DSYOLO-Trash: An Attention Mechanism-Integrated and Object Tracking Algorithm for Solid Waste Detection

Wanqi Ma \textsuperscript{a,b}, Hong Chen \textsuperscript{a,b,*}, Wenkang Zhang \textsuperscript{c}, Han Huang \textsuperscript{d}, Jian Wu \textsuperscript{a,b}, Xu Peng \textsuperscript{a,b}, Qingqing Sun \textsuperscript{d}

\textsuperscript{a} School of Business, Jiangnan University, Wuxi 214122, PR China
\textsuperscript{b} Research Institute of National Security and Green Development, Jiangnan University, Wuxi 214122, PR China
\textsuperscript{c} State Key Laboratory of Advanced Design and Manufacturing for Vehicle Body, College of Mechanical and Vehicle Engineering, Hunan University, Changsha 410082, PR China
\textsuperscript{d} School of Economics and Management, China University of Mining and Technology, Xuzhou 221116, PR China

* Correspondence: hongchenxz@163.com

Abstract: With the rising living standards, urban solid waste production has increased significantly, posing challenges for waste management. Improving waste classification efficiency is vital for resource utilization and environmental protection. However, current waste recycling standards are too complex for the general public. A real-time DSYOLO-Trash algorithm was developed based on YOLOv5, incorporating CBAM and CotNet attention modules. Using the TrashNet dataset, the model's performance was evaluated based on loss value changes, detection accuracy, AP values, and mAP values. Comparative studies were conducted against established algorithms under equivalent conditions. Results showed that DSYOLO-Trash achieved an mAP of 97.3\% and recall rate of 93.7\% for all categories, outperforming YOLOv5 and YOLOv7. The algorithm was integrated into an intelligent waste detection and sorting system, utilizing pneumatic grippers to place items in the appropriate recycling bins. DeepSORT tracking algorithm was used to enhance detection efficiency. DSYOLO-Trash contributes to intelligent waste management, promoting sustainable urban development.

Keywords: YOLOv5; DeepSORT; CBAM; CotNet; Attention mechanism; waste management

1. Introduction

Globally, as urban populations continue to expand, the complexity and volume of municipal solid waste (MSW) increase significantly, further exacerbating the environmental impact of urban living (Liao et al., 2022). World Bank data reveals that the global MSW totalled approximately 2.01 billion tons in 2016, and it is expected to rise to an annual production of 3.4 billion tons by 2050 and this rate of increase is more than double that of population growth (Kaza et al., 2018). This scenario precipitates environmental and health issues, which are further exacerbated by the current inadequacies in waste management practices (Malakahmad et al., 2017). Moreover, the vast
quantities of unrecycled MSW significantly deplete natural resources, posing a threat to the resource foundation of future generations (Zaman and Lehmann, 2011). These rapid changes have resulted in a dramatic surge in the generation of municipal solid waste, necessitating immediate attention and the development of sustainable solutions in both academia and policymaking. Waste segregation and recycling not only address the challenges of waste management but also yield socio-economic benefits (Ding et al., 2021). Material recovery effectively reduces net carbon emissions, enhances carbon sequestration in soil and forests, and serves as a catalyst for counterbalancing entropic energy losses. Therefore, the recycling and reprocessing of waste materials has been widely considered as an integrative approach that allows for the maximization of resource consumption while minimizing additional environmental costs, which can further mitigate the environmental repercussions of rapid urbanization. For instance, in Nigeria, Ayodele et al. (2018) conducted a comprehensive assessment in Nigeria, examining the energy, economic, and environmental benefits derived from recoverable resources within urban solid waste. The study utilized a model equation method, and the findings revealed that through recycling and reprocessing, an annual saving of approximately 89.99 toe (tons of oil equivalent) energy can be achieved, along with an economic gain of $110,000 and a reduction of about 71.307 tons in CO2 emissions. These findings underscore that waste recovery and reprocessing present an effective environmental strategy capable of maximizing resource utilization while minimizing additional environmental costs. In China, Liu et al. (2020) conducted a life-cycle assessment on the economic and environmental benefits of recycling waste paper. The results showed that in 2017, the economic benefits of waste paper recycling amounted to CNY 458.3 per ton while concurrently reducing CO2 emissions by approximately 901.1 kg. This study underscores the importance of material recovery in effectively reducing net carbon emissions, enhancing carbon sequestration in soil and forests, and serving as a catalyst in balancing entropy energy loss (Song et al., 2022). Similarly, in South Korea, Jang et al. (2020) estimated the environmental benefits of recycling plastic packaging using Material Flow Analysis (MFA). Their estimates showed that approximately 6.6 metric tons of CO2 equivalent greenhouse gas emissions can be saved per annum through plastic packaging recycling. Consequently, the efficient implementation of urban waste segregation and recycling utilization is of paramount importance for the collective benefit of environmental and economic systems.

At present, solid waste segregation primarily involves manual labor-intensive separation and mechanical separation techniques based on magnetic properties, sensor technology, screening, and eddy currents (Gundupalli et al., 2017). However, these methods exhibit certain limitations. For instance, the labor-intensive separation is associated with high costs and potential health hazards (Lu et al., 2022). Moreover, magnetic, sensor-based, and other mechanical separation techniques
are only applicable to specific material identification (Huang et al., 2010). Therefore, exploring alternative, efficient, and sustainable waste segregation approaches remains a critical area of research and development in the waste management field. Due to the substantial workload, proneness to errors, and low classification efficiency, computer vision (CV) has rapidly evolved and been applied to the recognition of various scene types. By utilizing machine learning and deep learning tools, computers can be trained to extract meaningful features from visual input, which can reduce labor costs, conserve human resources, and optimize resource reuse (Wu et al., 2023).

