Deep Learning Methods for Waste Classification

Jianchun Qi

A thesis submitted to the Auckland University of Technology
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Computer and Information Sciences

2024

School of Engineering, Computer & Mathematical Sciences

Abstract

Waste could be reduced, reused and recycled. Classifying waste and recycling play important roles in converting waste into valuable materials that help conserve land reduce pollution and optimize resource utilization. Effective waste management is vital for resource conservation, environmental protection, and sustainable human progress. However, there are difficulties in classifying and identifying recyclable materials due to the complex and diverse nature of waste combined with limited data on waste management. These challenges pose obstacles to the effectiveness of research, in this field.

The development of deep learning has promoted the advancement of pattern classification and visual object detection. Applying computer vision to waste classification and exploring high-efficiency, low-cost, and automated waste classification methods have important sustainability implications for society and the ecological environment.

Taken the current limitations of waste classification into account, this project makes use of deep learning methods to improve waste classification from several different perspectives. Firstly, we build two comprehensive and rich waste datasets, including four waste categories, namely recyclable waste, wet waste, hazardous waste, and dry waste. Afterwards, we have made innovations in data augmentation and proposed NUNI, a non-uniform data augmentation method, to improve the accuracy of waste classification. Finally, we propose a semi-supervised learning deep learning framework, called CISO, and utilize it on the waste classification task. We evaluate our models and provide the relevant comparisons.

Keywords: Deep learning, Waste classification, Data augmentation, Semi-supervised learning, Transformers

Table of Contents

List of Figures

List of Tables

Acronyms

AI: Artificial Intelligence

AP: Average Precision

API: Application Programming Interface

AR: Augmented Reality

ASPP: Atrous Spatial Pyramid Pooling

BBox: Bounding Box

BCE Loss: Binary Cross-Entropy Loss

BERT: Bidirectional Encoder Representations from Transformers

BYOL: Bootstrap Your Own Latent

CBAM: Convolutional Block Attention Module

CCT: Compact Convolutional Transformers

CIoU: Complete Intersection over Union

CNN: Convolutional Neural Network

DFL: Distribution Focal Loss

ELAN: Efficient Layer Aggregation Networks

ELECTRA: Efficiently Learning an Encoder that Classifies Token Replacements Accurately

EMA: Exponential Moving Average

EPSA: Efficient Pyramid Split Attention

FFNN: Feedforward Neural Network

GAN: Generative Adversarial Networks

GPT-4: Generative Pre-training Transformer 4

IoU: Intersection over Union

InfoNCE: Info Noise Contrastive Estimation

mAP: mean Average Precision

mIoU: mean Intersection over Union

NLP: Natural Language Processing

OhemCELoss: Online Hard Example Mining Cross-Entropy Loss

PaLM-E: Pathways Language Model with Embodied

PAN-FPN: Path Aggregation Network for Feature Pyramid

PR Curve: Precision-Recall Curve

PSA: Pyramid Split Attention

R-CNN: Region-CNN

ResNet: Residual Network

RFA: Residual Feature Augmentation

RNN: Recurrent Neural Network

RoBERTa: Robustly Optimized BERT Pretraining Approach

ROI: Region of Interest

RPN: Region Proposal Network

SENet: Squeeze-and-Excitation Networks

SE Weight: Squeeze-and-Excitation Weight

SGD: Stochastic Gradient Descent

SimCLR: Simple Contrastive Learning

SPP: Spatial Pyramid Pooling

SSD: Single Shot Multibox Detector

SSL: Semi-supervised Learning

SSOD: Semi-supervised object detection

SW-MSA: Shifted Window Multihead Self-Attention

TACO: Trash Annotation in Context

UAVs: Unmanned Aerial Vehicles

VGG: Visual Geometry Group

W-MSA: Window Multihead Self-Attention

XLNet: Extra Long Network

YOLO: You Only Look Once

Attestation of Authorship

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person (except where explicitly defined in the acknowledgments), nor material which to a substantial extent has been submitted for the award of any other degree or diploma of a university or other institution of higher learning.

Signature: Jianchun Qi Date: 1 March 2024

Acknowledgment

This thesis represents my long learning journey. During these three years of study, I would like to express my sincere gratitude to my supervisors. My heartfelt thanks go to my primary supervisor Minh Nguyen for his help and support. He provided me with great guidance and gave me a lot of constructive feedback on my research paper. He would also seriously give me a lot of advice and encourage me during my study process.

Furthermore, I would also like to extend my heartfelt thanks to my supervisor Wei Qi Yan and Yanbin Liu. They were very attentive in making effective suggestions on my research directions and research papers, and often shared their academic experience with me. They have done their best to help me when I encounter problems in my studies. I consider myself lucky to have benefited from the academics.

I am also grateful to the Auckland University of Technology (AUT) and all the staff for their support. And my friends, they also provided me with a lot of help.

Afterwards, I would like to thank my family for always supporting me the freedom to choose to do what I love. I will always love them no matter where they are.

Finally, I wish to thank my elder brother Mr. Qi. His encouragement, enlightenment, and support made me continue to become a better version of myself. Thousands of words, so much more was said in the unsaid.

Jianchun Qi

Auckland, New Zealand

March 2024

Chapter 1
Introduction

In this thesis, the first chapter contains five parts. In the first part, we introduce the background and motivation for this research project. The second part mentions the research questions related to the project. In the third part, our contributions are presented. Afterwards, our objectives and the structure of the thesis are summarized in part four and part five respectively.

1.1 Background and Motivation

Waste management is an increasingly important topic. It is predicted that the total amount of global waste will increase by one-fifth after 2030, with an imminent output of up to 2.6 billion tons of waste per year (Kaza, Yao, Bhada-Tata and Van, 2018). Inadequate waste management mechanisms pose a great challenge to the protection of the ecological environment, the improvement of public health, and the safeguarding of human health. For example, the current open waste dumps, this unhealthy waste management method is prone to produce hazardous chemicals (Mohanraj, Senthilkumar, Chandrasekar and Arulmozhi, 2023), pollute the soil and water, and damage the ecosystem (Chen et al., 2020), in addition to becoming a breeding ground for pathogens, which can easily lead to the spread of infectious diseases (Amasuomo and Baird, 2016; Ferronato and Vincenzo, 2019; Zaman, 2015; Zhang, Tan and Gersberg, 2010; Zhang, Hu, Zhang and Zhang, 2020). Proper waste disposal practices hold significant implications for ecological sustainability, resource efficiency, and public health enhancement (Meng and Chu, 2020).

Regarding waste classification, it should be grouped into categories according to its components, properties, value, and impact on the environment, depending on the type of disposal. In general, according to the characteristics of wastes, we group wastes into four major categories, namely hazardous waste, recyclable waste, wet waste, and dry waste. For example, valuable waste can be recycled and utilized, which can reduce the consumption of raw materials as well as the waste of resources and reduce costs. Finally, waste can be disposed, instead of incinerating it in a uniform manner, reducing CO_2 emissions and protecting the earth.

However, conventional methods of waste classification, which are typically semi-manual or semi-automatic, struggle to keep pace with the increasing volumes of waste, often resulting in inefficient sorting and adverse health effects on workers. Consequently, there is an urgent need to incorporate more sophisticated technologies,

such as artificial intelligence, into waste management. The integration of advanced AI-driven classification technologies can lead to more effective, efficient, and health-conscious waste management practices. This, in turn, supports economic growth and environmental protection, steering us toward the sustainable coexistence of humanity and nature (Qiu et al., 2022; Shi, Tan, Wang and Wang, 2021).

There are a plenty of algorithms for visual object detection in computer vision (Bachman, Alsharif and Precup, 2014; LeCun, Bengio and Hinton, 2015; Li, Wang, Hu and Yang, 2019; Lun et al., 2023; Pan and Yan, 2020; Wang et al., 2017; Yu et al., 2018; Yu, Jiang, Wang, Cao and Huang, 2016; Zhao, Zheng, Xu and Wu, 2019). The algorithms are grouped into two categories. One is the algorithm based on SSD (Liu et al., 2016), and YOLO nets (Redmon, Divvala, Girshick and Farhadi, 2016); the other is the algorithm based on R-CNN networks (Kang, Yang, Li and Zhang, 2020; Ren, He, Girshick and Sun, 2015; He, Gkioxari, Dollár and Girshick, 2017). The former is based on one-stage training with high speed, but the accuracy is not as high as the latter. These networks have been widely used (Tian, Shen, Chen and He, 2019). However, the Transformer models occupy an important position in the field of computer vision due to its superior processing capabilities and efficient computing performance (Vaswani et al., 2017). Later, various derivative models based on Transformer appeared one after another, leading the latest development in this field. For example, Vision Transformer (Dosovitskiy et al., 2020), DETR (Carion et al., 2020), and Swin Transformer (Liu et al., 2021).

Currently, deep learning models for waste classification are constantly being improved and have obtained significant classification and detection results. However, there is still room for improvement. A slew of waste classification models, such as the optimized DenseNet121 and ResNet-10 (Kashif, Khan and Al-Fuqaha, 2020) using fusion schemes, have waste classification accuracies as high as over 85% (Mao, Chen, Wang and Lin, 2021). However, the datasets they use are only simple recyclable waste categories, such as glass, cardboard, plastic bottle, paper, and metal, which cannot measure the real waste classification application scenarios. While the ETHSeg model

classifies the above four categories of waste based on X-rays, the classification accuracy of small objects in waste remains low (Qiu et al., 2022).

Thus, the lack of a waste dataset, waste objects are deformed and stacked, and the intensive manual annotation effort due to the wide variety of waste categories are the important challenges faced by artificial intelligence in waste classification tasks. Considering these limitations, the following three aspects need to be studied in-depth: how to build a rich and comprehensive waste dataset, how to use a small amount of data to achieve classification results, and how to improve the accuracy of waste classification.

With regard to computer vision tasks, the scale of training data could be employed as a factor that affects the performance of neural network models (Nayan, Saha, Mozumder and Mahmud, 2020; Neverova, Wolf, Taylor and Nebout, 2015). In general, the more data, the better the training effect of the model. Therefore, pre-training and fine-tuning are widely applied to improve model performance (Poth, Pfeiffer, Rücklé and Gurevych, 2021; Wang, Khabsa and Ma, 2020; Zhang, Zhao, Saleh and Liu, 2020). This can not only avoid overfitting but also speed up model convergence (Kohli, Sitzmann and Wetzstein, 2020). However, even if the performance of deep learning models based on supervised learning has been improved, the time-consuming and expensive work of collecting large-scale labelled datasets cannot be ignored. Therefore, the methods of applying semi-supervised learning to classification tasks are summarized (Olivier, 2006; Reddy, Viswanath and Reddy, 2018). Semi-supervised learning methods reduce the model's dependence on a large amount of labeled data, as well as the high labor and time costs caused by manual annotation (Berthelot et al, 2019; Chapelle, Schölkopf and Zien, 2010; Rasmus, Berglund, Honkala, Valpola and Raiko, 2015).

