective information that needs to be learned when the data fluctuates. This method can effectively alleviate deep network overfitting while reducing network redundancy.

3.3 Improved Model

YOLOv3 is significantly better than other neural networks in animals or people detection with the help of residual network structure and multi-scale feature fusion. However, there is a major feature difference between the domestic garbage data set and public data (such as ImageNet and PASCAL VOC etc.), which leads to poor detection performance in real application. In addition, the household garbage is

arranged in a mess, and the same kind of garbage contains many different objects. Therefore, the feature of the same class garbage is quite different, which produces a certain degree of data fluctuation. This data fluctuation makes the network overfitting during the deep residual network learning. The training results will lack sufficient generalization and ultimately affect the average accuracy. To solve the problem of over-fitting, we propose a Skip-YOLO model for the domestic garbage detection. This model uses dense blocks to extract high-dimensional feature maps and combines multi-feature fusion based on the YOLOv3 algorithm.

The Skip-YOLO model is shown in Fig. 5. First, a 7×7 convolution kernel is used to extract the original image information, which can enhance the receptive field of shallow feature. The sensitivity of the shallow network to the same garbage also can be improved. Second, the darknet framework is improved by the jump feature of dense blocks, which achieves the sharing of shallow and deep high-level semantic information. To improve the detailed information and the ability of the model expression, a deeper dense block is constructed to extract deep detailed features at different scales. Finally, the high-dimensional feature mappings of different scales are fused and sent to the YOLO layer to achieve end-to-end regression. Assuming that there are 4 convolution units in a dense block and each unit outputs m dimensional features, then each dense block will output $4m$ dimensional features. Among them, each layer of dense

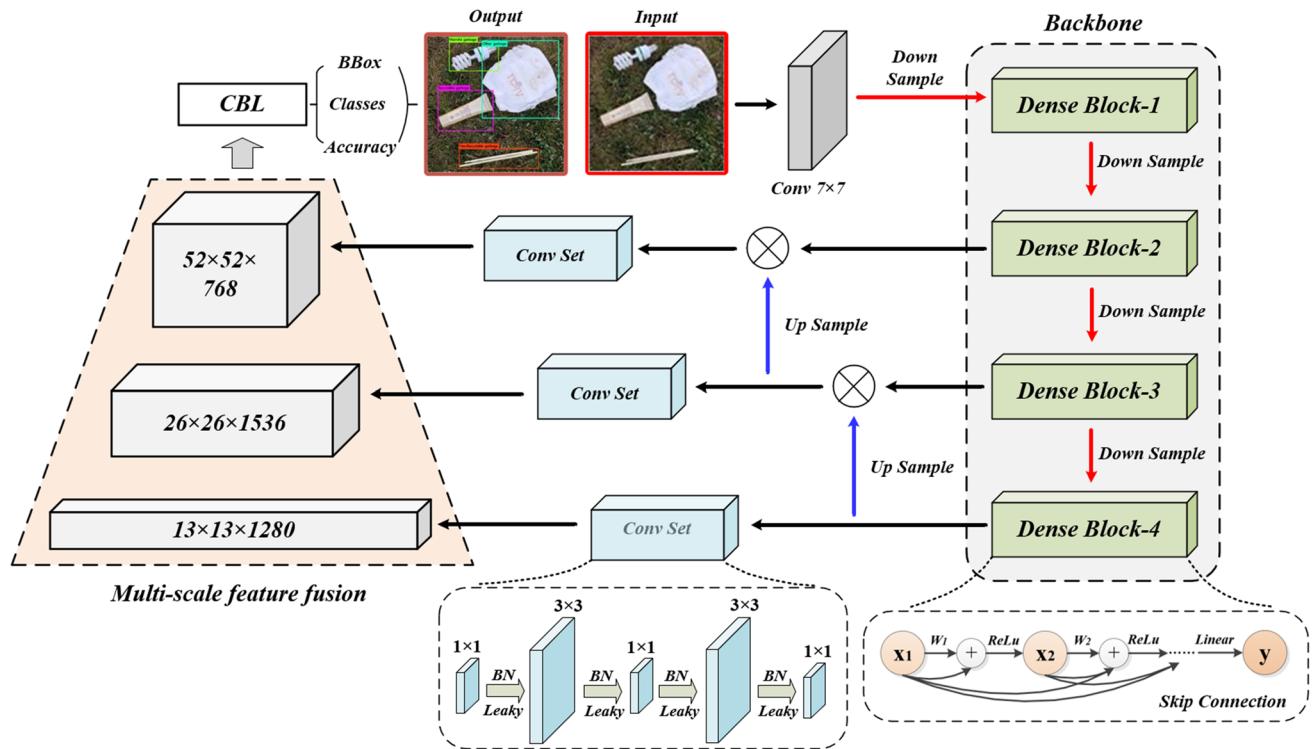


Fig. 5 Network structure of architecture

block learned fewer garbage features and the output information is relatively scattered. Therefore, the actual dimension of output feature mappings is much larger than the theoretical estimate. To make more effective use of the features extracted by dense blocks, this paper designs a CBL convolution block to compress scattered convolution information in the previous layer of the YOLO layer. CBL is consist of one 3×3 convolution layer and the number of convolution kernels is equal to the output feature convolution dimension. The BN operation and Leaky activation function will be performed after convolution.

4 Experiments and Analysis

4.1 Data Set

We shoot common domestic garbage images in the grass, pavement, dormitory and other scenes using the Nikon D5300 camera. The original size of each image is $3020 \times 3020 \times 3$ and the color channel are RGB. Among them, the outdoor background mainly includes dark grass, light grass, mud, concrete and streets. Likewise, the indoor background mainly includes dormitory and black garbage bag. In addition, this work divides garbage into recyclable garbage, non-recyclable garbage, harmful garbage, and other garbage. Therefore, current data set contains two or more kinds of

similar garbage. For example, harmful garbage includes batteries and medicine bottles etc. To better test the generalization ability of this approach, the sample distribution of the data set needs to be complicated. Therefore, this work data set contains 304 single-class images and 914 multi-class images. Among them, 80% of the data set images are used for training, 10% of the data set images are used for validation during the training process, and the remaining 10% of the data set images are used for testing after model training is completed. The results of the ablation experiment show that the deeper backbone increases the complexity of deep learn model, which helps to extract the characteristics of the garbage in different backgrounds. However, background information similar to garbage characteristics can easily be misidentified. The ability to distinguish background information and features can be improved by a suitable K value.