Currently, waste classification methods based on machine learning image recognition combine features such as color and texture to identify waste images. Nevertheless, traditional image recognition algorithms typically rely on relatively singular datasets, resulting in weak generalization capabilities (Rad et al., 2017). Consequently, there is room for improvement in accuracy and operational efficiency. In recent years, many scholars have conducted in-depth research on solid waste classification, making progress in waste image recognition and classification. The potential of CV for indirect waste separation has long been acknowledged (Gundupalli et al., 2017). With the continuous development of deep learning technology, waste detection using deep learning methods has become a primary research direction for scholars, attracting widespread attention in the waste management domain in recent years (Aral et al., 2018). However, the application of these methods in waste separation has been limited due to the relatively slow computational capacity and time delays. Thus, it is necessary to propose more advanced methods to identify and sort small-sized waste rapidly and online. The YOLO algorithm network, which is commonly acknowledged as one of the most efficient object detection methods (Redmon et al., 2016), can transform the target classification and localization problem into a regression problem, resulting in improved detection speed compared to other Convolutional Neural Networks (CNNs). With the evolution of the first four versions, YOLOv5 has achieved significant detection accuracy and speed, particularly for small objects (Yao et al., 2021). However, the existing YOLOv5 algorithm still fails to meet real-time detection requirements in practical environments. Hence, improving the accuracy of waste image detection algorithms is an urgent issue to be addressed.

To fill the research gaps mentioned above, this study establishes a real-time detection and tracking model for solid waste using CV. The model performs detection of different categories of solid waste, yielding real-time results. Once the waste types are rapidly and accurately detected, a robotic arm is deployed to pick them up and place them into the appropriate recycling bins. In cases where waste is missed or obscured, an image undergoing the target tracking algorithm can identify the corresponding waste type based on its waste ID. Ultimately, the detection algorithm and target tracking algorithm are seamlessly integrated into a real-time solid waste detection system for...
intelligent waste sorting. The main contributions of this work can be summarized as follows:

1. A detection algorithm incorporating convolutional block attention module (CBAM) and contextual transformer networks (CotNet) attention modules is proposed for real-time solid waste identification. It improves the detection accuracy while reducing detection time;
2. The deep simple online and realtime tracking (DeepSORT) target tracking algorithm is embedded in the DSYOLO-Trash detection algorithm for real-time tracking of solid waste;
3. A solid waste real-time detection and intelligent sorting system is designed, enabling real-time detection and intelligent sorting operations for solid waste management.

2. Literature review

2.1 Traditional machine learning algorithms for solid waste detection

In recent years, numerous scholars have conducted extensive empirical research in the field of solid waste detection. In earlier studies, traditional machine learning methods were predominantly adopted. Belongie et al. (2002) introduced an innovative approach for feature extraction based on a descriptor known as Shape Context, which was capable of precisely gauging the similarity among various shapes, thereby facilitating object recognition. Subsequently, Liu et al. (2010) proposed the augmented Latent Dirichlet Allocation (aLDA) model within the Bayesian generative framework. This model employed a comprehensive suite of mid to low-level features, meticulously capturing multiple facets of material appearance, culminating in object classification based on material identification. Aiming to further simplify the processing workflow, Adedeji et al. (2019) proposed an advanced intelligent waste sorting system which employed a 50-layer Residual pre-trained CNN model, a robust machine learning tool intended to function as a feature extractor. This system utilized Support Vector Machines (SVM) for waste classification, encompassing multiple categories, including glass, metal, paper, and plastic. Simultaneously, Gundupalli Paulraj et al. (2016) proposed a waste identification classification method based on thermal imaging. During the training phase, they constructed a feature bag model using the faster SURF features and learned "visual vocabulary" through the K-means clustering method (Zhang and Zhao, 2017). After vector quantization of these "visual words", a multi-class SVM classifier model was ultimately established using the clustering model. Liu, Y et al. (2018) combined the SURF algorithm with the Bag of Words (BoW) model to design a unique waste image feature extraction algorithm—SURF-BoW. During the training process, they employed this algorithm to extract feature vectors from training waste images, storing these feature vectors in a matrix. Subsequently, the matrix was divided into K categories through the K-means method, counting the frequency of different categories for classification. Salimi et al. (2018) fused the texture features extracted from the Gray-Level Co-occurrence Matrix (GLCM) with the
shape features obtained from HoG, and then used an SVM classifier to classify these features to distinguish between non-waste, organic waste, and non-organic waste. These innovative methods and techniques offer new perspectives for the detection and classification of solid waste and lay the foundation for future research.

2.2 Deep learning algorithms for solid waste detection

With the advancement of mobile devices and the emergence of deep learning, the latter has been extensively deployed in solving practical problems. The aim to elevate the level of intelligence and automation in waste recognition, thereby ensuring a more accurate and efficient classification, has led to significant shifts in the models used for target detection. Thung et al. (2016) manually assembled a dataset comprising around 3,000 waste images, which they analyzed using both SVM and CNN. Their SVM model yielded a classification accuracy of 63%, whereas the CNN model reached only 22% accuracy. Lu et al. (2022) employed the YOLOv3 image detection algorithm, training it on a self-curated dataset, and utilized it for detecting and recycling Printed Circuit Boards (PCBs) within electronic waste. Li J et al. (2022) built a real-time segmentation network based on Mask R-CNN for the automatic sorting of Construction and Demolition Waste (CDW). They achieved an accuracy rate of 96.22% and established an RGB-depet (RGB-d) detection platform for real-life construction waste detection operations. Sterkens et al. (2021) applied deep neural networks to annotate X-ray images of the internal structures of Waste Electrical and Electronic Equipment (WEEE). Experimental results demonstrated that this method yielded an accuracy of 89% for battery detection, reducing the high cost of manual battery separation and enhancing the safety of the process. Zhang Q et al. (2021) introduced a garbage image classification model based on transfer learning, utilizing the DenseNet169 architecture. Testing on their self-created NWNU-TRASH dataset, this model exceeded an accuracy rate of 82%, indicating its potential for solid waste detection in complex scenarios. A distinct study conducted Zhang et al. (2022) demonstrated the superior performance of the optimized YOLO-WASTE model in simultaneously recognizing and locating multi-label waste. Chen et al. (2021) employed unmanned aerial vehicles (UAVs) with a CNN framework to accurately detect scattered waste areas automatically. Tan et al. (2022) by combining the improved YOLOv5 algorithm with Near-Infrared Spectroscopy sensor technology, were able to implement automatic sorting of obsolete washing machine parts, achieving an average detection accuracy of 98.7%.