Consistency regularization is a key method in semi-supervised learning (Lee, 2021; Li, Liu, Zhao, Zhang and Fu, 2021; Miyato, Maeda, Koyama and Ishii, 2018; Sajjadi, Javanmardi and Tasdizen, 2016). Its core idea is that the model should respond to inputs data that are disturbed or changed, such as cropping, color transformation, rotation, and

given prediction results that are consistent with the original prediction. This method is particularly suitable for scenarios with large amounts of unlabeled data, helping to improve the model's generalization ability on unlabeled data (Sajjadi, Javanmardi and Tasdizen, 2016; Tarvainen and Valpola, 2017; Yang et al., 2022; Yang, Song, King and Xu, 2022).

In addition, data augmentation techniques are also closely related to semi-supervised learning methods (Kim et al., 2020; Sohn et al., 2020; Xie, Dai, Hovy, Luong and Le, 2020). Data augmentation not only increases the data volume by applying transformations to the original data, but also introduces perturbations and changes to the original data. In this way, the enhanced unlabeled data can provide more diverse data representations, assist the semi-supervised learning model improve its generalization ability, and enable the model to learn more robust feature representations (Krizhevsky, Sutskever and Hinton, 2017). Therefore, data augmentation plays a crucial role in driving progress in the field of semi-supervised learning. Currently, a great deal of improvements to semi-supervised models have verified the effectiveness of data augmentation methods (Suzuki, 2022; Xie, Luong, Hovy and Le, 2020).

In the research field of visual object detection, the application of semi-supervised learning has gradually emerged and achieved remarkable achievements (Girshick, 2015; Liu et al., 2021; Wu, Sahoo and Hoi, 2020; Bar et al., 2022; Jeong, Lee, Kim and Kwak, 2019). An important breakthrough in this field is the proposal of the STAC method (Sohn et al., 2020). Moreover, the Instant-Teaching method optimizes STAC, which has an important impact in the field of semi-supervised object detection (SSOD) and provides new directions for future SSOD research (Zhou, Yu, Wang, Qian and Li, 2021). The improvement of Instant-Teaching starts from the following two key points. First, the co-rectify strategy is used to solve the bias problem caused by pseudo labels. Second, the real-time pseudo label generation model is applied.

In this project, semi-supervised learning has gradually matured in visual tasks. Semi-supervised learning can learn from unlabeled data, reducing the need for a large

amount of manually labeled data and the high cost of data annotation, and maximizing the learning effect under limited resources (Liu et al., 2021; Tang, Chen, Luo and Zhang, 2021; Yang, Wei, Wang, Hua and Zhang, 2021). Besides, by learning from large amounts of unlabeled data, semi-supervised models are also able to capture richer and more general data representations, thereby demonstrating better generalization performance on a variety of tasks.

In this thesis project, we created the waste classification model from three aspects: Dataset, data augmentation, and model structure. Firstly, two waste datasets named WasteData and WasteNet, which contain 1,560 and 1,326 images respectively, are collected. Secondly, non-uniform data augmentation methods that meet the characteristics of waste classification is developed. Finally, the semi-supervised learning structure is studied and applied to the waste classification model. Our method improves the efficiency of waste classification.

1.2 Research Questions

Waste classification is one of the most important ways to protect our environment. There is a great deal of benefits to dispose wastes. For example, it can improve sustainability and pollution, protect the ecological environment. Besides, it can enlarge land space and improve the utilization. The recycling of waste can also make effective use of natural resources. It is important that waste is disposed efficiently and cost-effectively, and automated waste disposal is one of the solutions. Since deep learning has achieved significant results on visual object classification, detection, and segmentation tasks, we also choose to use deep learning to study waste classification from digital images and videos. However, automatic waste classification has not been widely used. Even though many deep learning algorithms have been applied in the field of waste classification, there is still much room for improvement in terms of disposing efficiency and detection accuracy. Moreover, the collection of waste data is currently challenged by a wide variety of waste types. Therefore, it is necessary to construct an efficient deep learning model and collect effective waste data to improve the efficiency

of waste classification. Therefore, our research questions are as follows:

Question 1:

How does a deep learning model distinguish between different types of waste (such as hazardous waste, recyclable waste, wet waste, dry waste)?

Question 2:

How do different deep learning architectures perform in waste classification?

Question 3:

How to improve the accuracy of deep learning models for waste classification in complex backgrounds?

Question 4:

How to improve the accuracy of waste classification in the case of limited or imbalanced data?

1.3 Contributions

In this thesis, we firstly validate the potential of the waste datasets to improve waste classification using deep learning. In the later stage, the focus of this thesis is on how to use less data to obtain better waste classification performance. With this in mind, we improve the accuracy of our proposed model in identifying waste by improving data augmentation methods and semi-supervised learning methods.

1.3.1 Datasets

In this thesis project, we collected two waste datasets. The first dataset is WasteData. We selected a number of waste images from a large number of images and divided them into four categories, totaling 1,560 images. According to the waste categories, we manually set the labels of the images as “recyclable”, “hazardous”, “wet”, and “dry”, as

shown in Figure 1.1. Besides, each class contains a different kind of waste. As an example, the recyclable waste category includes glass, cardboard, and plastic. To enrich the dataset, we also annotated multiple perspectives of the same object as well.

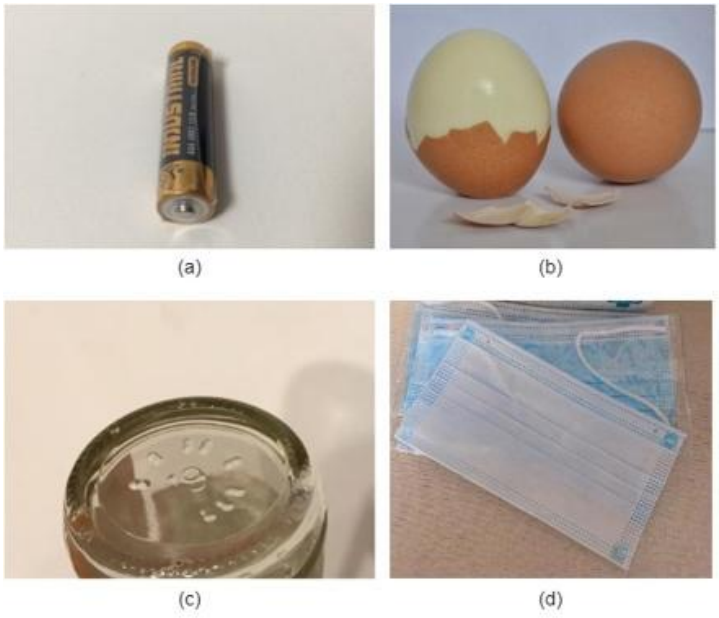

Figure 1.1. The example of waste image in WasteData

Figure 1.2. The example of waste image in WasteNet

The second dataset is called WasteNet, and the waste images in the dataset are all taken by ourselves. WasteNet contains 1,326 images and four categories of waste classified according to waste classification standards, namely “recyclable waste”,

“hazardous waste”, “dry waste”, and “wet waste”. We employed the LabelMe annotation tool to manually annotate images. Each image contains multiple types of waste, so the number of labeled wastes in the dataset exceeds the number of images. Some waste data images are shown in Figure 1.2. From Figure 1.2, our dataset details the stacking phenomenon between waste objects, and some waste is also deformed, which is consistent with real waste classification scenarios.

Simultaneously, we also utilized the world’s first waste dataset employed in industry, called ZeroWaste (Bashkirova et al., 2022). This dataset contains approximately 6,212 images, which were taken in real waste classification scenes. It vividly reflects a variety of waste forms and polluted waste environments, significantly enriching the diversity of the waste dataset. Figure 1.3 illustrates some images from the ZeroWaste dataset. Although the ZeroWaste dataset has the advantages of rich data and large data volume, the reason why this thesis still uses the WasteNet dataset is that the ZeroWaste dataset only includes recyclable waste. In comparison, the WasteNet dataset contains a very comprehensive range of waste categories, which can not only help with waste collection, but also identify hazardous waste or non-recyclable waste.

Figure 1.3. The example of waste image in ZeroWaste

1.3.2 Data Augmentation

How to reduce the negative impact on waste classification models caused by the scarcity of waste datasets and the complexity of waste data scenarios? How to make up for the shortcomings of waste datasets? How to alleviate the arduous work of waste data collection and annotation? These are our concerns. Through research, some existing data augmentation methods have different improvement directions, such as ignoring part of the pixel information of the image, resulting in the loss of important features. Therefore, this thesis not only adopted some existing advanced data augmentation methods, such as cropping, rotating, flipping, translation, CutMix (Yun et al., 2019), Mixup (Zhang, Cisse, Dauphin and Lopez-Paz, 2018), and Mosaic Augmentation (Bochkovskiy, Wang and Liao, 2020), but also designed a non-uniform data augmentation method and an adaptive weighted loss function for high-complexity real waste classification scenarios. Our method not only improves the accuracy of waste classification, but also saves the cost of manual annotation. Furthermore, the adaptive weighted loss function can make up for the imbalance of results caused by the large difference in the number of samples of the two data categories.

1.3.3 Semi-supervised Learning

In practical applications, obtaining a large amount of effective data and labeling it will consume expensive labor costs. Semi-supervised learning has great potential to solve this problem. Semi-supervised learning can improve the performance of learning models by incorporating unlabeled data, especially when labeled data is scarce. Therefore, in this thesis, we propose a novel approach: Co-Iteration Semi-Supervised Learning (CISO) to provide a more flexible and adaptable semi-supervised learning framework. This framework allows the model to be trained for multiple iterations (IoU values greater than the mean IoU value) and retain pseudo-label data during the first iteration. Moreover, although we set the data to make the IoU value greater than the mean IoU value for training, this does not mean that the mean IoU is a fixed value.

Because the pseudo-label is updated every iteration of the model, it means that the mean IoU value is also constantly changing, so that unlabeled data can be fully utilized.

1.4 Objectives of This Thesis

This thesis introduces basic deep learning methods and various advanced deep learning models for waste classification in detail, and conducts a comprehensive evaluation and comparison of their results. The characteristics and performance of these deep learning models are explored in detail to demonstrate their effectiveness and limitations on waste classification tasks. Secondly, to achieve efficient and effective waste classification tasks, this project collects new waste datasets. Finally, this thesis also proposes the deep learning methods suitable for waste classification for further understanding.

Therefore, focusing on the existing difficulties in waste classification, the main objectives of this thesis are divided into five parts, covering the entire process from waste data collection to model training. The first is the collection of waste data. The key is to collect diverse waste samples and ensure the quality of training data. The second is data augmentation. The nonuniform data augmentation is introduced to increase the data diversity during model training and reduce the risk of overfitting caused by the lack of waste samples. The third part is the innovation of deep learning methods. The main research content is related to semi-supervised learning to improve the accuracy and efficiency of waste classification. The fourth part is model training, including selecting appropriate parameters to adapt to the specific dataset and classification task. Finally, the experimental results analysis section will evaluate the performance of the model, which includes a detailed comparison and analysis of different deep learning methods and our proposed method specifically for the waste classification task.

In addition, this thesis will use a variety of performance metrics to verify the effectiveness and robustness of different methods. Through these comparisons and analyses, we hope to provide more effective and efficient solutions to the waste classification problem, thereby promoting the development of this field.