4.2 Implementation Details

All the experiments in this paper were performed on a laptop computer containing features, such as [Intel (R) Core (TM) i7-9700H CPU @ 2.6GHz, GPU GeForce RTX 1660Ti (6G), 16GB RAM, Windows10-64bit. Deep learning framework is Darknet with CUDA10.1 version and cudnn7.6.4 neural network acceleration library]. The size of all images is resizing to 416×416 before training. The training process uses multi-scale and iterative learning strategies. Among

them, all experiments set the initial learning rate to 0.001, the learning rate decay coefficient to 0.1, the weight decay coefficient to 0.0005, and the momentum to 0.9.

4.3 Evaluation Metrics

In the field of object detection, common model evaluation standards include different specifications, such as accuracy (P), recall rate (R), $F1$ score, average accuracy (AP , mAP , etc.). The calculation formula is as follows:

$$P = \frac{TP}{TP + FP} \times 100\% \quad (3)$$

$$R = \frac{TP}{TP + FN} \times 100\% \quad (4)$$

$$F1 = 2 \times \frac{P \times R}{P + R} \quad (5)$$

where TP is the correct number of positive samples predicted, FP is the number negative samples mistaken as positive samples. FN is the number of positive samples mistaken as negative samples. Both AP and mAP can represent the average accuracy of the model. The difference is that AP can measure the performance of the model in some category, while mAP can measure the overall performance of the model. In practical applications, AP_{50} , AP_{75} and other indicators are commonly used to evaluate model performance. For example, AP_{50} represents the detection accuracy of the model when the IOU threshold is 50%. The relation between AP and mAP is as follows:

$$mAP = \frac{1}{N} \sum_{i=1}^n AP_i \quad (i = 1, 2, \dots, n) \quad (6)$$

4.4 Ablation Study

This work explores the performance of the Skip-YOLO through the ablation experiments. The backbone of Skip-YOLO consists of four different dense blocks, which can set a different growth rate K . At the same scale, the dense hop

network can stack the input characteristics in order on the channel dimension. The influence of the amount of information contributed by different feature stacks on the model is further analyzed through appropriate K value. Therefore, the ablation experiment in this article first obtains three different models by adjusting the backbone depth. Second, based on the backbone with the best performance, other different models are obtained by adjusting the growth rate K . Finally, we analyzed the qualitative and quantitative comparison results of the experiments.

The results of the quantitative comparison are shown in Table 1 and Fig. 6. It can be seen that the Skip-YOLO-0 with the deepest layer has an AP_{50} slightly lower than the Skip-YOLO-4 by 0.26%. However, the mAP_{50} reached 90.38%, indicating that a deeper backbone is good for fitting complex data in complex background. From Fig. 6a, c, it is found that the average accuracy of the Skip-YOLO-0 is steadily increasing and the loss function fluctuation is small. It can be seen from Fig. 6b, d that the changes of accuracy curves after adjusting the K value are not much different. When the K value is 32, the change of the loss curve is the most stable. The K value has a more significant impact on model complexity compared to the depth of backbone.

As shown in the first line of Fig. 7, Skip-YOLO-0 with the deepest network layer and Skip-YOLO-4 with the maximum K value both detect the white light background as recyclable garbage and hazardous garbage. Thus, both types of garbage have some characteristics similar to white light. Therefore, when the model has a deeper backbone or a larger K value, its ability to express detailed information will be strong. However, irrelevant background information may be error detected as the object. As the results shown in the third row of Fig. 7, when the Skip-YOLO-0, 1 and 2 models detect a large single object, the shallower backbone has poor regression performance. Among them, shallowest network (Skip-YOLO-1) even has a false detection. Comparing Skip-YOLO-0, 3 and 4, it can also find that when the K value decreases, the prediction box becomes smaller. However, when the value of K increases, the model can eventually learn more features. It is possible to mistake the redundant background as the feature of the detection object. As shown in the second row of Fig. 7, the false and missed detection

Table 1 Experiment results of ablation study

Models	Backbone	K	AP ₅₀ (%)	mAP ₅₀ (%)	BFLOPS	Param (M)
Skip-YOLO-0	DenseBlock (6–12–48–32)	32	93.79	90.38	40.425	26.32
Skip-YOLO-1	DenseBlock (6–12–24–16)	32	92.80	89.64	32.694	21.13
Skip-YOLO-2	DenseBlock (6–12–32–32)	32	93.22	89.53	36.271	23.42
Skip-YOLO-3	DenseBlock (6–12–48–32)	24	93.69	89.83	29.948	26.32
Skip-YOLO-4	DenseBlock (6–12–48–32)	40	94.05	89.67	53.362	26.32

(6–12–48–32) represent the number of each dense block layers are 6, 12, 48 and 32