In summary, through the investigation and analysis of the aforementioned domestic and international literature, it is evident that models based on image recognition and deep learning have been extensively applied across various industries. However, there is currently a limited amount of research focusing on the application of image recognition technology in waste detection.
Simultaneously, there are few publicly available datasets for waste classification, and the image recognition techniques employed are relatively simplistic. Consequently, integrating image detection technology with deep learning algorithms and applying this combination to waste detection is an inevitable research trend.

3. Materials and methods

3.1 Dataset construction

3.1.1 Dataset of the TrashNet

TrashNet is a small-scale recyclable waste image dataset created by Thung et al. (2016). This dataset comprises images that were captured by placing objects on a white poster board and using either natural light or indoor lighting. All images in the dataset have been resized to a spatial resolution of $512 \times 384$ pixels. It consists of RGB images of six waste categories, with each image exclusively featuring a single type of waste: glass, paper, cardboard, plastic, metal, and general trash. Currently, the dataset contains a total of 2,528 images, distributed among the categories as follows: 501 for glass, 594 for paper, 403 for cardboard, 483 for plastic, 410 for metal, and 137 for general trash. The dataset encompasses six categories of solid waste: glass, paper, cardboard, plastic, metal, and other trash. Fig. 1 presents example images extracted from the TrashNet dataset.

![Fig. 1. Example images in the TrashNet dataset.](image)

3.1.2 Dataset annotation

During the process of model training, the discrepancy between the predicted and true values is characterized by a loss function. The search for the optimal solution involves iteratively adjusting the weight parameters through the technique of backpropagation. As a result, prior to conducting the experiments, it is imperative to annotate the dataset in order to acquire the accurate labels. In this study, the LabelImg tool was employed to annotate all images according to the VOC2007 dataset standard format, generating XML formatted label files. The annotation interface can be seen in Fig. 2. A total of six target objects were subjected to annotation: glass, paper, cardboard, plastic,
metal, and other trash.

Fig. 2. Schematic diagram of TrashNet dataset annotation.

### 3.1.3 Dataset distribution

The detailed data of the TrashNet dataset is presented in Table 1, which encompasses a total of approximately 2,528 images, including 501 images of glass, 594 images of paper, 403 images of cardboard, 483 images of plastic, 410 images of metal, and 137 images of other trash. The division of the dataset into training and testing sets followed a ratio of 7:3.

<table>
<thead>
<tr>
<th>Categories</th>
<th>Numbers</th>
<th>Training sets (70%)</th>
<th>Testing sets (30%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paper</td>
<td>594</td>
<td>475</td>
<td>119</td>
</tr>
<tr>
<td>Glass</td>
<td>501</td>
<td>401</td>
<td>100</td>
</tr>
<tr>
<td>Plastic</td>
<td>483</td>
<td>386</td>
<td>97</td>
</tr>
<tr>
<td>Metal</td>
<td>410</td>
<td>328</td>
<td>82</td>
</tr>
<tr>
<td>Cardboard</td>
<td>403</td>
<td>322</td>
<td>81</td>
</tr>
<tr>
<td>Trash</td>
<td>137</td>
<td>110</td>
<td>27</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>2528</strong></td>
<td><strong>2022</strong></td>
<td><strong>506</strong></td>
</tr>
</tbody>
</table>

### 3.2 the DSYOLO-Trash model

In the field of CV, attention mechanisms have been widely recognized and extensively employed for various tasks, such as classification, detection, and segmentation. In the context of CNNs, attention mechanisms are applied to feature maps to extract salient attention information. As illustrated in Fig. 3, the DSYOLO-Trash algorithm incorporates the CBAM after the Backbone and before the feature fusion in the Neck network, effectively bridging the gap between these two stages. Additionally, the CotNet attention module is integrated during the Head stage, which not only enhances detection accuracy but also reduces the detection time, making it well-suited for real-time solid waste detection environments.
3.2.1 the YOLOv5 model

The YOLOv5 is a prominent representative of One-Stage detection algorithms, with YOLOv5s known for its compact model size and rapid detection speed. As illustrated in Fig. 4, its network structure primarily comprises four sections: Input end, Backbone, Neck, and Prediction output end. The Input end employs Mosaic data augmentation to enrich the dataset, where four images are randomly selected, scaled, cropped, and concatenated. When the image size differs, adaptive image scaling is used to reduce computational load and enhance detection speed. The Backbone utilizes the CSPDarkNet network structure. The first layer replaces the prior version's Focus structure with a $6 \times 6$ convolutional layer, which proves more efficient given current GPUs and optimization algorithms. The Spatial Pyramid Pooling (SPP) (He et al., 2015) structure is modified into SPPF, where it passes serially through the MaxPool layer. This ensures both efficiency and detection speed. The Neck incorporates an FPN (Lin et al., 2017) and PAN (Liu, S. et al., 2018) structure. The FPN layer communicates strong semantic features from top to bottom, while the feature pyramid conveys strong positioning features from bottom to top. It aggregates parameters from different backbone layers to different detection layers, thereby enhancing the network's feature extraction. The Prediction output end screens the overlapping area of the confidence bounding box with other candidate boxes. Finally, it predicts targets of different sizes on feature maps of different sizes, undergoes Non-Maximum Suppression (NMS) (Neubeck and Van Gool, 2006) and outputs the category with the highest confidence score.
3.2.2 CotNet attention mechanism

In CotNet, a novel Transformer-style module, namely the Contextual Transformer (CoT) block, is developed for visual recognition (Li, Y. et al., 2022). This design effectively exploits the contextual information between input keys to guide the learning process of dynamic attention matrices, thereby enhancing the capability of visual representation. The CoT block initially encodes the input keys through contextual convolution, resulting in static contextual representations of the inputs. Furthermore, the encoded keys are connected with input queries through two consecutive convolutions. By multiplying the learned attention matrix with the input values, dynamic contextual representations of the inputs are achieved. The fusion of static and dynamic contextual representations is ultimately employed as the output. Finally, the CotNet module is integrated into the Head end of the YOLOv5 algorithm to improve the prediction accuracy, as illustrated in Fig. 5.