1.5 Structure of This Thesis

The structure of this thesis is detailed as follows:

In Chapter 2, we present a detailed literature review, including traditional deep learning methods and their applications, and critical analysis of these fields. Based on relevant literature research results, this chapter mainly discusses different deep learning methods, data augmentation methods, semi-supervised learning strategies, and various waste datasets, and introduces the current advanced waste classification models based on deep learning.

In Chapter 3, we explain the datasets and evaluation metrics used in this thesis, as well as the new waste datasets, proposed in this thesis. It mainly includes the following four aspects: data collection methods and processes, data augmentation, and evaluation metrics.

In Chapter 4, we utilize traditional deep learning methods to classify waste, and summarizes the advantages and limitations of these neural network models.

In Chapter 5, taken characteristics of the complexity and diversity of waste datasets, new data augmentation methods and loss functions are proposed, called non-uniform color data augmentation, non-uniform offset data augmentation, and adaptive weighted loss function, respectively.

In Chapter 6, potential solutions are elaborated. Semi-supervised learning has the advantages of reducing labeling costs, improving model efficiency, and enhancing model generalization capabilities. It can effectively reduce the negative impact of waste classification tasks due to the wide variety of waste types and the lack of waste datasets. Therefore, this chapter innovates a semi-supervised learning method, called Co-Iteration Semi-Supervised Learning (CISO).

In Chapter 7, we discuss the experimental results and compares them with other state-of-the-art model results. In addition, comprehensive ablation experiments are

designed to verify the effectiveness and application potential of the proposed method. This chapter includes three parts, namely, basic waste classification results, waste classification results based on non-uniform data augmentation, and waste classification results based on semi-supervised learning.

Finally, in Chapter 8 we summarize the methods and research results used in this thesis. Then, research limitations and future work are discussed.

Chapter 2
Literature Review

In this chapter, we will provide an in-depth exploration of the work related to the research theme of this thesis, such as the state-of-the-art deep learning methods, waste classification models, and waste datasets. Additionally, since this thesis also involves data augmentation and semi-supervised learning, this chapter will also analyze related work in these two fields to provide background information, development history, main theories, and research status of these fields.

2.1 Introduction

Solid waste is one of the top problems in the world today. Along with progress of economy and production activities in modern society, the amount of waste also generated gradually increases (Kang, Yang, Li and Zhang, 2020). Waste management is an increasingly important topic. It is predicted that the total amount of global waste will increase by one-fifth after 2030, with an imminent output of up to 2.6 billion tons of waste per year (Kaza, Yao, Bhada-Tata and Van, 2018). Some studies show that about 600 million tons of plastic waste still exist in the world's oceans. By 2020, there are still some countries where the waste recyclability rate will not be higher than 35% (Ferronato and Torretta, 2019; Mohanraj, Senthilkumar, Chandrasekar and Arulmozhi, 2023; Xiao, Dong, Geng and Brander, 2018). The disposal and recycling of waste is an indispensable topic in protecting the environment. How to deal with waste efficiently and reasonably is an increasingly arduous task. Waste is not waste, recycling, reducing, and reusing are strongly needed (Yang and Thung, 2016).

At present, inadequate waste management mechanisms pose a great challenge to the protection of the ecological environment, the improvement of public health, and the safeguarding of human health. For example, the current main methods for open waste dumps are landfilling or incineration. However, these two unhealthy ways of disposing waste is prone to pollute the environment (Mohanraj, Senthilkumar, Chandrasekar and Arulmozhi, 2023). The amount of waste is positively correlated with the number of carbon emissions, which will lead to environmental pollution and exacerbate global warming (Chen et al., 2020). In another way, in the landfill process, a large amount of harmful chemicals, such as acid, alkaline and other toxic substances, will be produced, causing groundwater pollution and crop yield reduction (Chen et al., 2020). This will affect the utilization of water resources and increase the risk of virus transmissions, in addition to becoming a breeding ground for pathogens, which can easily lead to the spread of infectious diseases, such as malaria and diarrhea (Amasuomo and Baird, 2016; Ferronato and Vincenzo, 2019; Zaman, 2015; Zhang, Tan and Gersberg, 2010; Zhang,

Hu, Zhang and Zhang, 2020).

Table 2.1 The summary of four classifications of waste and their characteristics.

Categories	Examples	Characteristics
Recyclable waste	• Pape • Glass • Plastic • Cardboard • Metal	Recyclable waste can be converted into raw materials through the recycling process and used again in production.
Wet waste	• Food scraps • Fruit peels	Wet waste can be disposed of through biodegradable methods such as composting.
Hazardous waste	• Battery • Medicines • Bulbs	Hazardous waste contains substances harmful to the environment and humans. This type of waste requires special handling to reduce harm to the environment.
Dry waste	• Ceramics • Cigarette • Mask	Dry waste is usually not easily recycled or biodegradable and needs to be disposed of by landfill or incineration.

There is a myriad of domestic wastes. To realize the harmless and reuse of waste, waste classification is an effective solution, improving the possibility of recycling and reuse (Meng and Chu, 2020). Typically, waste classification involves sorting various types of waste according to specific classification criteria and converting them into reusable resources. Proper waste disposal practices hold significant implications for ecological sustainability, resource efficiency, and public health enhancement. Currently, waste classification is a crucial step in waste management, aiding in resource recycling and minimizing resource depletion. It contributes to reducing the reliance on waste

incineration and landfills, thereby lessening pollution and safeguarding ecosystems, helping to achieve the goal of peaceful coexistence between human and nature. The composition, characteristics, value, and treatment methods of waste are the basic basis for waste classification. Waste can be divided into four major categories, as shown in Table 2.1.

Waste recycling can save living resources, reduce production costs, and alleviate environmental pollution, which is necessary for modern society. However, the implementation of waste classification at this stage faces a series of challenges. Traditional methods of waste classification, which are typically semi-manual or semi-automatic, struggle to keep pace with the increasing volumes of waste, often resulting in inefficient sorting and adverse health effects on workers. Furthermore, it is difficult to carry out household waste classification activities especially for the aged and children. They may not be able to fully comply with the waste classification rules due to mobility problems, resulting in low efficiency and accuracy of waste classification. Consequently, there is a pressing need to automate waste classification and incorporate more sophisticated technologies, such as artificial intelligence, into waste management (Gundupalli, Hait and Thakur, 2017; Pan, Li and Yan, 2018; Qi, Nguyen and Yan, 2022). The integration of advanced AI-driven classification techniques can lead to more effective, efficient, and health-conscious waste management practices (Pan et al., 2021; Qi, Nguyen and Yan, 2022). This, in turn, supports economic growth and environmental protection, stimulating us toward the sustainable coexistence of humanity and nature (Qiu et al., 2022; Shi, Tan, Wang and Wang, 2021).

2.2 Deep Learning

The development of deep learning has brought great scientific and societal convenience to our community, especially computer vision (Janiesch, Zschech and Heinrich, 2021; Joseph, Khan, Khan and Balasubramanian, 2021; Shen, Chen, Nguyen and Yan, 2018). There are plenty of algorithms for visual object detection in computer vision. They have been widely applied to industry, agriculture, and services (Papageorgiou and Poggio,

2000; Shrivastava, Gupta and Girshick, 2016; Wu, Sahoo and Hoi, 2020).

2.2.1 Basic Deep Learning Methods

Convolutional Neural Network (CNN)

CNN is an essential convolutional architecture in the field of deep learning, which mainly extracts features in images through linear operations (Cai, Fan, Feris and Vasconcelos, 2016; Cao et al., 2019; Dalal and Triggs, 2005; Fran, 2017; Fu, Sun, Wang and Fu, 2020; Fu, Li, Ma, Mu and Tian, 2020; Geirhos et al., 2020; Passalis and Tefas, 2018; Yan, 2021). Generally, CNN consists of four core parts, namely convolutional layer, activation layer, pooling layer, and fully connected layer, which has the advantages of translation invariance and reduced computational load (Abdel-Hamid et al., 2014; Almahairi et al., 2016; Ba, Kiros and Hinton, 2016; Liu, Yan and Kasabov, 2024; Simonyan and Zisserman, 2014; Xin, Nguyen and Yan, 2020). Currently, the visual object detection method using CNN as the backbone network can be grouped into two categories. Followed the classification of one-stage network and two-stage network, as seen in Table 2.2, the one-stage networks can extract visual features directly to classify visual objects, the model is fast and suitable for mobiles, but the accuracy is lower than that of the two-stage networks (Prince, 2012). They have four main types.

The core idea of YOLO is to apply the entire graph as the input of the network, and directly return to the position of a bounding box and the class to which the bounding box belongs in the output layer (Bochkovskiy, Wang and Liao, 2020; Le, Nguyen, Yan and Nguyen, 2021; Li, Xu and Yan, 2023; Rezatofighi et al., 2019). It is fast and can meet real-time requirements. By using the full image as context information, we are able to reduce the errors of detecting the background as an object, and it has a strong generalization ability (Cui et al., 2020; Li, 2016; Oliva and Torralba, 2007). The mAP is about 63%. But YOLO has minor errors. For example, each grid can only predict one object, which is easy to cause missed detection; and it is relatively sensitive to the scale of visual object. SSD improves these two aspects by using multiscale feature maps and

convolution for detection, bringing mAP to 71% (Liu et al., 2016).

Table 2.2 The summary of four different one-stage networks.

References	Networks	Approaches
(Girshick and Farhadi, 2016)	YOLOv1	YOLO is trained on a loss function that directly corresponds to detection performance and the entire model is trained jointly. Fast YOLO is the fastest general-purpose object detector in the literature and YOLO pushes the state-of-the-art in real-time object detection, bringing mAP to 63.4%.
(Liu et al., 2016)	SSD	Combine bboxes, use NMS to suppress some overlapping or incorrect bboxes. The data augmentation method also played a key role in the SSD algorithm, making mAP change from 65.5% to 71.6%.
(Najibi, Rastegari and Davis, 2016)	G-CNN	The object proposal stage in the CNN-based object detection framework is removed, and the object detection problem is modeled as an iterative regression problem.
(Kong et al., 2017)	RON	Use deconvolution to deconvolution from the last layer and connect the previous layer to make the feature map semantics of the previous layer richer. It also filters most negative samples and solve the problem of imbalance between positive and negative samples in the default recommendation box.

Table 2.3 The summary of four different two-stage networks.

References	Networks	Approaches
(Girshick, Donahue, Darrell and Malik, 2014)	RCNN	RCNN has three improvements: selection of candidate regions, CNN feature extraction, and classification. When a large amount of labeled data is lacking, neural network transfer learning can be performed, and then fine-tuned.
(He, Zhang, Ren and Sun, 2015)	SPP-net	An SPP layer is added between the last convolutional layer and the fully connected layer. The input of the network can be of any scale, and each pooling filter in the SPP layer will be resized according to the input to ensure that the input to the next fully connected layer is fixed.
(Ren, He, Girshick and Sun, 2015)	Faster R-CNN	The RPN algorithm is used to replace the original selective search method to generate candidate frames. The CNN network that generates the candidate frames and the CNN network for object detection are the same CNN network.
(He, Gkioxari, Dollár and Girshick, 2017)	Mask R-CNN	The proposed multi-task structure improves the performance of instance segmentation. In addition, RoIAlign solves the misalignment problem of the RoI pooling algorithm, and ensures that the original image is aligned with the feature map, and the pixel from the

		feature map to the ROI is aligned to improve the accuracy of object detection. This is a general framework for instance segmentation.