BFLOPS is calculation complexity of model

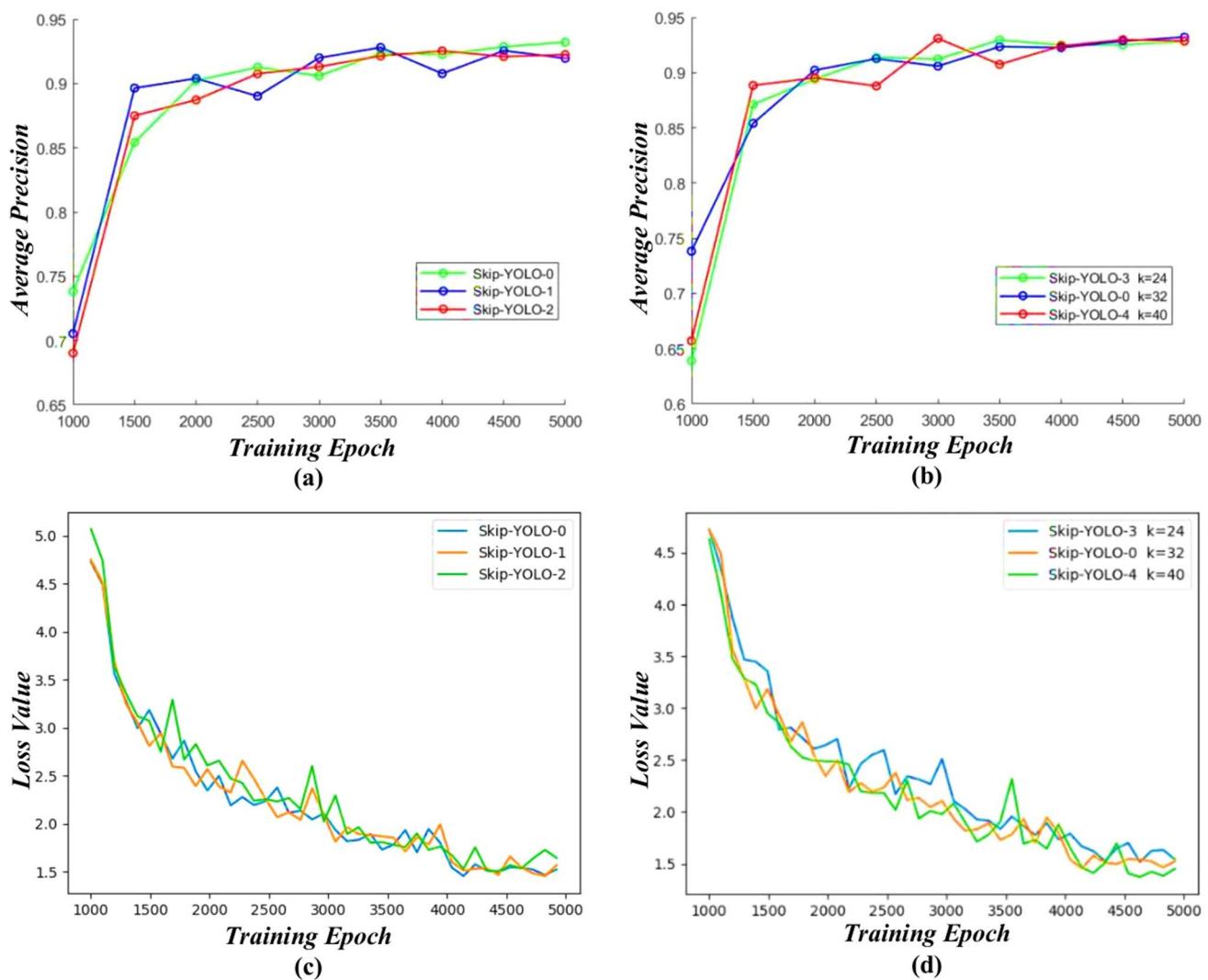


Fig. 6 Quantitative comparison of ablation study

of Skip-YOLO-1 are obviously if the object has centralized distribution. When the number of network layers decreases, the probability of false detection will be increased. However, the Skip-YOLO-3 has the best detection effect, because the K value is the smallest. The fourth row of Fig. 7 has evenly object distributed. Compared with Skip-YOLO-0, 1 and 2, it can be found that the Skip-YOLO-1 has no false detection. Comparing Skip-YOLO-0, 3 and 4 can also find that reducing the K value can effectively improve the detection performance in the deep network.

4.5 Analysis of Different Models

As presented in Table 2, the average accuracy of $AP-N$ and $AP-O$ is lower. The reason is that the characteristics of similar garbage are very different, which makes the data fluctuate greatly during the training process, and

ultimately affects the detection accuracy of the model. In addition, the improved model is much deeper than the YOLOv3, which makes the detection more time-consuming, but the detection accuracy is better. It can also be seen from Fig. 8 that the classification and recognition ability of this approach is significantly better than YOLOv3. Comparing the first column and the second column of Fig. 9, it is found that YOLOv3 mistakenly detects hazardous waste (expired drugs, batteries, etc.) as other waste in the single-category detection. When the object is small and the position is relatively scattered, the redundant background will be wrongly detected as the garbage. Comparing the third column and the fourth column of Fig. 9 shows that the YOLOv3 has serious missed detection. Although current approach has some error detection, the improvement of generalization performance makes more targets successfully identified.

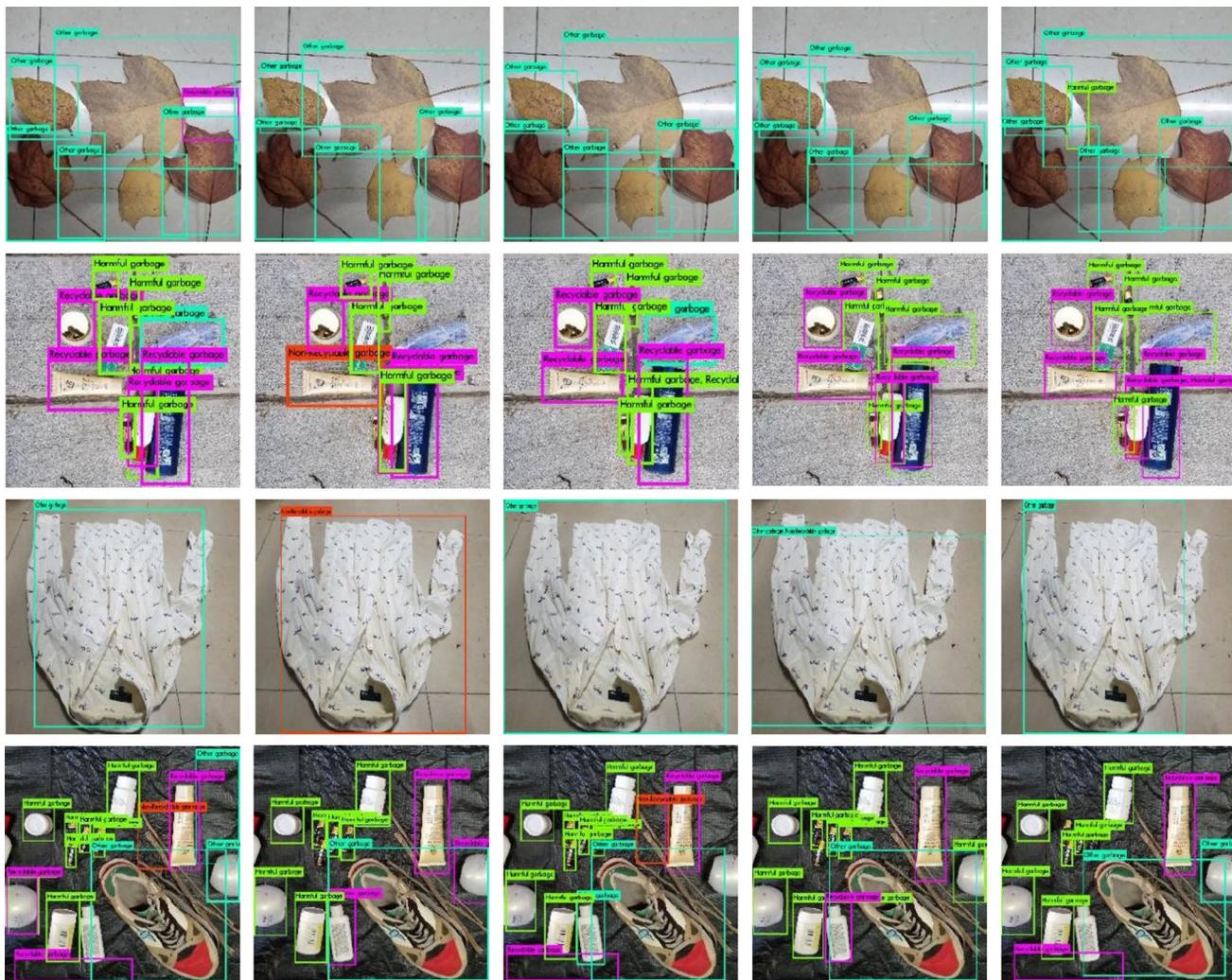


Fig. 7 Qualitative represents comparison of ablation study. From left to right: Skip-YOLO-0, Skip-YOLO-1, Skip-YOLO-2, Skip-YOLO-3, and Skip-YOLO-4