3.2.3 CBAM attention mechanism

![Fig. 4. YOLOv5 network architecture.](image)

![Fig. 5. CotNet architecture.](image)
CBAM, which was initially introduced by Woo et al. (2018). In 2018, is an attention mechanism that combines channel attention and spatial attention. The structural representation of CBAM is shown in Fig. 6. The channel attention module processes the input feature map through parallel max-pooling and average-pooling layers. This process transforms the feature map into a size of $C \times 1 \times 1$. Subsequently, it traverses through a Multi-layer Perceptron (MLP) (Taud and Mas, 2018) to reduce the channel count by a factor of $r$. Afterwards, the compressed channel count is expanded back to the original number. After activation by a ReLU function, two outputs are obtained. These outputs are element-wise added and fed into a Sigmoid activation function. The activated output is multiplied by the original feature map to obtain the output result of the channel attention module, as shown in Fig. 6. The spatial attention module takes the output result of the channel attention module as input. After max-pooling and average-pooling, two $1 \times H \times W$ feature maps are obtained. These feature maps are concatenated and passed through a convolutional layer to generate a feature map with a channel count of 1. Subsequently, the feature map undergoes activation by a Sigmoid function and is multiplied by the original feature map to obtain the input for the spatial attention module, as shown in Fig. 6. This input also serves as the output for the entire CBAM module.

Fig. 6. Structure of CAM and SAM Modules: (a) Channel Attention Module; (b) Spatial Attention Module.

3.3 DeepSORT object tracking algorithm

DeepSORT (Wojke et al., 2017), an advancement over the Simple Online and Realtime Tracking (SORT) framework, facilitates multi-object online tracking. It captures raw video frames representing solid waste, applies the YOLO-Trash detection algorithm for real-time analysis, extracts appearance and motion features from detected targets, calculates inter-frame target correlations, and assigns unique identifiers to each tracked object.

The principal challenge in detection-based tracking involves prediction matching and subsequent updates after detection, essentially requiring the precise correlation and localization of
targets from one frame to the next. To address this issue in tracking solid waste identified by the
detection model, DeepSORT employs an 8-dimensional variable $x$ to encapsulate both the
appearance and motion information of the solid waste within the image, as denoted in equation (1):

$$x = (u, v, r, h, \dot{x}, \dot{y}, \dot{r}, \dot{h})$$ (1)

where $(u, v)$ represents the central coordinates of the solid waste, $r$ denotes the aspect ratio of
the waste detection box, and $h$ signifies the height of the waste detection box. Consequently,
$(u, v, r, h)$ corresponds to the velocity information of $(\dot{x}, \dot{y}, \dot{r}, \dot{h})$.

DeepSORT incorporates the motion and appearance information of solid waste on a conveyor
belt, utilizing the Hungarian algorithm (Mills-Tettey et al., 2007) for matching prediction boxes and
tracking boxes. In terms of motion information for solid waste, the Mahalanobis distance
(McLachlan, 1999) is employed to depict the correlation between the Kalman filter (Welch, 2020)
prediction results and the detection results of the YOLO-Trash algorithm, as indicated in equation
(2):

$$d^{(1)}(i, j) = (d_j - y_i)^T S_i^{-1} (d_i - y_j)$$ (2)

where $d^{(1)}(i, j)$ represents the degree of motion matching between the $j$-th detected target and
the $i$-th trajectory. Here, $S_i$ is the covariance matrix of the predicted observation space at the current
time after using the Kalman filter on the trajectory; $y_i$ denotes the predicted observation quantity
of the trajectory at the current time, while $d_j$ refers to the position information of the $j$-th detected

target.

In cases where two or more solid waste objects overlap or obstruct each other, the appearance
model comes into play. At this point, the feature extraction network computes a 128-dimensional
feature vector for each detection box, subject to the constraint $||r_j|| = 1$. Simultaneously, an
appearance feature vector of 100 frames is constructed for the detected solid waste, establishing a
definitive trajectory. The minimum cosine distance between these two entities is calculated using
equation (3):

$$d^{(2)}(i, j) = \min \{1 - r_j^T r_k^l | r_k^l \in R_l \}$$ (3)

where $r_j$ represents the feature vector corresponding to the detection box, while $r_k$ represents the feature vector associated with 100 successfully correlated frames. Finally, these two
are combined in a weighted fashion, serving as the ultimate measurement standard, as shown in
equation (4):

$$C_{ij} = \lambda d^{(1)}(i, j) + (1 - \lambda) d^{(2)}(i, j)$$ (4)

where $\lambda$ denotes the weight coefficient. If $C_{ij}$ falls within a specified range, a correct
correlation is deemed to have been achieved.

The Hungarian algorithm is leveraged to ascertain the congruity between a target in the current
frame and one in the prior frame. The obstacle of probability dispersion due to the Kalman filter
prediction after prolonged obstruction or significant frame loss is circumvented through cascade matching. This technique ensures that the predicted target corresponds to the pre-obstruction or pre-loss target. Implementing this strategy, the DeepSORT algorithm achieves precise positioning of detected targets in subsequent frames.