Regarding G-CNN, it is similar to the YOLO algorithm, but the focus is on the reduction of the number of initial proposals, so that a large number of proposals become very few initial grids, the final more accurate bbox is obtained through subsequent iterations (Najibi, Rastegari and Davis, 2016). The last one-stage network is RON (Kong et al., 2017), which solves the problem of the imbalance of positive and negative sample ratios, the detection results reached up to 77%.

Pertaining to one-stage network, Darknet-53 is utilized as YOLOv3-Darknet backbone network to dete0ct municipal waste (Cui, Zhang, Green, Zhang and Yao, 2019), with an average accuracy of 87.4%, which could also be detected in complex environments. However, the detection of small targets needs to be considered. In addition, L-SSD was proposed (Ma, Wang and Yu, 2020). ResNet-101 as the backbone network, its accuracy for object detection reaches 83.5%. The L-SSD model is able to detect small targets well, but the detection speed could be improved. Similarly, YOLOv3 was also used for waste classification (De Carolis, Ladogana and Macchiarulo, 2020). Due to the size of small dataset, the average accuracy of model training was 68%. Therefore, it is feasible to improve the accuracy of YOLOv3 model by increasing the number of images.

Another category is the algorithms based on two-stage network, as shown in Table 2.3. It has more steps than the one-stage network to generate region proposals based on anchor points, then merge the detected bounding boxes, and find the object locations. This kind of structures have slow speed but high precision.

The proposed R-CNN proves that the neural network can be applied to the candidate area from the bottom-up, so that the target classification and target positioning can be carried out (Girshick, Donahue, Darrell and Malik, 2014). It

increased mAP to 53.3%, which is an improvement of 30% from the previous results. Then, Faster R-CNN creatively employed the convolutional network to generate the suggestion box by itself, and shared it with the target detection network, reduced the number of suggestion boxes from the original ones from 2,000 to 300, the quality of the suggestion box is also essential improve. After that, based on Faster R-CNN, Mask R-CNN adds a mask branch to realize instance segmentation, and takes use of RoIAlign instead of RoI pooling to improve the accuracy of instance segmentation (He, Gkioxari, Dollár and Girshick, 2017). It is a general instance segmentation framework. Finally, SSP-net (He, Zhang, Ren and Sun, 2015) is propounded, which solves the problem that CNN needs to fix the size of the input image, so that it can be correctly transmitted to the network regardless of the scales of the input image.

Transformer Models

Most of CNN models are mature and have achieved good results in visual classification, detection, and segmentation tasks (Chen, Kornblith, Norouzi and Hinton, 2020; Romera-Paredes and Torr,2016; He, Fan, Wu, Xie and Girshick, 2020; He, Zhang, Ren and Sun, 2016). However, in the computer vision area, more publications were shifting from CNN to the Transformer model (Vaswani et al., 2017; Zagoruyko and Komodakis, 2016; Zhu, Cheng, Zhang, Lin and Dai, 2019). It is a recently proposed network based on the attention model. Transformer is the most primitive attention mechanism model, which is mainly employed in machine translation (Guo et al., 2022; Jaderberg, Simonyan and Zisserman, 2015; Li, Liu, Zhang and Cheng, 2020; Lieskovská, Jakubec, Jarina and Chmulík, 2021; Niu, Zhong and Yu, 2021).

Later, the combination of Natural Language Process (NLP) (Radford et al., 2019; Sakalle, Tomar, Bhardwaj, Acharya and Bhardwaj, 2021; Srivastava, Geoffrey, Alex, Ilya and Ruslan, 2014) and Transformer gradually deepened, making Transformer the mainstream model of nonlinear programming. This model does not rely on a convolutional neural network, encodes input, and calculates output based on attention mechanism. Transformer breaks the limitation that RNN models cannot be trained in

parallel compared to CNN. The Transformer also avoids the increase in the number of operations required to correlate between locations as the distance grows (Vaswani et al., 2017). It has six encoders and decoders with the same structure but different parameters, respectively. In the encoder, there are two modules. The first one is self-attention modules, and the second module is Feedforward Neural Network (FFNN). It is worth noting that the first module is the attention mechanism. After the data passes through the first module, weighted eigenvector attention is obtained, which is shown in Eq. (2.1).

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)v \tag{2.1}$$

where *Q, K, V* are query vector, key vector, and value vector, respectively, with a vector length 64. If the input data is transformed into embedded vectors, three vectors *q, k, v* are obtained. Then, the attention mechanism calculates the score for *q, k,* and *v,* that is *q* × *k*. To stabilize the gradient, it will be normalized, which is reflected in Eq. (2.1). Next, the score function is activated with the Softmax function and dot product (Mikolov, Kombrink, Burget, Černocký and Khudanpur, 2011; Yan, 2019; Luo, Nguyen and Yan, 2022). Finally, the output resultant attention thus is obtained.

While the encoder outputs data, the data enters the decoder. The decoder has one more step: "encoder-decoder attention" than the encoder after the self-attention step, mainly focusing on the feature vector. If the data is output from the decoder, it will become a real vector, which needs to go through the linear transformation and Softmax layers to become the final output data.

Moreover, another important concept is multi-head attention, which is the ensemble of *h* different self-attention. Its principle is to input the data into *h* to obtain characteristic matrices. Then, the characteristic matrices are introduced in the order of columns to form a new matrix, and finally obtain a new vector attention through a layer of FFNN. In addition to the encoder and decoder, there is also a part of data preprocessing. For positional encoding, the model uses sine and cosine functions with

different frequencies for calculation.

We see that Transformer has a few advantages, such as parallel computing and being more explanatory. However, the Transformer is much suitable for application in the NLP field than in the computer vision area. Thus, the proposal of Vision Transformer model makes the Transformer much widely in the field of vision. Compared with the Transformers, Vision Transformer has one more step, that is, the input picture is divided into patches with a size of 16×16 (Dosovitskiy et al., 2021). Each patch is input into the embedding layer to obtain the corresponding token. After that, we add a position embedding on each token and input it all into the encoder layer of the Transformer. The Transformer of each layer is applied to calculate the mutual self-attention between patches and passes them layer by layer. The frame of the Vision Transformer model is shown in Figure 2.1. It makes use of the Transformer to get ideal results in the applications of computer vision.

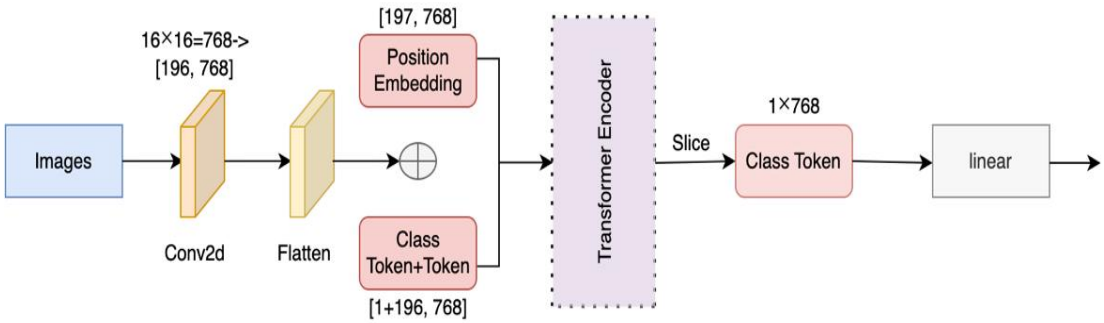

Figure 2.1. The framework of Vision Transformer

Subsequently, various Transformer derivative models have emerged and have widely used in the field of computer vision, such as DETR (Carion et al., 2020) and Swin Transformer (Liu et al., 2021). Furthermore, there are other kinds of Transformer models employed for visual object detection. For example, BoTNet makes use of the Transformer to enhance the remote dependency of the CNN backbone, then PiT, a kind of Hierarchical Transformer (Heo et al., 2021; Srinivas et al., 2021). The Transformer network-based model has tremendous advantages, such as parallel computing and more explanatory power.

Large Language Models

Large language model refers to deep learning models with a huge number of parameters and complex structures, which can contain millions or even hundreds of millions of parameters and are able to handle large-scale data, learn more complex features, and have more powerful generalization ability and higher accuracy (Atallah, Banda, Banda and Roeck, 2023; Kasneci et al., 2023; Kozachek, 2023; Liu, et al., 2023; Radford, Narasimhan, Salimans and Sutskever, 2018; Yang et al., 2019).

At present, large language models are gradually becoming the mainstream development direction in the field of deep learning (Thirunavukarasu et al., 2023; Waisberg et al., 2023). The large-scale multilingual and multimodal machine translation model named SeamlessM4T, which can perform speech translation between up to 100 languages (Barrault et al., 2023). Afterwards, a large language model that can role-play animated characters was proposed, called ChatHaruhi (Li et al., 2023). In addition to the domains of speech translation and role-playing chat robots, large language models also have excellent performance in writing code, modifying code bugs, and textual question and answer, such as the highly regarded ChatGPT model (ChatGPT, personal communication, 2023).

The large language models were explored in specific fields and utilize large language models for intelligent change in healthcare are of much interest. The ability to write postoperative patient discharge summaries and recognize images of patient lesions using GPT-4 was demonstrated to have the potential to aid medical innovation (ChatGPT, personal communication, 2023; Liu, 2023; Waisberg, et al., 2023). Subsequently, the application of GPT-4 in biomedical engineering has also been explored, and it has demonstrated excellent performance in the areas of medical devices, bioinformatics, and medical imaging (Cheng et al., 2023). Finally, large language model can also be applied to healthcare, such as providing users with healthcare-related information support for weight loss and mental health (Egli, 2023; Hendrycks and Gimpel, 2016). It is thus conjectured that GPT-4 has unlimited potential to help other

domains.

Large language models combined with vision tasks also gradually being developed. Continuous improvements in large language models have evolved their functionality from processing text to processing visual images, bringing significant benefits to many text-to-image interaction tasks.

Visual ChatGPT connects a range of visual foundation models into ChatGPT, enabling users to interact with ChatGPT in the form of text and images (Chen, Guo, Yi, Li and Elhoseiny, 2022). It also provides complex visual instructions that allow multiple models to work together. Visual ChatGPT can also understand and respond to both text-based and vision-based inputs, reducing the barriers to accessing text-to-image models. Then, Google proposed the multimodal visual language model PaLM-E, which has 562 billion parameters (Driess et al., 2023). Based on the language model, PaLM-E performs continuous observation, e.g., receives image or sensor data and encodes it into a series of vectors of the same size as the language token. In this way, PaLM-E can continue to understand sensory information in the same way it processes language. The success of these models validates the future possibilities of such multimodal models.