Table 2 Quantitative displays comparison of different models

Models	AP-R (%)	AP-N (%)	AP-H (%)	AP-O (%)	mAP50 (%)	mRecall	T/ms
YOLOv3	91.12	56.32	90.32	54.93	73.17	75.73%	163.85
Skip-YOLO-0	96.07	81.48	95.20	88.77	90.38	89.81%	286.68
Skip-YOLO-4	95.68	79.49	96.89	86.61	89.67	88.72	303.41

The recyclable garbage, non-recyclable garbage, harmful garbage, and other garbage are denoted by R, N, H, and O, respectively. In addition, average accuracy is represented by AP-R, AP-N, AP-H and AP-O, respectively

The comparison of generalization performance is shown in Fig. 10. The detection effects of YOLOv3 and this approach both perform relatively well when the trained images are selected for testing. However, the YOLOv3 has different degrees of error or missed detection when using untrained images for the same testing. For example, in the second column, other garbage is detected as recyclable

garbage. Based on the confusion matrix generated from the untrained image, it can be seen that this approach has a better recognition effect on various types of garbage. Because dense block can selectively learn a small amount of effective information when fitting a data set, which can effectively alleviating data fluctuations and improve the generalization performance.

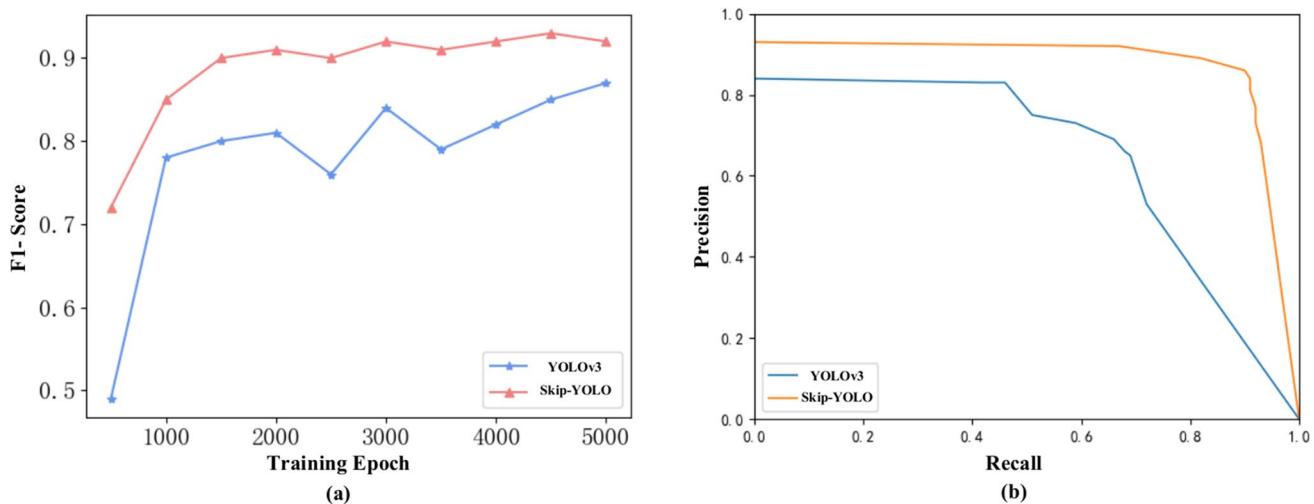


Fig. 8 Quantitative shows comparison of different models. **a** F1-Score curves. It is a common indicator for classification problems. **b** PR curves. The curve with good performance will completely cover the curve with poor performance