4 Experimental Design and Results Analysis

4.1 Experimental Environment

Due to the high number of layers and complexity in the DSYOLO-Trash algorithm developed in this study, a substantial amount of complex iterative calculations is required, necessitating a relatively high-performance experimental environment. The hardware configuration consists of an Intel Core i9-12900K processor and an NVIDIA RTX 3090 graphics card with 24 GB of video memory. In terms of software infrastructure, the experimental platform is based on the Windows 11 operating system and the PyTorch deep learning framework, utilizing CUDA 11.1 and cuDNN 8.0.5 for high-performance parallel computing. The dataset used for the DSYOLO-Trash model comprises a total of 2,528 images, which are divided into training, validation, and testing sets. The training and testing sets are split in a 7:3 ratio, with 90% of the images in the training set used for network training, and the remaining 10% for validation.

4.2 Model Training

For the TrashNet solid waste image dataset, adjustments need to be made to the dataset placement when training the DSYOLO-Trash model. Before training, the VOCdevkit folder, which contains the dataset, should be placed in a designated file path. The label files and image files should be placed separately in the Annotation and JPEGImages folders under the VOC2007 folder within the VOCdevkit folder. Once the dataset is properly placed, further processing is required to obtain yolo_train.txt and yolo_val.txt for training. At this point, modifications to the code are necessary. More specifically, the trash yaml file corresponding to the DSYOLO-Trash model in the models folder should be altered to change the number of training categories to 6, corresponding to the six different solid waste categories in the TrashNet dataset. Finally, the maximum number of training iterations and the initial learning rate must be modified in the train.py file. During the training process, the generated weight files will be saved in the logs folder, facilitating subsequent image prediction tasks. For the training of the solid waste image detection algorithm, 2528 waste images from the TrashNet dataset will be used. These images will be split into training and testing sets in a 7:3 ratio. The PyTorch deep learning framework is employed. The input solid waste image size is set to 640 × 640, and the initial learning rate is set to 0.001. Also, the training process starts from epoch 0 for a total of 100 epochs.

4.3 Evaluation Metrics
This study primarily utilizes Average Precision (AP) and mean Average Precision (mAP) as evaluation metrics to assess the algorithm's performance. The computation of Average Precision (AP) encompasses the notions of Precision and Recall, as demonstrated in Equation (5) and Equation (6):

\[
\text{Precision} = \frac{TP}{TP + FP} \times 100\% \\
\text{Recall} = \frac{TP}{TP + FN} \times 100\%
\]

where TP refers to the count of samples correctly predicted as positive by the model, FP signifies the count of samples incorrectly predicted as positive by the model while being negative, and FN represents the count of samples predicted as negative by the model but are in fact positive, as provided in Equation (6) and Equation (7):

\[
AP = \int_0^1 P(R) dR
\]

\[
mAP = \frac{\sum_{i=1}^{C} AP_i}{C}
\]

where \( i \) denotes different categories, totaling \( C \) types; \( AP_i \) represents the average precision of the \( i \)-th category.

A Precision-Recall (P-R) curve can be generated by plotting precision against recall at various recall levels. The y-axis represents the highest precision achieved at each recall level, while the x-axis represents the corresponding recall level. Equation (5) and Equation (6) demonstrate that the area under the P-R curve represents the average precision (AP). If there are \( C \) categories of interest, the mean average precision (mAP) can be calculated using Equation (7) and Equation (8). Therefore, by integrating the P-R curve, the AP can be obtained, and by averaging the AP values across all categories, the mAP can be obtained.

### 4.4 Experimental results

Fig. 7 illustrates the confusion matrix (Zhang et al., 2020) used to evaluate the prediction accuracy of the DSYOLO-Trash solid waste detection algorithm. The horizontal axis of the matrix represents the true annotated categories, namely cardboard, glass, metal, waste paper, plastic, and other waste materials. At the same time, the vertical axis represents the predicted categories for cardboard, glass, metal, waste paper, plastic, and other waste materials. The matrix shown in Fig. 7 contains data reflecting the predicted probability values. Notably, the data on the diagonal represent the True Positives probability values, i.e., the probability of correctly classifying positive samples. By examining the provided detailed information, the probability values for each category can be clearly analyzed.
In this section, the DSYOLO-Trash model was used to conduct experiments on the TrashNet dataset created by the authors. Table 2 summarizes the performance metrics obtained during testing on the TrashNet dataset. The experimental results show that the model achieved a detection precision of 92.6% and a recall rate of 93.7% for detecting different types of garbage objects. Furthermore, conducted on individual sample sets of each category during the experiments. The DSYOLO-Trash model demonstrated high accuracy rates, with 94% for waste paper, 93.1% for glass, 96.4% for plastic, 96.8% for metal, 95.4% for cardboard, and 79.7% for other garbage.

<table>
<thead>
<tr>
<th>Categories</th>
<th>Numbers</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paper</td>
<td>594</td>
<td>94</td>
<td>97.1</td>
</tr>
<tr>
<td>Glass</td>
<td>501</td>
<td>93.1</td>
<td>95.6</td>
</tr>
<tr>
<td>Plastic</td>
<td>483</td>
<td>96.4</td>
<td>91.7</td>
</tr>
<tr>
<td>Metal</td>
<td>410</td>
<td>96.8</td>
<td>97.1</td>
</tr>
<tr>
<td>Cardboard</td>
<td>403</td>
<td>95.4</td>
<td>97</td>
</tr>
<tr>
<td>Trash</td>
<td>137</td>
<td>79.7</td>
<td>83.8</td>
</tr>
<tr>
<td>Total</td>
<td>2528</td>
<td>92.6</td>
<td>93.7</td>
</tr>
</tbody>
</table>

To evaluate the performance of the DSYOLO-Trash solid waste detection algorithm, the P-R curve is utilized to display the relationship between precision and recall. The area enclosed by the P-R curve and the horizontal and vertical axes represents the AP value. Typically, a higher the AP value indicates the better performance of the algorithm model. In Fig. 8, the precision-recall curves for each type of garbage are presented after training the DSYOLO-Trash model on the TrashNet dataset. As shown in Fig. 8, the AP values for glass, cardboard, metal, waste paper, plastic, and other garbage are 98%, 98.9%, 99.1%, 98.9%, 98.7%, and 90.2%, respectively. Based on the AP
values for each type of garbage detection, the DSYOLO-Trash model demonstrates good performance in detecting various types of garbage. The mAP for the DSYOLO-Trash garbage detection algorithm is the average of the AP values obtained for all categories of waste, as shown in Fig. 8. This figure also presents the overall P-R curve for the model, which attains an mAP value of 97.3%.