2.2.2 Data Augmentation

Data augmentation is an important method in deep learning, which refers to improve the diversity, richness, and data volume of a limited data set without increasing the amount of original data, so that the neural network can learn a wider range of image features (Liang and Yan, 2022; Xin, Nguyen and Yan, 2020; Zoph et al., 2020). In general, CNNs have invariance to operations such as image size and displacement, so data augmentation can help neural networks to understand and interpret the performance of the same object under different positions and scaling conditions (Pan and Yan, 2018). Therefore, data augmentation is widely used in image processing tasks, such as image classification, object detection, and semantic segmentation (Yu et al., 2018).

Currently, data augmentation can be grouped into two categories: Supervised data

augmentation and unsupervised data augmentation. Supervised data augmentation is mainly used on labeled data, and it can be further classified into single-sample data augmentation and multiple-sample data augmentation. Among them, single- sample data enhancement focuses on operations on a single data sample, including flipping, rotation, random cropping, translation, adjusting image saturation, Elastic deformation, scaling, color transformation, Jitter, noise injection, Cutout (DeVries and Taylor, 2017; Krizhevsky, Sutskever and Hinton, 2023; Liu et al., 2022; Redmon and Farhadi, 2017; Wan, Zeiler, Zhang, Cun and Fergus, 2013). These techniques simply change the appearance of images to increase the size of the dataset (Liu, Neuyen and Yan, 2020). Multi-sample data augmentation uses multiple sample data to generate new data samples. Representative technologies include CutMix, Mixup, Mosaic augmentation, and SamplePairing (Bochkovskiy, Wang and Liao, 2020; Yun, 2019; Zhang, Cisse, Dauphin and Lopez-Paz, 2018). CutMix refers to cutting a certain part of one image at the corresponding position in another image, and its labels are also mixed. After that, Mixup mixes two different images according to the proportion to form a new training sample. Then, Mosaic augmentation stitches four different images into a new image. Finally, SamplePairing combines the average of the pixel values of two different images to generate a new sample. These methods enhance the model's ability to adapt to data variations by introducing more diverse data combinations.

The other category is unsupervised data augmentation, which is mainly suitable for unlabeled data and is represented by methods such as Generative Adversarial Networks (GAN) and AutoAugmentation (Cubuk, Zoph, Mane, Vasudevan and Le, 2019; Karras, Laine and Aila, 2019). GAN consists of a generator and a discriminator, which generates new images through an adversarial process. AutoAugmentation is different from traditional data augmentation methods. It can automatically perform data augmentation operations through search algorithms, focusing on optimizing existing data.

By performing data augmentation, the neural network model can avoid learning many irrelevant features and reduce the risk of overfitting during the training process,

thereby improving the overall performance of the model and enhancing the generalization ability and robustness of the model. The most important thing is that data augmentation can adapt to different task requirements and reduce the cost of acquiring new data and annotating data. It is a key strategy in the field of deep learning.

Among these data augmentation methods, we take use of the recently popular Mixup method as an example for in-depth study. The main idea of Mixup is to use linear interpolation to randomly merge two different images to form a new image, which helps to reduce the risk of model overfitting. Assume two training samples A and B, let "X_A" and "X_B" be the labels corresponding to these two samples respectively. We select λ as the random parameter, ranging from 0 to 1.0. The calculation method of Mixup is as shown in Eq. (2.2). and Eq. (2.3).

$$X_{A\,1} = \lambda * X_A + (1 - \lambda) * X_A \tag{2.2}$$

$$X_{B\,1} = \lambda * X_B + (1 - \lambda) * X_B \tag{2.3}$$

where '*' represents the multiplication of corresponding pixels, and '+' illustrates the weighted sum of probabilities of corresponding category labels.

It can be seen from the Mixup method that since it generates new images through linear interpolation, it increases the diversity of data, reduces label noise, and improves the robustness of the model when the data is limited. For other state-of-the-art data augmentation methods, although the performance of Cutout and Cutmix is slightly inferior to Mixup, they also have the same advantages as Mixup. However, there is room for improvement in these data augmentation methods. For Mixup, some pixel information of the original image may be lost. This may have a negative impact on tasks that require retaining the detailed features of the sample. Cutout simply blocks some pixels of the sample, such as creating a black rectangular part. This may cause the model to fail to learn complete image information, affecting the model training effect.

Similarly, Cutmix makes use of another image to fill the occluded parts of the

original image based on Cutout. This can result in pixel discontinuities in new samples. In addition, since the new sample is generated by mixing two images, more annotation information is required, increasing the cost of data annotation. For waste classification tasks, the waste data has problems such as high data labeled costs and more accurate image information required for model training. This project speculates that the potential to utilize these data augmentation methods such as Cutout to improve model accuracy and efficiency is limited. Therefore, this thesis attempts to study and improve these two aspects in Chapter 5.

2.2.3 Semi-supervised Learning

In machine learning, self-supervised learning, unsupervised learning, and semi-supervised learning are three different learning paradigms. Supervised learning requires a large amount of manually labelled data to train the model (Atito, Awais and Kittler, 2021; Caron et al., 2021; Chen, Xie and He, 2021; Jing and Tian, 2020; Kolesnikov, Zhai and Beyer, 2019; Wang and Gupta, 2015; Sermanet et al., 2018). Labelling data and predicting results are taken into consideration before performing actions such as the calculation of gradients, after which learning is continuously performed as a way to obtain the ability to identify new samples, such as classification and regression tasks (Luo, Yan and Nguyen, 2022; Xiao, Nguyen and Yan, 2021).

However, unsupervised learning does not require labelled data which looks for relationships among data through its features as data clustering task. It mainly makes use of a pretext to construct supervised information from large-scale unlabeled data and obtains a pre-training model by training the network, then commences fine-tuning for the parameters obtained through transfer learning, to get useful visual representation (Bhunia et al., 2021). We see that the main difference between supervised learning and unsupervised learning is whether or not the labelled data or ground truth is required. Recently, contrastive learning methods are more popular in unsupervised learning for visual tasks. For example, SimCLR and MoCo series both utilize the contrastive learning method for research (Chen, Kornblith, Norouzi and Hinton, 2020; He, Fan, Wu,

Xie and Girshick, 2020). The contrastive learning method makes use of Siamese net and the core idea is to construct a positive sample and negative sample so that the distance between the sample and the positive sample is much greater than the distance between the sample and the negative sample (Qi and Su, 2017). These unsupervised learning models using a contrastive learning method have achieved impressive results. However, the introduction of the BYOL model eliminated the role of the negative sample (Grill et al., 2020). After discarded the negative sample, it also trained an unsupervised learning model that was superior to other models.

Besides data augmentation, another important function of unsupervised learning is the loss function (Cheng and Wang, 2019; Gonzalez and Miikkulainen, 2020). For the visual field, one of the loss functions of unsupervised learning is InfoNCE, which belongs to contrast learning and is applied to measure similarity (Oord and Vinyals, 2018). Its design is to crop a region from an image. If the cropped image is from the same one for another image, then the image is considered as a positive sample of the current image, otherwise, it is regarded as a negative sample. The sampled image is therefore represented by the query. At the same time, all the images will be saved to form a set of images and merged into a dictionary. The feature of these images is the key issue. Therefore, the InfoNCE is obtained as shown in Eq. (2.4).

$$L_q = -log \frac{exp(q \bullet k_+ / \tau)}{\sum_{i=0}^{K} exp\,(q \bullet k_i / \tau)} \tag{2.4}$$

where q is expressed as the feature vector of the query, $k+$ indicates a positive sample vector of q in the dictionary, k_i is the feature vector of the key in the dictionary, τ is a hyperparameter vector which is employed to adjust the loss. InfoNCE obtains similarity by dot-multiplying query with each key, the result of dot multiplication is controlled by the τ coefficient. By comparing positive and negative samples, it helps capture the structure of the data and can effectively learn data features. Therefore, the loss function also plays a key role in unsupervised learning.

Unsupervised learning can make full use of unlabeled data and reduce the cost of

data annotation. Similarly, for waste classification tasks, if the model dependence on labeled data can be reduced, the efficiency of waste classification can be greatly improved. However, unsupervised learning is not the optimal choice. In the waste classification task, the model needs to know the type of classification required in the image, and unsupervised learning is more suitable for the task of discovering the intrinsic structure of the data. At this time, the advantages of semi-supervised learning are revealed (Zhai, Oliver, Kolesnikov and Beyer, 2019). Waste classification tasks usually involve complex data samples. Using a small portion of labeled data can allow the model to better capture sample features.

Semi-supervised learning combines supervised learning and unsupervised learning techniques, and can use both labeled and unlabeled data to promote model learning. The core idea of semi-supervised is to generate a large number of pseudo-labels of unlabeled samples from a small number of labeled samples to provide additional supervision signals, the amount of labeled data is much less than that of the unlabeled data. In semi-supervised learning, two popular strategies are consistency regularization and entropy minimization (Cascante-Bonilla, Tan, Qi and Ordonez, 2021; French, Laine, Aila, Mackiewicz and Finlayson, 2019; Mahajan et al., 2018). The goal of consistency regularization is to ensure that the model can produce consistent prediction results even when the data is perturbed (Pham, Dai, Xie and Le, 2021; Xie, Dai, Hovy, Luong and Le, 2020).

Then, entropy minimization, from a certain perspective, serves consistency regularization. The loss function of consistency regularization can be designed by the concept of entropy minimization. Entropy minimization can calculate the entropy of the prediction results of unlabeled data. The lower the entropy, the more credible prediction of the samples (An and Yan, 2021; Chen, Jin, Jin, Zhu and Chen, 2021; Gong, Wang and Liu, 2021). These two paradigms can aid semi-supervised learning make better use of unlabeled samples, improve the generalization ability and robustness of the model, and solve the problem of limited data.

Currently, semi-supervised learning methods are applied to many visual tasks (Arazo, Ortego, Albert, O'Connor and McGuinness, 2020; Hinton, Vinyals and Dean, 2015; Iscen, Tolias, Avrithis and Chum, 2019). The CSD method takes use of the idea of contrastive learning to train the model from unlabeled samples and generate pseudo-label (Jeong, Lee, Kim and Kwak, 2019). This method belongs to soft pseudo-labeling. Corresponding to hard pseudo labels, a typical representative method is STAC (Sohn et al., 2020). The STAC method introduces an active correction strategy and combines pseudo-labeling and self-supervised learning to improve the performance of semi-supervised learning models. Afterwards, the instant-teaching method emphasizes real-time training of labeled samples, unlabeled samples, and pseudo-label, which is helpful for the optimization of semi-supervised learning (Zhou, Yu, Wang, Qian and Li, 2021).

Finally, the unbiased teacher method was also proposed to improve the bias problem caused by the generation of pseudo-label (Liu et al., 2021). Although these methods have improved on the key point of pseudo-label, there are still many challenges in the generation of pseudo labels. The confidence of pseudo-label affects the performance of the model in two aspects. At first, excessive confidence can easily lead to model overfitting. Second, pseudo-label with insufficient confidence will introduce noise and affect the training result of the model. Based on semi-supervised learning, which can improve the results of waste classification, this thesis will propose potential solutions to the previous two problems in Chapter 6.