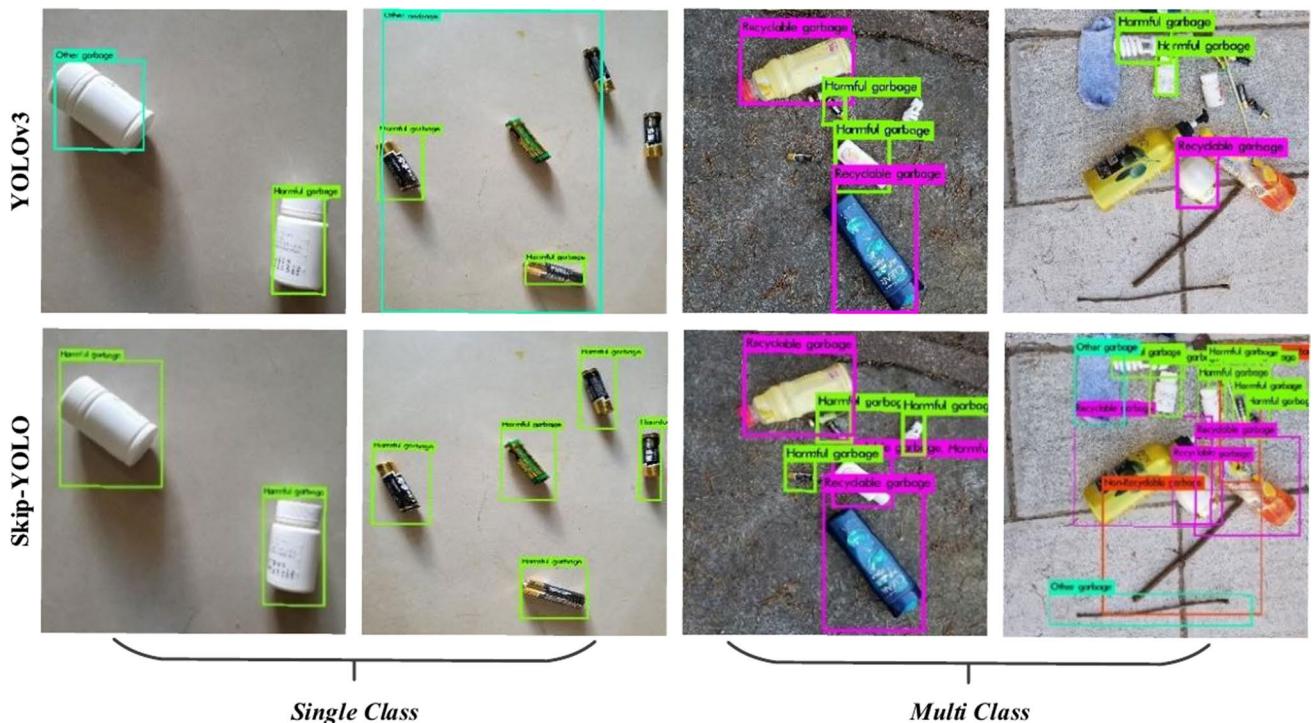


Fig. 9 Qualitative demonstrates comparison of different models

4.6 Comparison of Current Advance Detection Methods

In this paper, the method proposed (Skip-YOLO-0) in the present study is compared with current models to show the importance and advantages, and the model performance is compared in Table 3. Among them, FasterDet is a

lightweight new model that combines ShuffleNetV2 [37] and multi-scale fusion. DETR is an object detection framework that uses Transformer to encode and decode. YOLOv5 and YOLOv7 are the current advanced detection method of the YOLO series. All experiments use the same hyperparameters and train on the same data set. When the loss value iteration converges, the training is terminated. From the comparison

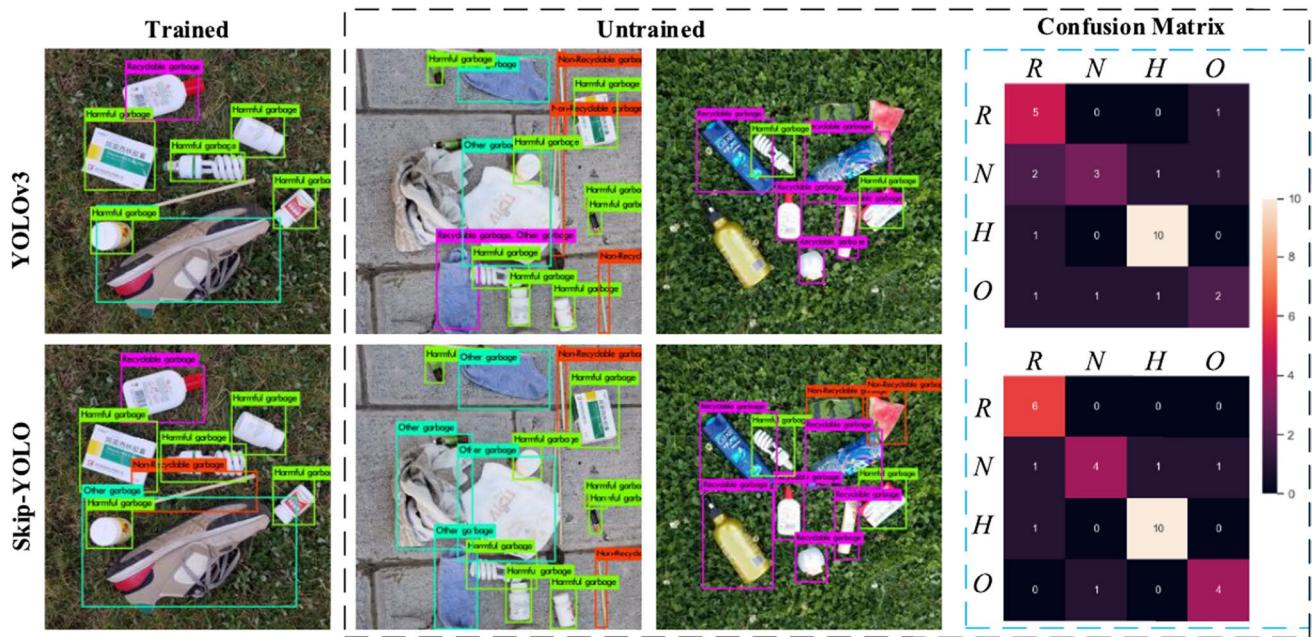


Fig. 10 Generalization performance comparison of different models. From top to bottom: YOLOv3, Skip-YOLO

Table 3 Quantitative comparison of current models

Models	mAP50	mRecall	BFLOPS	# Param (M)	T/ms
YOLOv3 [24]	73.17%	75.73%	65.304	58.67	163.85
DETR [39]	84.70%	74.91%	57.43	39.37	233.10
FasterDet [40]	70.30%	52.90%	4.60	1.31	10.95
YOLOv5	81.50%	76.60%	108.2	21.2	46.1
YOLOv5-s	68.90%	62.70%	15.9	6.69	6.9
YOLOv5-n	57.21%	56.43%	4.20	1.68	5.8
YOLOv7 [41]	83.30%	79.70%	103.2	36.9	201.6
YOLOv7-X	80.80%	70.20%	188.0	35.47	87.4
YOLOv7-tiny [42]	78.70%	71.50%	13.0	5.74	17.7
Faster-RCNN [43]	86.35	80.72	252.66	136.75	82.64
Skip-YOLO-0	90.38%	89.81%	53.362	26.32	286.68

results in Table 3, YOLOv5-n achieves the fastest detection speed of 5.8ms per image, but its detection accuracy is only 57.21%. The DETR, YOLOv5, YOLOv7 and Faster-RCNN achieve good detection performance of 84.70%, 81.50%, 83.30% and 86.35%, respectively, with a large of BFLOPS. Compared with other advanced algorithms, the proposed method (90.38% mAP50) achieves the highest detection

accuracy, exceeding YOLOv3 (73.17% mAP50) by 17.21%, YOLOv5-s (68.90% mAP50) by 21.48%, YOLOv7-X (80.80% mAP50) by 9.58%, YOLOv7-tiny (78.70% mAP50) by 11.68% and it significantly exceeded the fastest model FasterDet (70.30% mAP50) by 20.08%.