Fig. 8. Precision-recall curves and the mAP value of the DSYOLO-Trash model: (a) The precision-recall curve for glass; (b) The precision-recall curve for cardboard; (c) The precision-recall curve for metal; (d) The precision-recall curve for paper; (e) The precision-recall curve for plastic; (f) The overall P-R curve.
Upon completion of model training, the test set images were employed to evaluate the performance. The specific detection results are illustrated in Fig. 9. In this figure, boxes of distinct colors represent various waste types. A comprehensive list of waste categories corresponding to different labels can be found in Fig. 9 (a). Each bounding box is accompanied by two pieces of information: one indicating the waste category and the other denoting the detection accuracy. The findings suggest that the DSYOLO-Trash model is capable of accurately identifying diverse waste items within the target dataset. To provide a visual interpretation of the detection process in the DSYOLO-Trash model, this study employs the Grad-CAM visualization method (Selvaraju et al., 2017). Grad-CAM expresses the fusion weights of the target feature maps as gradients, and it calculates the weights using the global average of the gradients. After obtaining the weights for each class across all feature maps, a weighted sum is computed to generate a heatmap. The heatmap offers an intuitive representation of the focal points of the model during feature extraction. Darker colors indicate a higher level of attention, with the red regions (the darkest areas) signifying the most critical locations, as demonstrated in Fig. 9 (b).

Fig. 9. Detection results of the DSYOLO-Trash model: (a) Detection results of the DSYOLO-Trash model; (b) Heatmap of prediction results for the DSYOLO-Trash model.

An analysis of the loss values during waste detection using the DSYOLO-Trash model was conducted concurrently. Fig. 10 presents the loss values during model training, with the horizontal axis representing the number of training epochs, totaling 100. Fig. 10 (a) displays the localization loss (Box_loss), representing the error (GIoU) between the predicted bounding boxes and the ground-truth boxes. Fig. 10 (b) shows the confidence loss (Obj_loss), which calculates the network’s confidence; and Fig. 10 (c) illustrates the classification loss (Classification_loss), which evaluates the correctness of the anchor box’s corresponding ground-truth classification. According to Fig. 10 (a) to Fig. 10 (c), the vertical axis indicates loss values. The train_loss represents the model’s loss during training, the smooth_train_loss corresponds to the smoothed training loss curve, the val_loss
signifies the model’s validation loss during training, and the smooth_val_loss pertains to the smoothed validation loss curve. During each training epoch, the model automatically partitions the TrashNet dataset into training and testing subsets, with a 70% and 30% allocation, respectively. The DSYOLO-Trash model undergoes simultaneous training and validation to enable optimization throughout the training process. As the number of training epochs increases, the loss values decrease gradually, as shown in Fig. 10.

Fig. 10. Loss curves of the DSYOLO-Trash model: (a) Box_loss Curve; (b) Obj_loss Curve; (c) Classification_loss Curve.

By integrating the improved DSYOLO-Trash model with the DeepSORT algorithm for solid waste tracking, the tests on best_mobilenetv2_x1_45, best_mobilenetv2_x1_46, and best_mobilenetv2_x1_47 were conducted, and the visualized results are provided in Fig. 11. In the video, the target disappears at the 5th frame due to a missed detection by the DSYOLO-Trash model, rendering it untrackable. However, at the 10th frame, the DeepSORT algorithm re-establishes tracking of the target, successfully restoring its identity and maintaining the tracking process.

Fig. 11. Solid waste ID recovery using the DeepSORT model: (a) Frame 1; (b) Frame 5; (c) Frame 10.

4.5 Comparative experiments

4.5.1 Detection Algorithm Comparison Experiment

To evaluate the effectiveness of the proposed DSYOLO-Trash solid waste detection algorithm, this section replicates these methods by using the same experimental settings for training and testing to ensure a fair comparison. A comprehensive analysis of eight mainstream object detection algorithms at the current stage, including SSD (Liu et al., 2016), Faster-RCNN (Ren et al., 2015), EfficientDet (Tan et al., 2020), YOLOv3 (Redmon and Farhadi, 2018), YOLOv4 (Bochkovskiy et
YOLOX (Ge et al., 2021), YOLOv7 (Wang et al., 2023) and YOLOv5. The YOLO series algorithms and Faster-RCNN represent the classical single-stage and two-stage object detection networks at the current stage, respectively. The YOLO series algorithms adopt the CSPDarknet network as the backbone architecture, while Faster-RCNN employs the ResNet50 network. Table 3 presents a detailed comparison of mAP@0.5 (mAP when the IoU threshold is set to 0.5) and Recall on the test set of the TrashNet dataset for different stage models. It can be observed that the YOLOv3 algorithm model exhibits the poorest performance, with an mAP of 81.4%. Also, the mAP values of YOLOv3 algorithm are only 72.4% and 59.6% when the IoU is set to 0.6 and 0.7, respectively. Among the two-stage models, the predictive performance of Faster-RCNN surpasses the SSD model, with an mAP value of 90.5%, a 0.8% improvement over VGG. That also outperforms single-stage models YOLOv3, YOLOv4, and YOLOv7. However, the DSYOLO-Trash model proposed in this study achieves the best mAP value of 97.3% and a Recall of 93.7% among the eight comparison models, with a 3.7% increase in mAP and a 4.4% increase in Recall compared to YOLOv5. Additionally, its mAP value is 7.6%, 6.8%, 8.6%, 15.9%, 14.8%, 4.3%, and 9% higher than those obtained by SSD, Faster R-CNN, EfficientDet, YOLOv3, YOLOX, YOLOv4, and YOLOv7 models, respectively. This demonstrates that the practical application of DSYOLO-Trash in solid waste detection is feasible and reliable.