Another concept to understand about semi-supervised learning is semi-supervised semantic segmentation. Semantic segmentation is different from object detection. It assigns detailed semantic information (such as the category and position of the pixel) to each pixel of the image. It is a deep learning method that can be accurate to the pixel level (Miyato, Maeda, Koyama and Ishii, 2018; Pham, Dai, Xie and Le, 2021; Xie, Dai, Hovy, Luong and Le, 2020; Wang, Cai, Liang and Ye, 2020). Semantic segmentation has the advantage of being more sensitive to the size, position, and orientation of objects in images, and has demonstrated significant value in many application scenarios,

such as autonomous driving, plant leaf disease identification, medical image analysis, and Augmented Reality (AR). However, the challenge of achieving pixel-level segmentation is greater. This is because each pixel needs to be accurately labeled, and the cost of data labeling will substantially raise. In addition, the problem of model overfitting will also occur, causing the performance of the model to decline on the test set.

Based on the existing problems, the proposal of semi-supervised semantic segmentation provides new possibilities for this field. Similar to semi-supervised learning, semi-supervised semantic segmentation only labels part of the data and then processes other data in an unsupervised or semi-supervised manner. This can significantly reduce the labor cost of labeling data.

Generative adversarial networks (GAN) emerged in the early field of semi-supervised semantic segmentation. Its characteristic is that it can continuously improve the generator and discriminator. GAN can not only extract valuable features from unlabeled data, but also distinguish generated data from real data (Goodfellow et al., 2020), thereby effectively improving the robustness of the model (Hung, Tsai, Liou, Lin and Yang, 2018; Ouali, Hudelot and Tami, 2020). Currently, more semi-supervised learning strategies are applied in semi-supervised semantic segmentation, including consistency regularization and entropy minimization (Cascante-Bonilla, Tan, Qi and Ordonez, 2021; Chen, Jin, Jin, Zhu and Chen, 2021; Gong, Wang and Liu, 2021). They improved the performance and efficiency of semi-supervised semantic segmentation to varying degrees.

Besides, data augmentation technology is also indispensable (French, Laine, Aila, Mackiewicz and Finlayson, 2019). By performing transformation operations on images (such as rotation and scaling), the diversity of data is enriched, training samples are expanded, and the model is helped to better handle different images.

In this thesis, we also explore the application of semi-supervised semantic segmentation on waste classification tasks.

2.3 Waste Classification

The rapid development of computer vision has led to the rise of many fields, such as intelligent driving and medical diagnosis (Gedara, Nguyen and Yan, 2023; Ji, Liu, Yan and Klette, 2019; Rabano, Cabatuan, Sybingco, Dadios and Calilung, 2018; Zhang et al., 2021). Its emergence has significantly improved the efficiency of tasks that require processing large amounts of image or video data (Vallayil, Nand, Yan and Allende-Cid, 2023). The waste classification task requires processing a large amount of complex waste data. Therefore, combining computer vision with waste classification makes waste classification gradually automated, harmless, and efficient, which is of great value and significance to the ecological environment and social economy.

2.3.1 Waste Datasets

Visual object detection has made remarkable progress in recent years, which is essential to use deep learning to replace manual labor for an automated approach to waste classification (Ji, Liu, Yan and Klette, 2019). However, the accuracy of waste classification models is lower than that of other classification task, such as fruit classification. We speculate that the reason for this performance difference is that the waste data has complex and dirty backgrounds and a wide variety of objects, which results in low quality of the waste dataset. However, the dataset is an important factor affected the training of the algorithm model, and a low quality of the dataset can seriously affect the accuracy of the model prediction. Therefore, how to improve model performance by improving the quality of waste datasets is needed.

In recent years, many waste datasets have been created. Labeled Waste in the Wild Dataset contains 1,002 data (Sousa, Rebelo and Cardoso, 2019). In this dataset, the collected waste objects based on food trays in shopping malls and homes, including plastic bottle, glass bottle, paper napkin, and metal can. Therefore, the collection of this dataset is not limited to the laboratory, the images taken in the real environment contain the conditions that exist in the real world, such as lighting, increasing the authenticity of

the sample. However, this dataset focuses on waste in trays and lacks many waste types such as batteries and sponges. This makes the dataset lack some diversity.

Another TrashNet dataset focuses on the recyclable waste category and contains a large number of labeled recyclable waste samples, with more than 2,500 images (Yang and Thung, 2016). The characteristics of this dataset are that data are collected under controlled conditions, all image backgrounds are white, and each image contains only one waste object. Figure 2.2 shows an example. Although TrashNet contains a comprehensive range of recyclable waste categories: Metal, glass, plastic, paper, cardboard, and garbage, this dataset only has one category of recyclable waste, and cannot train the model to identify other waste categories.

Figure 2.2. An example image of TrashNet dataset

Then, UAVVaste dataset released (Kraft, Piechocki, Ptak and Walas, 2021). This dataset is different from other waste datasets in that it utilizes Unmanned Aerial Vehicles (UAVs) to fly in outdoor environments (cities) and capture waste images at different heights, angles, and perspectives. Among them, there are 770 waste samples and approximately 3,700 waste objects collected. UAVVaste collected wastes from the perspective of aerial photography, restores the true state of the waste, and can provide the background environment where the waste is located, providing diversified data. It is highly beneficial for tasks that require waste management's outdoors (such as some

areas that are difficult for humans to access or require a lot of cost to access). Some examples are illustrated in Figure 2.3.

Figure 2.3. An example image of UAVVaste dataset

Figure 2.4. An example image of TACO dataset

Finally, Trash Annotation in Context (TACO) dataset is more popular (Proença and Simoes, 2020). It has a total of 1,500 waste images, and each image contains multiple waste objects, with approximately 60 annotation categories and 4,800 annotations. Furthermore, compared to UAVVaste, TACO's samples have richer

backgrounds, such as city streets, indoor environments, woods, and tropical beaches. Therefore, the TACO dataset is more suitable for on-site waste detection scenarios. Figure 2.4 shows a sample image of TACO.

By providing high-quality and diverse data, it contributes to the development of automated waste classification technology, thereby helping to alleviate environmental pollution problems.

2.3.2 Waste Classification Using Deep Learning

Waste detection is becoming popular. Using deep learning for waste classification has many advantages, such as scalability, high accuracy, and convenience (Altikat, Gulbe and Altikat, 2022; Fan, Cui and Fei, 2023; Huang, He, Xuan and Huang, 2020; Olugboja and Wang, 2019). The waste classification models based on deep neural networks have also been continuously proposed (Zhou et al, 2022; Ziouzios and Dasygenis, 2019; Zhang, Chen, Yang and Gong, 2021).

Lightweight network MobileNet-v2 was trained for waste classification with an accuracy 82.92% (Yong, Ma, Sun and Du, 2023). After that, the EnCNN-UPMWS model generated by combining CNN with an unequal precision measurement weighting strategy improves the waste classification accuracy to 92.85% by the two key points of weight coefficients and predicted probability vectors (Hua and Gu, 2021). Although EnCNN-UPMWS model can improve the performance of the ensemble learning model, it is trained based on the TrashNet dataset. This results in the waste features learned by the model being based on simple backgrounds, which may affect the robustness of the model.

Then, the improved Mask Scoring R-CNN algorithm based on Mask R-CNN was applied to detect waste and achieved an accuracy of 65.8% (Li, Yan and Xu, 2020). Although the accuracy of the Mask Scoring R-CNN algorithm in waste classification is 36.5%, higher than that of general objects, the model needs a large number of datasets for training, and the training results need to be improved. Besides, Faster R-CNN model

has been proposed to carry out object detection for waste classification. The results show that the accuracy of the model is 89.7% (Nie, Duan and Li, 2021). However, Faster R-CNN model could also be improved in the object detection of small target waste.

Next, a method based on the combination of CNN model with metal detector and recorder was proposed (Funch, Marhaug, Kohtala and Steinert, 2021). It detects metal and glass in waste bags with an accuracy of 98.0%. Although the detection accuracy of this method is very high, it utilizes a metal detector and recorder, which are not convenient that still needs a multi-sample dataset for training.

In addition to image classification, an object detection system based on YOLOv3 for real-time identification of waste in video streams was proposed, and accuracy 68% was obtained (Carolis, Ladogana and Macchiarulo, 2020). The advantage of YOLOv3 model is that it can apply cameras for real-time waste monitoring, and it can be employed in waste bins or waste classification factories in the future. However, the training of this model is still based on the existing waste data in the network, and there is an imbalance in the number of images and the number of annotations, which affects the performance of the model. Hence, a lightweight network-based waste classification model LW-GCNet was proposed. It utilizes depthwise separable convolution for feature extraction and adopts adaptive maximum pooling to reduce the number of parameters, and the waste detection accuracy reaches 77.17% (Xia, Xu and Tan, 2022). Finally, ResNet-34 algorithm was also applied to waste classification and an automatic classification bin was briefly designed, including the hardware structure (Kang, Yang, Li and Zhang, 2020).

Other than the proposed models based on the traditional CNN structure, the up-and-coming Transformer model is also applied to the waste classification task. Compared to CNN, the advantage of the self-attention mechanism, which is not limited by local interactions, allows Vision Transformer to achieve 96.98% in waste classification (Huang, Lei, Jiao and Zhong, 2021).

Although deep learning models for waste classification are constantly being improved and have obtained significant classification and detection results, there is still room for improvement (Adedeji and Wang, 2019; Ahmad, Khan and Al-Fuqaha, 2020). A slew of waste classification models, such as the optimized DenseNet-121 and ResNet-10 using fusion schemes, have waste classification accuracies as high as over 85% (Mao, Chen, Wang and Lin, 2021; Kashif, Khan and Al-Fuqaha, 2020). However, the datasets they use are only simple recyclable waste categories, such as glass, cardboard, plastic, paper, and metal, which cannot measure the real waste classification application scenarios. According to the waste classification standard, waste should be divided into four categories, namely “recyclable waste”, “wet waste”, “dry waste”, and “hazardous waste”. While the ETHSeg model classifies the four categories of waste based on X-rays, the classification accuracy of small objects in waste remains low (Qiu et al., 2022). Table 2.4 summarizes the characteristics of above waste classification models. Thus, the lack of a waste dataset and the intensive manual annotation due to the wide variety of waste categories is the important challenge faced by our algorithms in waste classification tasks.

Table 2.4 Comparative summary of different waste classification models.