FasterDet, DETR, YOLOv5, and Skip-YOLO-0 are selected for qualitative comparison in this study. As shown in the first line of Fig. 11, the FasterDet lightweight model has a lot of error recognition due to the limited features extracted. The DETR using Transformer for end-to-end prediction has the best recognition effect, and other methods perform well. However, both the lightweight model and the transformer-based method have partially missed detection can be observed from the second line of Fig. 11. Besides, the YOLO-based method performs well in dense object detection in this case of study. From the third line of Fig. 11, the Transformer is not friendly for large object detection, while Skip-YOLO has the best recognition effect. Other methods have different degrees of error recognition. Compared with the fourth line in Fig. 11, when the object is relatively scattered, all methods have a good recognition effect. The proposed method performs well in large object recognition and dense multi-object classification. Therefore, the proposed method is more applicable and robust in real scenarios.

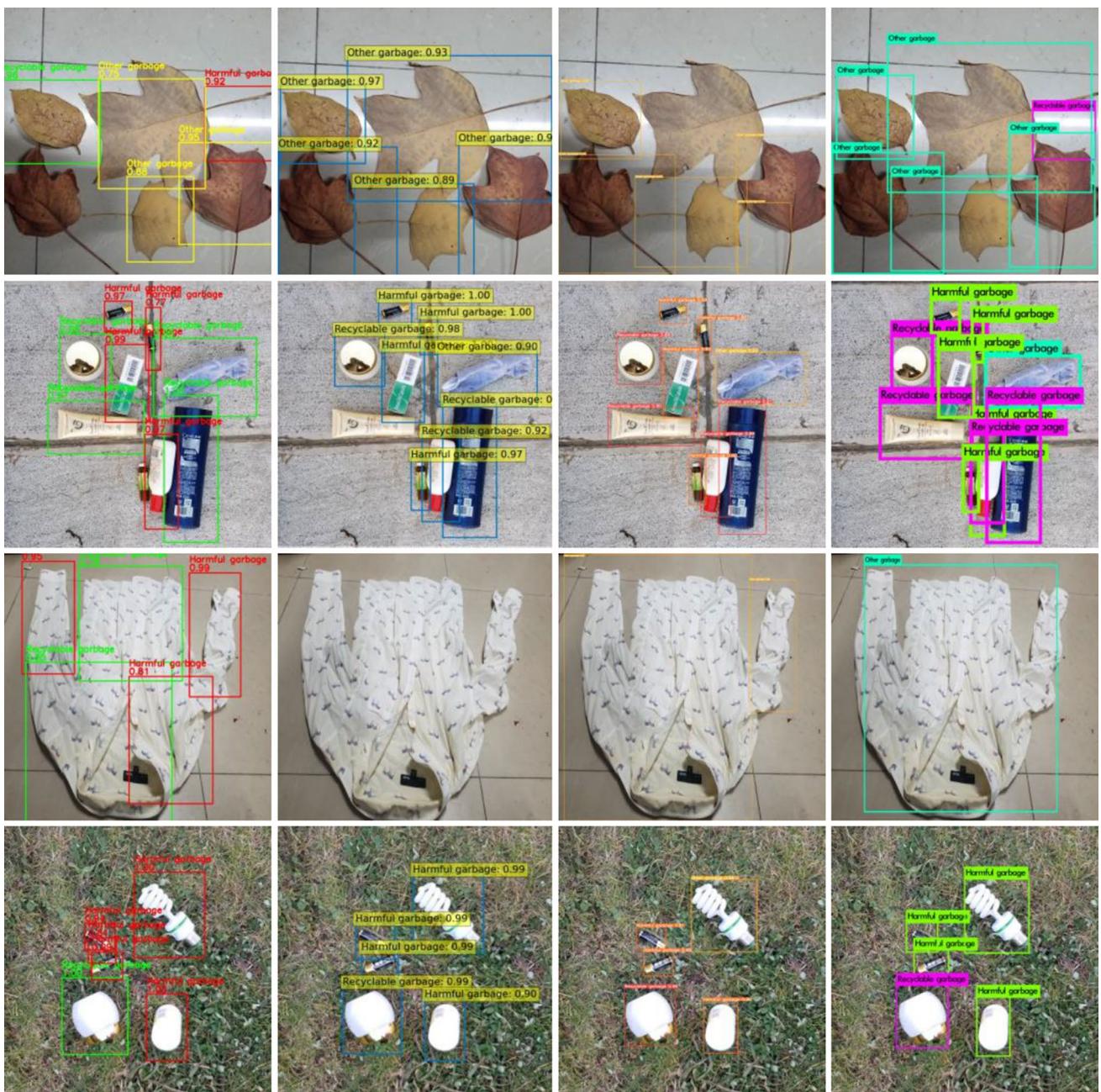


Fig. 11 Comparison of different object detection methods. From left to right: FasterDet, DETR, YOLOv5, Skip-YOLO-0

5 Instance of Application

As shown in Fig. 11, a common industrial camera is used to detect the moving domestic waste on the conveyor belt. It is also used to simulate the different application for the waste sorting production line. Part of the detection effect is shown in Fig. 12. Although some objects and the backgrounds have not been trained, the deep learning method still has a certain

detection effect. Among them, YOLOv3 and FasterDet mistake the background as other garbage, and the error detection is more serious. In contrast, Skip-YOLO is not affected by unfamiliar backgrounds, although there are some missed detections and false detections. DETR also has some false detections and missed detection. Figure 13 shows part of the test results, where it explains the sequence from top to bottom as Skip-YOLO, YOLOv3, FasterDet and DETR.



Fig. 12 Simulation work bench

6 Conclusion

This paper reports a Skip-YOLO model for the intelligent detection of domestic waste by aiming at the problems of low similarity of domestic waste characteristics and complex scenes. First, this paper visualizes the feature mappings in different neural networks. Second, the backbone network has been improved by dense blocks, which

helps to extract high-quality high-dimensional feature mappings and suppress deep network overfitting. Finally, high-dimensional feature mappings of different scales are fused and garbage detection is completed through the YOLO layer. Through ablation experiments, it is found that a deeper backbone has stronger ability of feature expression. However, there is a risk of mis-checking the redundant background at the same time. Therefore, setting a reasonable growth rate of dense blocks can prevent excessive learning of background features and control the size of bounding box. The experimental results indicate compared to the results obtained by the YOLOv3 model that this approach increases the mAP50 by 22.5% and the average recall rate increases by 18.6%. Among them, the precision of non-recyclable garbage and other garbage reached 81.48% and 88.77%, respectively. The qualitative experiments and the results of waste sorting production line are well-performed during this approach.

The proposed method also has set on the following improvements in the future work. For example, the essence of parameter jump is to further optimize the back propagation path of neural network. Therefore, deep learning strategy optimization can also be achieved by connecting different network parts or embedding similar attention mechanisms. In addition, complex data set labeling is time-consuming, which makes it necessary to develop an intelligent



Fig. 13 Part of the test results. From top to bottom: Skip-YOLO, YOLOv3, FasterDet, DETR

Waste sorting recognition algorithm based on unsupervised and weak Supervised learning.