### Table 3. mAP and Recall of different models on the TrashNet dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>Backbone</th>
<th>IOU</th>
<th>mAP (%)</th>
<th>Recall (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSD</td>
<td>VGG</td>
<td>0.5</td>
<td>89.7</td>
<td>-</td>
</tr>
<tr>
<td>Faster-RCNN</td>
<td>ResNet50</td>
<td>0.5</td>
<td>90.5</td>
<td>-</td>
</tr>
<tr>
<td>EfficientDet</td>
<td>EfficientNet</td>
<td>0.5</td>
<td>88.6</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.5</td>
<td>81.4</td>
<td>86</td>
</tr>
<tr>
<td>YOLOv3</td>
<td>Darknet53</td>
<td>0.6</td>
<td>72.4</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.7</td>
<td>59.6</td>
<td>70</td>
</tr>
<tr>
<td>YOLOv4</td>
<td>CSPDarknet53</td>
<td>0.5</td>
<td>82.5</td>
<td>80.9</td>
</tr>
<tr>
<td>YOLOX</td>
<td>CSPDarknet</td>
<td>0.5</td>
<td>93.4</td>
<td>-</td>
</tr>
<tr>
<td>YOLOv7</td>
<td></td>
<td>0.5</td>
<td>88.3</td>
<td>84.2</td>
</tr>
<tr>
<td>YOLOv5</td>
<td>CSPDarknet</td>
<td>0.5</td>
<td>93.5</td>
<td>89.3</td>
</tr>
<tr>
<td>Ours(DSYOLO-Trash)</td>
<td></td>
<td>0.5</td>
<td><strong>97.3</strong></td>
<td><strong>93.7</strong></td>
</tr>
</tbody>
</table>

#### 4.5.2 Classification Algorithm Comparison Experiment

In this study, the performance of the DSYOLO-Trash model was compared with previously proposed models, including SVM-SIFT, CNN, CNN+SVM, RecycleNet, ResNet18, PSO, GA. Table 4 displays the comparison of test accuracy among different models after training with the TrashNet dataset. Thung and Yang's model, utilising Scale-Invariant Feature Transform (SIFT) and SVM, achieved a test accuracy of 63% under an 80%/20% train/test split (Yang and Thung, 2016).
Meanwhile, Adedeji and Wang's model, which leverages a pre-trained 50-layer ResNet-based CNN and SVM for waste categorisation, reached an accuracy of 87% (Adedeji and Wang, 2019). Bircanoğlu et al.'s RecycleNet model attained 81% accuracy (Bircanoğlu et al., 2018). Notably, Ahmad et al.'s approach, integrating Particle Swarm Optimization (PSO) and Genetic Algorithm (GA), achieved remarkable test accuracies of 94.11% and 94.58% (Ahmad et al., 2020). The experiment was conducted on the TrashNet dataset using ResNet18+SMM, resulting in a classification accuracy of 95.87% (Zhang, Qiang et al., 2021). Our DSYOLO-Trash model demonstrated 92.6% accuracy on the TrashNet dataset, surpassing most models but slightly lower than the PSO-GA hybrid and ResNet18+SMM. These results affirm DSYOLO-Trash's efficacy in waste classification and the effectiveness of our model design and training approach.

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM-SIFT</td>
<td>63%</td>
</tr>
<tr>
<td>CNN</td>
<td>22%</td>
</tr>
<tr>
<td>CNN+SVM</td>
<td>87%</td>
</tr>
<tr>
<td>RecycleNet</td>
<td>81%</td>
</tr>
<tr>
<td>ResNet18+SMM</td>
<td>95.87%</td>
</tr>
<tr>
<td>PSO</td>
<td>94.11%</td>
</tr>
<tr>
<td>GA</td>
<td>94.58%</td>
</tr>
<tr>
<td><strong>Ours (DSYOLO-Trash)</strong></td>
<td><strong>92.6%</strong></td>
</tr>
</tbody>
</table>

### 4.6 Ablation experiments

To verify the improvement in detection performance brought by the modification experiments conducted in this study, the YOLOv5 algorithm based on the CSPDarknet backbone network is used as a comparative experimental model, and CotNet and CBAM were employed as validation modules for ablation experiments. The training results are shown in Table 5. It can be observed that adding either CotNet or CBAM attention modules separately to the model enhances the detection performance. Specifically, the introduction of the CotNet attention module alone results in an mAP value of 94.7%, a 1.1% improvement over the unmodified YOLOv5 algorithm. However, the Recall decreases by 1.3%. In contrast, the standalone introduction of the CBAM attention module yields an mAP value of 95.5%, a 1.9% increase compared to the unmodified YOLOv5 algorithm, and a 1.4% improvement in Recall. Different combinations contribute to a certain extent in enhancing the model's detection performance. The training results obtained by the DSYOLO-Trash model show an mAP of 97.3%, a 3.7% improvement compared to the unmodified YOLOv5 model, and a 4.4% increase in Recall, both achieving optimal results. These findings demonstrate that the DSYOLO-Trash model is more accurate in detecting solid waste compared to YOLOv5-CotNet and YOLOv5-CBAM, thereby improving the efficiency of real-time detection.
Table 5. Ablation experiment results for the DSYOLO-Trash model on the TrashNet dataset

<table>
<thead>
<tr>
<th>Number</th>
<th>Method</th>
<th>mAP (%)</th>
<th>Recall (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>YOLOv5 (CSPDarknet)</td>
<td>93.6</td>
<td>89.3</td>
</tr>
<tr>
<td>B</td>
<td>A+CotNet</td>
<td>94.7</td>
<td>88</td>
</tr>
<tr>
<td>C</td>
<td>A+CBAM</td>
<td>95.5</td>
<td>90.7</td>
</tr>
<tr>
<td>D</td>
<td>B+C (DSYOLO-Trash)</td>
<td>97.3</td>
<td>93.7</td>
</tr>
</tbody>
</table>

The superior performance of DSYOLO-Trash over standalone module implementations of YOLOv5-CotNet and YOLOv5-CBAM indicates the effectiveness of the integrated module fusion. This endorses the thoughtful design principles of DSYOLO-Trash. Under consistent experimental conditions and utilizing the TrashNet dataset, a comparative analysis of prediction outcomes, as depicted in Table 6, was conducted. The results demonstrate that DSYOLO-Trash provides a more robust solution to the challenges posed by overlapping and occluded solid waste. This points to its high level of robustness.