References	Models	Summary	Characteristics
(Li, Yan and Xu, 2020)	Mask Scoring R-CNN	The improved Mask Scoring R-CNN based on Mask R-CNN and achieved an accuracy of 65.8%.	This model needs a large number of datasets for training.
(Carolis, Ladogana and Macchiarulo, 2020)	YOLOv3	Real-time identification of waste in video streams was proposed, and the accuracy 68% was obtained.	Real-time video waste classification provides a good start for the application of waste classification, but this model has the problem

			of data imbalance.
(Xia, Xu and Tan, 2022)	LW-GCNet	LW-GCNet is a lightweight network with a waste classification accuracy of 77.17%.	It uses adaptive maximum pooling to reduce the number of parameters and improve classification efficiency.
(Kang, Yang, Li and Zhang, 2020)	ResNet-34	This model briefly designs an automatic waste classification bin and uses the ResNet-34 algorithm.	This method further expands the application of waste classification in real society.
(Yong, Ma, Sun and Du, 2023)	MobileNet-v2	MobileNet-v2 is a lightweight network model that can achieve 82.92% accuracy when used for waste classification.	The lightweight network model can be applied to mobile terminals, such as mobile phones, to realize the function of real-time waste classification.
(Nie, Duan and Li, 2021)	Faster R-CNN	Faster R-CNN model has been proposed to carry out waste classification. The results show that the accuracy is 89.7%.	The model does not perform well in identifying small waste objects in waste datasets.
(Mao, Chen, Wang, Lin, 2021)	DenseNet-121	DenseNet-121 divides the data into two stages for processing. First, it roughly identifies the data, and then extracts	DenseNet-121 is also trained based on TrashNet dataset and cannot measure real waste classification

		features from possible target areas in the data.	scenarios.
(Kashif, Khan and Al-Fuqaha, 2020)	ResNet-10	This model uses a feature fusion scheme to achieve a waste classification accuracy of 94.58%.	This model validates the advantages of the fusion model.
(Hua and Gu, 2021)	EnCNN-UPMWS	By introducing UPMWS to design the weight coefficient, using the TrashNet dataset, the accuracy is 92.85%.	It can improve the performance of the ensemble learning model. However, the waste features learned by the model being based on simple backgrounds, which may affect the robustness of the model.
(Huang, Lei, Jiao and Zhong, 2021)	Vision Transformer	The Vision Transformer model uses a self-attention mechanism, and the accuracy of waste classification reaches 96.98%.	Waste classification based on Vision Transformer avoids being limited by the receptive field (characteristics of the CNN model), and the classification accuracy is better than other models using CNN architecture.

(Funch, Marhaug, Kohtala and Steinert, 2021)	CNN	This method based on the combination of CNN model with metal detector and recorder.	The detection accuracy of this method is high. However, the application of metal detector and recorder is not convenient.
(Qiu et al., 2022)	ETHSeg	ETHSeg model classifies the waste based on X-rays.	It looks at the waste classification model from a new perspective, and the use of X-rays greatly improves model performance.

Chapter 3
Datasets and Evaluation Metrics

The core idea of this chapter is a detailed description of the dataset collection and preparation process. After that, the data augmentation technology used in this thesis is briefly explained. Finally, the model evaluation metrics are also shown.

3.1 Data Collection

Datasets are crucial for training and improving performance of deep learning algorithms (Everingham, Van Gool, Williams, Winn and Zisserman, 2010). Deep learning models need to learn correct patterns from a large amount of data to classify and predict unclassified data. A suitable dataset should have multiple characteristics: Rich samples, accurate labels, sufficient scale, and high-quality images, which can help the algorithm better understand the nature of the deep learning problem. This section details the data collection methods and processes, including the sources of data and the tools and methods of data collection.

Many research groups have released waste datasets, the more popular ones are Labeled Waste in the Wild, TrashNet, UAVVaste, TACO, and ZeroWaste. Table 3.1 briefly summarizes the contents of these datasets.

In this thesis, the experiments on the direction of semi-supervised semantic segmentation are conducted based on the ZeroWaste dataset. The ZeroWaste dataset is the world's first industrial-level waste dataset. It is collected in a real waste classification factory scenario, as shown in Figure 3.1. We see that all waste is placed on the classification conveyor belt, and each image contains several waste objects. This dataset contains three sub-datasets, namely fully-supervised ZeroWaste-f, weakly-supervised ZeroWaste-w, and semi-supervised ZeroWaste-s. They have 4,503, 1,410, and 6,212 images respectively. Since we are applying this dataset to a semi-supervised learning algorithm, we are use of the ZeroWaste-s dataset. We chose the ZeroWaste dataset because it is the first one collected based on real waste classification industrial scenarios, and its advantage is that it can increase the authenticity and diversity of the dataset. However, this dataset only contains four types of recyclable waste: “Metal”, “Cardboard”, “Rigid Plastic”, and “Soft Plastic”, but the management and processing of other waste types such as hazardous waste, dry waste, and wet waste are equally important. Only using ZeroWaste will make the model have poor generalization ability or even be unable to identify the other three categories of

waste, which may have a negative impact on the waste classification results.

Table 3.1 The summary of the different datasets.

Datasets	Summary
Labeled Waste in the Wild	The collected waste objects based on food trays in shopping malls and homes, including plastic bottle, glass bottle, paper napkin, and metal can.
TrashNet	The data are collected under controlled conditions, all image backgrounds are white, and each image contains only one waste object. This dataset only contains recyclable waste samples.
UAVVaste	It utilizes Unmanned Aerial Vehicles (UAVs) to fly in outdoor environments (cities) and capture waste images at different heights, angles, and perspectives.
TACO	It contains multiple waste object and has richer backgrounds, such as city streets, indoor environments, woods, and tropical beaches. Therefore, this dataset is more suitable for on-site waste detection scenarios.
ZeroWaste	It is the first real-world industrial-scale waste dataset, including recyclable waste such as cardboard.

Therefore, we also collected our own waste datasets, they are WasteData and WasteNet. The purpose of collecting our own datasets is to make the model more robust and stable in identifying waste. Compared with other datasets, our datasets contain four

waste categories, namely recyclable waste, wet waste, dry waste, and hazardous waste, which meets the diversity and comprehensiveness of the data.

Figure 3.1 An image example of the ZeroWaste dataset

Figure 3.2 An image example of the WasteData dataset

For the WasteData dataset, we marked 1,560 images in total. Taken recyclable waste as an example, Figure 3.2 shows some of the recyclable waste images in this

dataset. The advantage of this data set is that waste objects are clearly visible and can be well labeled. In addition, our dataset complies with the waste classification standards and contains four waste categories, and each category is rich in different wastes. This greatly enriches the diversity of waste datasets. However, we also studied the ZeroWaste dataset and realized the shortcomings of our dataset, that is, it ignored the dirty background of waste. Therefore, we collected another waste dataset, named WasteNet.

In WasteNet, a variety of wastes in the image are stacked together, which not only meets the waste classification scene, but also meets the conditions for the diversity of waste forms, such as being distorted, compressed, and folded. Although WasteNet was not collected in a waste classification factory, apart from this problem, this dataset is infinitely close to real-world scenarios, this allows deep learning models to more accurately capture and understand the real relationships and patterns of the data. Moreover, considering that the waste classification factory is indoors, we also collected data indoors, as shown in Figure 3.3.

Figure 3.3 An image example of the WasteNet dataset

Although we believe that WasteNet can more accurately represent the real appearance of waste, enabling the model to have stronger generalization capabilities,

and effectively cope with complex real-world environments. However, this does not imply that the WasteData dataset should be disregarded. The WasteData dataset facilitates a more straightforward processing approach, can reduce interference during model identification, and achieve faster model training and verification. How to combine the advantages of these two datasets requires further research work. In the future, we will also expand our datasets so that each image fits the real cluttered waste classification background, and can ensure the effectiveness of feature extraction to further help waste detection tasks (Nixon and Aguado, 2019).

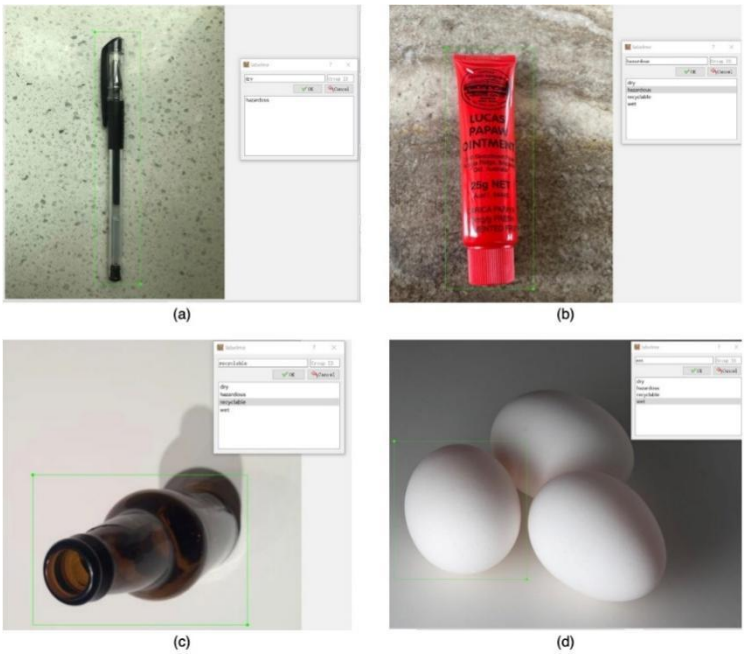

Figure 3.4 Visualization of annotation using Labelme

Moreover, the annotation of the dataset is completed with the help of Labelme software. We label the four classes of images as “Dry”, “Wet”, “Hazardous”, and “Recyclable”, respectively. For each class of waste image, the manual annotation is carried out based on the shape of visual object itself, which is shown in Figure 3.4. All images are marked and the bounding box (bbox) of the waste object in the image is found, which can also be called the region of interest (ROI). In this thesis, we take advantage of a rectangular frame to define the bbox, and the obtained information will be stored in JSON format. Figure 3.5 also shows an example of the ROI of a random

image in the dataset being stored in JSON format. However, in Chapter 4, some of our experiments used the YOLO series models. Therefore, the JSON format files need to be converted into txt version to be suitable for training of the YOLO series models. Take Figure 3.6 as an example. Finally, all data will be split into two groups, namely training set and test set.


```
{
    "segmentation": [
        [
            400.4745762711865,
            91.37288135593221,
            396.2372881355933,
            200.6949152542373,
            774.2033898305085,
            195.61016949152543,
            769.9661016949153,
            84.59322033898306
        ]
    ],
    "iscrowd": 0,
    "area": 41173.51335823038,
    "image_id": 26,
    "bbox": [
        396.0,
        84.0,
        378.0,
        116.0
    ],
    "category_id": 0,
    "id": 27
},
```


Figure 3.5 The JSON format for labeling


```
2 0.35859375 0.60234375 0.34765625 0.35703125
2 0.4875 0.52265625 0.91484375 0.59609375
2 0.465625 0.5046875 0.56875 0.61953125
2 0.27421875 0.6390625 0.49296875 0.42265625
2 0.38125 0.38359375 0.26015625 0.17421875
2 0.44921875 0.58515625 0.49140625 0.43359375
2 0.42734375 0.7734375 0.09140625 0.19140625
2 0.2890625 0.496875 0.38515625 0.35859375
2 0.32890625 0.39375 0.13125 0.18203125
2 0.74375 0.4984375 0.478125 0.69453125
2 0.21328125 0.35625 0.3640625 0.24375
2 0.32421875 0.45703125 0.3921875 0.29140625
2 0.5125 0.50703125 0.38203125 0.3609375
2 0.5796875 0.66796875 0.22890625 0.17265625
2 0.515625 0.59453125 0.28671875 0.2
```


Figure 3.6 The txt format for labeling

3.2 Data Augmentation

In addition to data collection, data processing is also a crucial step. The main purpose is to optimize the original data to improve the performance of the model. Generally, data processing covers many methods:

(1) Delete or fill data containing missing values

(2) Apply methods such as undersampling to deal with sample imbalance problems

(3) Utilize data augmentation to address problems with dataset class imbalance or small datasets

(4) Use smoothing algorithms to reduce image noise

(5) Remove duplicate data

Figure 3.7 Visualization of crop operating

Figure 3.8 Visualization of rotating operating

According to the characteristic that the number of samples in our datasets is lower than that of ZeroWaste dataset, data augmentation is the best way we choose to expand the dataset and enrich the diversity of training samples. Flipping, translation, cropping, and rotation are more popular data augmentation methods (Yan, Zhang, Wang, Paris and Yu, 2016). Therefore, in this thesis, our experiments also adopt these four data augmentation methods to improve model performance. Figure 3.7 and Figure 3.8 show a banana image after data augmentation.