Authors' Contributions Conceptualization, SW and MSI; data curation, ZL, YP and SY; formal analysis, ZL, SW and ZA; funding acquisition, SW; investigation, SY; methodology, ZL, YP and SW; project administration, SW; resources, ZL; supervision, SW; validation, ZL, YP and SY; visualization, ZA; writing—original draft, ZL and YP; writing—review and editing, ZA and MSI.

Funding The research is supported by Scientific Research Startup Fund for Shenzhen High-Caliber Personnel of SZPT (No.6022310046K) and National Natural Science Foundation of China (No.12104324), the National Natural Science Foundation of China (52065035) and Characteristic innovation project of the Department of Education of Guangdong Province (2022KTSCX261).

Data Availability Not applicable.

Declarations

Conflict of Interest The authors declare that they have no competing financial interests.

Informed Consent Not applicable.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Kutty, A.A., Wakjira, T.G., Kucukvar, M., Abdella, G.M., Onat, N.C.: Urban resilience and livability performance of European smart cities: a novel machine learning approach. *J. Clean Prod.* **378**, 134203 (2022). <https://doi.org/10.1016/j.jclepro.2022.134203>
- Rossit, D.G., Toutouh, J., Nesmachnow, S.: Exact and heuristic approaches for multi-objective garbage accumulation points location in real scenarios. *Waste Manage.* **105**, 467–481 (2020). <https://doi.org/10.1016/j.wasman.2020.02.016>
- Leeabai, N., Areeprasert, C., Khaobang, C., Viriyapanitchakij, N., Bussa, B., Dilinazi, D., Takahashi, F.: The effects of color preference and noticeability of trash bins on waste collection performance and waste-sorting behaviors. *Waste Manage.* **121**, 153–163 (2021). <https://doi.org/10.1016/j.wasman.2020.12.010>
- Das, S., Bhattacharyya, B.K.: Optimization of municipal solid waste collection and transportation routes. *Waste Manage.* **43**, 9–18 (2015). <https://doi.org/10.1016/j.wasman.2015.06.033>
- Cheng, C., Zhu, R., Thompson, R.G., Zhang, L.: Reliability analysis for multiple-stage solid waste management systems. *Waste Manage.* **120**, 650–658 (2021). <https://doi.org/10.1016/j.wasman.2020.10.035>
- Long Zhao, H., Liu, F., Liu, H.Q., Wang, L., Zhang, R., Hao, Y.: Comparative life cycle assessment of two ceramsite production technologies for reusing municipal solid waste incinerator fly ash in China. *Waste Manag.* **113**, 447–455 (2020). <https://doi.org/10.1016/j.wasman.2020.06.016>
- Alibardi, L., Astrup, T.F., Asunis, F., Clarke, W.P., De Gioannis, G., Dessì, P., Lens, P.N.L., Lavagnolo, M.C., Lombardi, L., Muntoni, A., Pivato, A., Polettini, A., Pomi, R., Rossi, A., Spagni, A., Spiga, D.: Organic waste biorefineries: Looking towards implementation. *Waste Manage.* **114**, 274–286 (2020). <https://doi.org/10.1016/j.wasman.2020.07.010>
- Xu, F., Huang, Q., Yue, H., He, C., Wang, C., Zhang, H.: Reexamining the relationship between urbanization and pollutant emissions in China based on the STIRPAT model. *J. Environ. Manag.* **273**, 111134 (2020). <https://doi.org/10.1016/j.jenvman.2020.111134>
- Xiao, S., Dong, H., Geng, Y., Tian, X., Liu, C., Li, H.: Policy impacts on municipal solid waste management in Shanghai: a system dynamics model analysis. *J. Clean Prod.* **262**, 121366 (2020). <https://doi.org/10.1016/j.jclepro.2020.121366>
- Wen, Z., Hu, S., De Clercq, D., Beck, M.B., Zhang, H., Zhang, H., Fei, F., Liu, J.: Design, implementation, and evaluation of an Internet of Things (IoT) network system for restaurant food waste management. *Waste Manage.* **73**, 26–38 (2018). <https://doi.org/10.1016/j.wasman.2017.11.054>
- Sunayana, S., Kumar, R.: Kumar, forecasting of municipal solid waste generation using non-linear autoregressive (NAR) neural models. *Waste Manage.* **121**, 206–214 (2021). <https://doi.org/10.1016/j.wasman.2020.12.011>
- Funch, O.I., Marhaug, R., Kohtala, S., Steinert, M.: Detecting glass and metal in consumer trash bags during waste collection using convolutional neural networks. *Waste Manage.* **119**, 30–38 (2021). <https://doi.org/10.1016/j.wasman.2020.09.032>
- Mao, W.L., Chen, W.C., Wang, C.T., Lin, Y.H.: Recycling waste classification using optimized convolutional neural network. *Res. Conservat. Recycl.* **164**, 105132 (2021). <https://doi.org/10.1016/j.resconrec.2020.105132>
- He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2016: 770–778 (2016). Doi: <https://doi.org/10.1109/CVPR.2016.90>.
- Adedeji, O., Wang, Z.: Intelligent waste classification system using deep learning convolutional neural network. *Procedia Manufact.* **35**, 607–612 (2019). <https://doi.org/10.1016/j.promfg.2019.05.086>
- Rahman, M.W., Islam, R., Hasan, A., Bithi, N.I., Hasan, M.M., Rahman, M.M.: Intelligent waste management system using deep learning with IoT. *J. King Saud. Univer-Comp. Inform. Sci.* **34**, 2072–2087 (2022). <https://doi.org/10.1016/j.jksuci.2020.08.016>
- Sterkens, W., Diaz-Romero, D., Goedemé, T., Dewulf, W., Peeters, J.R.: Detection and recognition of batteries on X-Ray images of waste electrical and electronic equipment using deep learning. *Res. Conservat Rec.* (2021). <https://doi.org/10.1016/j.resconrec.2020.105246>
- Nowakowski, P., Pamuła, T.: Application of deep learning object classifier to improve e-waste collection planning. *Waste Manage.* **109**, 1–9 (2020). <https://doi.org/10.1016/j.wasman.2020.04.041>
- Davis, P., Aziz, F., Newaz, M.T., Sher, W., Simon, L.: The classification of construction waste material using a deep convolutional neural network. *Automat. Const.* **122**, 103481 (2021). <https://doi.org/10.1016/j.autcon.2020.103481>
- Panwar, H., Gupta, P.K., Siddiqui, M.K., Morales-Menendez, R., Bhardwaj, P., Sharma, S., Sarker, I.H.: Aquavision: automating the