Table 6. Detection results of different experiment algorithms on the TrashNet dataset

<table>
<thead>
<tr>
<th>Experimental algorithms</th>
<th>Detection Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>YOLOv5</td>
<td>![YOLOv5]</td>
</tr>
<tr>
<td>YOLOv5 with CotNet</td>
<td>![YOLOv5 with CotNet]</td>
</tr>
<tr>
<td>YOLOv5 with CBAM</td>
<td>![YOLOv5 with CBAM]</td>
</tr>
<tr>
<td>DSYOLO-Trash</td>
<td>![DSYOLO-Trash]</td>
</tr>
</tbody>
</table>

4.7 Smart detection and automatic sorting system for solid waste

Fig. 12 introduces an intelligent detection and automated sorting system based on solid waste. This sorting system includes a vision detection component, mechanical grasping component, and a
hardware control system, designated to detect solid waste, ascertain its location on the conveyor belt, and subsequently, convey it via a robotic arm into specific recycling bins. An industrial camera is positioned directly above the conveyor belt's front end. As the solid waste set for inspection passes through the image capture area on the belt, this camera relays the real-time image data to a computer server for detection processing. The well-trained DSYOLO-Trash algorithm is employed to detect individual solid waste items and determine their categories and positions. In the event of target obstruction or camera angle jitter, the DeepSORT target tracking algorithm is employed for real-time comparison of the position IDs of the frames before and after the solid waste images captured by the industrial camera, thus tracking whether the corresponding waste has been detected and grasped. This detection information is used for hardware execution in the physical world. A Programmable Logic Controller (PLC) is utilized to receive category data from the computer and control the sorting operations. As depicted in Fig. 12, the upper half illustrates real-time transmission and mechanical gripper grasping on the conveyor belt, whereas the lower half shows the operations of the robotic arm, controlled by the PLC receiving the results from the DSYOLO-Trash algorithm. Upon arrival of the solid waste at its categorized monitoring area, a mechanical pneumatic gripper device is capable of seizing the object and placing it into its respective container. Lastly, these unidentified solid waste parts will fall into the residual container at the belt's end for further processing. This arrangement facilitates the collection of specific categories of solid waste, thereby promoting a more efficient process of solid waste recycling and classification.

**Fig. 12.** Intelligent detection and automatic sorting system for solid waste.
5 Conclusion

With the advancement of household waste classification systems, the categorization of solid waste has become a research hotspot. However, the limited understanding of classification standards among community residents and the lack of automated and auxiliary tools further intensify the categorization challenge. In this study, we propose a solid waste detection algorithm based on DSYOLO-Trash, utilizing the public dataset TrashNet, to address these issues. The main findings of our research are as follows:

(1) Our algorithm integrates the CBAM attention module into the Neck portion of the base YOLOv5 network. This effectively mines channel and spatial attention features, enhancing the model's feature fusion capability, which is crucial for enhancing the network expressiveness of the DSYOLO-Trash algorithm.

(2) In the Head portion of the YOLOv5 base network, our algorithm embeds the CotNet attention module. This guides the learning of the attention matrix through leveraging the contextual information of input keys, thereby enhancing prediction accuracy and improving the detection performance of the DSYOLO-Trash algorithm.

(3) By incorporating the DeepSORT object tracking algorithm, real-time comparisons of positional IDs in previous and subsequent frames can be conducted to address potential omissions and obstructions, thereby improving detection efficiency.

(4) Experimental results indicate that the detection accuracy and recall rate of the DSYOLO-Trash algorithm on the TrashNet public solid waste dataset surpass those of eight classic detection algorithms, such as YOLOv7, YOLOv5, and Faster-RCNN, with a mAP reaching 97.3%, and a recall rate of 93.7%.

(5) Ablation study results reveal that, compared with the integration of either the CotNet attention module or the CBAM module alone, the detection accuracy of the DSYOLO-Trash algorithm increases by 2.6% and 1.8% respectively, with recall rates also improving by 5.7% and 3% respectively, demonstrating its superiority in practical detection performance.

(6) We designed a comprehensive system for intelligent detection and automatic sorting of solid waste, implementing the DSYOLO-Trash algorithm in a practical application, contributing meaningfully to solid waste management.

In the future, based on the existing public dataset TrashNet, we aim to create a multi-label solid waste dataset to further verify the high precision and universality of the DSYOLO-Trash algorithm. Simultaneously, we will promote the practical application of the intelligent detection and sorting system, making further contributions to solid waste recycling.
Disclosure statement

No potential conflict of interest was reported by the authors.

Data availability statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgements

This work was supported by the Major Project of National Social Science Foundation of China (No. 21&ZD166), the National Natural Science Foundation of China (No. 71874178), the Special Project for Research Institute of National Security and Green Development, Jiangnan University (No. 1095219032210060), the Postgraduate Research & Practice Innovation Program of Jiangsu Province (No. 1092050205238300).

References


Stanford University: Stanford, CA, USA.


Zhang, S., Zhao, H., 2017. Algorithm research of optimal cluster number and initial cluster center. Application Research of Computers 34(6), 1617-1620.