- detection of waste in water bodies using deep transfer learning. *Case Stud. Chem. Environ. Eng.* **2**, 100026 (2020). <https://doi.org/10.1016/j.cscee.2020.100026>
- 21. Bai, J., Lian, S., Liu, Z., Wang, K., Liu, D.: Deep Learning Based Robot for Automatically Picking Up Garbage on the Grass. *IEEE Trans. Consum. Electron.* **64**, 382–389 (2018). <https://doi.org/10.1109/TCE.2018.2859629>
 - 22. Redmon, J., Farhadi, A.: YOLOv3: An Incremental Improvement, (2018). <http://arxiv.org/abs/1804.02767>.
 - 23. Kong, S., Tian, M., Qiu, C., Wu, Z., Yu, J.: IWSCR: an intelligent water surface cleaner robot for collecting floating garbage. *IEEE Trans. Syst. Man. Cybernet. Syst.* **51**, 6358–6368 (2021). <https://doi.org/10.1109/TSMC.2019.2961687>
 - 24. Xia, R., Chen, Y., Ren, B.: Improved anti-occlusion object tracking algorithm using Unscented Rauch-Tung-Striebel smoother and kernel correlation filter. *J. King Saud. Univer. Comp. Inform. Sci.* **34**, 6008–6018 (2022). <https://doi.org/10.1016/j.jksuci.2022.02.004>
 - 25. Zhang, J., Feng, W., Yuan, T., Wang, J., Sangaiah, A.K.: SCSTCF: Spatial-channel selection and temporal regularized correlation filters for visual tracking. *Appl. Soft. Comp.* **118**, 108485 (2022). <https://doi.org/10.1016/j.asoc.2022.108485>
 - 26. Chen, Y., Xia, R., Zou, K., Yang, K.: FFTI: Image inpainting algorithm via features fusion and two-steps inpainting. *J. Vis. Commun. Image Rep.* **91**, 103776 (2023). <https://doi.org/10.1016/j.jvcir.2023.103776>
 - 27. Chen, Y., Tao, J., Liu, L., Xiong, J., Xia, R., Xie, J., Zhang, Q., Yang, K.: Research of improving semantic image segmentation based on a feature fusion model. *J Amb Intell Human Comp* **13**, 5033–5045 (2022). <https://doi.org/10.1007/s12652-020-02066-z>
 - 28. Zhang, X., Zhang, X., Zhao, L., Jiang, R., Huang, P., Xu, J.: Multi-level Feature Fusion Network for Single Image Super-Resolution. In: Proceedings-2020 IEEE International Conference on Big Data, Big Data 2020: 3361–3368 (2020). Doi: <https://doi.org/10.1109/BigData50022.2020.9377776>
 - 29. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks, Proceedings-30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017. 2017: 2261–2269 (2017). Doi: <https://doi.org/10.1109/CVPR.2017.243>.
 - 30. Li, Y., Lu, Y., Chen, J.: A deep learning approach for real-time rebar counting on the construction site based on YOLOv3 detector. *Automat. Const.* **124**, 103602 (2021). <https://doi.org/10.1016/j.autcon.2021.103602>
 - 31. Wu, D., Wu, Q., Yin, X., Jiang, B., Wang, H., He, D., Song, H.: Lameness detection of dairy cows based on the YOLOv3 deep learning algorithm and a relative step size characteristic vector. *Biosys. Eng.* **189**, 150–163 (2020). <https://doi.org/10.1016/j.biosystemseng.2019.11.017>
 - 32. Sri Jamiya, S., Esthar Rani, P.: LittleYOLO-SPP: a delicate real-time vehicle detection algorithm. *Optik* **225**, 165818 (2021). <https://doi.org/10.1016/j.ijleo.2020.165818>
 - 33. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: (2017) Feature pyramid networks for object detection. In: Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition. 2017: 936–944. Doi: <https://doi.org/10.1109/CVPR.2017.106>.
 - 34. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: 32nd International Conference on Machine Learning 1: 448–456 (2015)
 - 35. Glorot, X., Bordes, A., Bengio, Y.: Deep sparse rectifier neural networks. *J. Mach. Learn. Res.* **15**, 315–323 (2011)
 - 36. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings. 1–14 (2015)
 - 37. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-End Object Detection with Transformers, Lecture Notes in Computer ScienceIn: N. Carion (eds). Computer Vision – ECCV 2020: 16th European Conference, Glasgow, August 23–28, 2020, Proceedings, Part I. Springer International Publishing, Cham UK (2020)
 - 38. Ma, N., Zhang, X., Zheng, H.T., Sun, J.: Shufflenet V2: Practical guidelines for efficient cnn architecture design, Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 11218 LNCS (2018) 122–138. Doi: https://doi.org/10.1007/978-3-030-01264-9_8.
 - 39. Wang, C.-Y., Bochkovskiy, A., Liao, H.-Y.M.: YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. (2022) 1–15. <http://arxiv.org/abs/2207.02696>.
 - 40. Gilani, I.H., Amjad, M., Khan, S.S., Khan, I., Larkin, S., Raw, B., Abbas, Z.: PEMFC application through coal gasification along with cost-benefit analysis: a case study for South. Africa (2021). <https://doi.org/10.1177/0144598721999720>
 - 41. Wang, D., Abbas, Z., Du, Z., Du, Z., Lu, L., Zhao, K., Zhao, X., Yuan, Y., Zong, H., Cui, Y., Suo, L., Liang, J.: Phase field simulation of electrohydrodynamic jet droplets and printing microstructures on insulating substrates. *Microelect Eng.* **261**, 111817 (2022). <https://doi.org/10.1016/j.mee.2022.111817>
 - 42. Abbas, Z., Harijan, K., Shaikh, P.H., Walasai, G.D., Ali, F.: Effect of ambient temperature and relative humidity on solar pv system performance: a case study of quaid-e-azam solar park. Pakistan, Sindh Univer Res J-Sci Ser. **49**, 721–726 (2017). <https://doi.org/10.26692/surj/2017.12.47>
 - 43. Ren, S., He, K., Girshick, R., et al.: Faster R-CNN: towards real-time object detection with region proposal networks. *Comp. Vis. Patt. Recog.* (2023). <https://doi.org/10.48550/arXiv.1506.01497>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.