The Crop method needs to randomly select an area in the original image and crop it out, but it should be noted that the height and width of the cropped sample should be smaller than the height and width of the original image respectively. The detailed pseudocode of the cropping method is introduced in Algorithm 3.1. In Algorithm 3.1, given an input parameter is the original image, define *randomCrop*() as a function, where the width and height of the original image and the cropped image are the parameters of this function. Then, return the height and width of the original image.

Next, we let *final_width* and *final_height* of the cropped image smaller than the original image respectively, where x represents randomly select a starting point for cropping along the width of the original image. Likewise, y refers to randomly selecting a starting point for cropping along the width of the original image. Finally, a new cropped image is generated and stored in the *newImage* variable and returned, where x and y determine the starting coordinate point of the upper left corner of the cropping area, and x+*final_width* and y+*final_height* represent the coordinate points of the lower right corner of the cropping area.

Algorithm 3.1 Crop method for images

Input: The original images

Output: The new images which changed by Crop method

```
def randomCrop(image, final_width, final_height):
    width, height = image.get_width(), image.get_height()
    assert final_width <= width and final_height <= height
    x = random_select(0, width – final_width)
    y = random_select(0, height – final_height)
    newImage = image.crop(x, y, x+final_width, y+final_height)
    return newImage
```

Algorithm 3.2 introduced the pseudocode of the translation method. The translation changes the position of the image content by randomly selecting some pixels in the image and moving them vertically or horizontally. In Algorithm 3.2, given an input parameter as the original image, we define *randomTranslation*() as a function, including three parameters: the original image, the maximum vertical movement distance, and the maximum horizontal movement distance. The second step is to get the width and height of the original image. After that, *moveX* and *moveY* represent randomly obtained horizontal and vertical moving distances respectively.

Note that the use of *max_x* and *max_y* does not mean that the moving distance is a negative value, but represents the difference in the moving direction. The fifth line of pseudocode creates a new image with the same size as the original, making it easy to copy the translated image on top. In addition, the width and height of new images cannot exceed that of the original image. Next, the two loop statements for *x* in *range*(*width*) and for *y* in *range*(*height*) function to traverse all pixels of the original image. Then, *newX* and *newY* calculate the new position information obtained by translation of each pixel. Finally, copy the image obtained through Translation to *newImage* and return it.

The last aspect to introduce is Algorithm 3.3, the pseudocode of the flipping method. Flipping plays the role of flipping the original image vertically or horizontally in data augmentation (usually flipping along the horizontal and vertical lines in the middle of the image), the model does not rely on image features in a specific direction.

In Algorithm 3.3, we define the flipping function with two parameters: The original image and the flipping direction.

After that, get the width and height of the original image. Let *newImage* be a new image of the same size as the original image. Iterate through all pixels of the original image through the loop function for *x* in *range*(*width*) and for *y* in *range*(*height*). Next, set up *if* and *elif* conditional statements to distinguish the flip type. Among them, horizontal and vertical represent vertical flipping and horizontal flipping respectively. Finally, copy the image obtained by Flip to *newImage* and return.

Algorithm 3.2 Translation method for images

Input: The original images

Output: The new images which changed by Translation method

```
def randomTranslation(image, max_x, max_y):
    width, height = image.get_width(), image.get_height()
    moveX = random_select(-max_x, max_x)
    moveY = random_select(-max_y, max_y)
    newImage = create_new_image(width, height)
    for x in range(width):
        for y in range(height):
            newX = x + moveX
            newY = y + moveY
            if 0 <= newX < width and 0 <= newY < height:
                newImage.set_pixel(newX, newY, image.get_pixel(x, y))
    return newImage
```

Algorithm 3.3 Flip method for images

Input: The original images

Output: The new images which changed by Flip method

```
def flip(image, flipDirection):
    width, height = image.get_width(), image.get_height()
    newImage = create_new_image(width, height)
    for x in range(width):
        for y in range(height):
            if flipDirection == "horizontal":
                newX = width – 1 – x
                newY = y
            elif flipDirection == "vertical":
                newX = x
                newY = height – 1 – y
            newImage.set_pixel(newX,  newY, image.get_pixel(x, y))
    return newImage
```

Data augmentation is one of the key strategies to improve the performance of deep learning models in the context of limited data. It has the capability to expand the dataset, effectively improve the robustness and generalization ability of the model, and reduce the cost and time of data collection.

3.3 Evaluation Metrics

The evaluation metrics of deep learning is a key tool for quantifying and comparing model performance, which can assist users to understand the performance of the model more easily. With these metrics, the strengths and weaknesses of the model can be found out and optimized and improved in a targeted manner. Therefore, choosing

appropriate evaluation metrics is crucial for model training and improvement. This thesis will employ reliable evaluation metrics that are adopted by a great deal of models, such as confusion matrix, Average Precision (AP), and mean Average Precision (mAP).

Most of these evaluation metrics are derived from the confusion matrix, which is represented by a matrix and includes both real and predicted two categories. The trained model is firstly employed to obtain a confidence score for all the test samples, and a set of confidence scores and ground truth labels are obtained for each category. This is summarized in Table 3.2.

Table 3.2 The confusion matrix.

		Prediction	
		Positive	Negative
Actuality	Positive	True Positive (TP)	False Negtive (FN)
	Negative	False Positive (FP)	True Negtive (TN)

T and **F** respectively indicate that the prediction result is true and the prediction result is false as the Boolean values, while P and N represent positive samples and negative samples respectively. Therefore, the definition and calculation method of each metric can be seen in Table 3.3.

Table 3.3 The definition of matrices.

Metric	Definition	Calculation method
TP	The proportion of positive samples that are predicted correctly	TP/(TP+FN)
FP	The proportion of positive samples that are predicted incorrectly	FP/(FP+FN)
FN	The proportion of negative samples that are predicted incorrectly	FN/(TP+FN)
TN	The proportion of negative samples that are	TN/(FP+TN)

	predicted correctly	

The meaning of Precision is how many of the results predicted to be positive samples are correctly classified, see Eq. (3.1).

$$Precision = \frac{TP}{TP+FP} \tag{3.1}$$

Eq. (3.2) details the calculation method of Recall. It is the number of correctly predicted positive samples accounting for the total number of positive samples.

$$Recall = \frac{TP}{TP+FN} \tag{3.2}$$

Intersection over Union (IoU) represents the ratio of the intersection and union of the prediction result of a prediction result and the true value. It is between 0 and 1.0. The closer IoU is to 1.0, the closer the bonding box predicted by the model is to the real bounding box. The mean Intersection over Union (mIoU) is the average value calculated after accumulating the IoU of each category. Their formulas are shown in Eq. (3.3) and Eq. (3.4) respectively.

$$IoU = \frac{TP}{TP++FP+FN} \tag{3.3}$$

$$mIoU = \frac{1}{C_n}\sum_{i=C_1}^{C_n} IoU_i \tag{3.4}$$

where *C* is the category, *n* is the number of categories.

For Average Precision, it is to calculate the area under the PR Curve (Precision-Recall Curve), so it needs to be calculated using integrals,

$$AP = \int_0^1 p(r)dr \tag{3.5}$$

It follows that mean Average Precision (mAP) is the average of AP on all categories, and the formula is,

$$mAP = \frac{\sum_{i=1}^{n} AP_i}{N} \tag{3.6}$$

where N (generally greater than 1) refers to the number of categories.

Chapter 4
Basic Methods of Waste Classification

In this chapter, we consider the advantages of deep learning algorithms and applies them to the task of waste classification. Models based on CNN and Transformer are included, such as the popular models YOLOv7, YOLOv8, Swin Transformer, and emerging star large language model. In this chapter, the experiments on YOLOv7, YOLOv8, and Swin Transformer models are based on WasteData, and the experiments on large language model are based on WasteNet. This is the reason why large language model can generate detailed descriptions of multiple objects in complex scenes. WasteNet may be more suitable for large language model because it provides richer contextual information and helps the model better understand waste objects in images.

4.1 YOLOv7

The backbone of YOLOv7 makes use of ELAN structure to enhance the model performance without destroying the original gradient path. In contrast, for the MP structure, downsampling is achieved by using both convolution and max pooling with the same number of channels. Next, the detection head is an anchor-based structure, mainly using SPPCSPC structure, Rep structure, and ELAN structure (different from the one in the backbone). We have chosen YOLOv7 model as our baseline and improved the model with the following (Wang, Bochkovskiy and Liao, 2022).

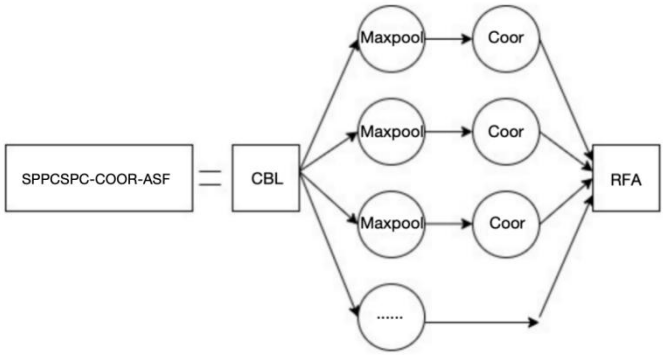

Figure 4.1 The structure of SPPCSPC-COOR-ASF module

In order to improve the model performance by making it much more capable of feature representation, we select RFA component of the AugFPN model (Guo, Fan, Zhang, Xiang and Pan, 2020). In general, the RFA component can generate contextual features in receptive fields through pooling, which can expand the receptive field whilst keeping the depth of the network structure constant. It can be considered a feature enhancer. In detail, the ASF module in the RFA component makes use of a similar approach to spatial attention. It can generate adaptive spatial weights for visual features of multiple receptive fields by using convolutions. This enables an efficient fusion of features from various receptive fields and enhances feature representation. Therefore, we add the RFA component into YOLOv7 to enhance the feature representation capability. In YOLOv7, the original SPPCSPC module is employed as a method to obtain contextual feature maps of a slew of receptive fields, which is also harnessed to perform pooling operations. Therefore, we combine the RFA component with the

original SPPCSPC module to obtain a new SPPCSPC-COOR-ASF module to generate adaptive feature maps with a plethora of spatial weights for multiple scales. Figure 4.1 illustrates the structure of the SPPCSPC-COOR-ASF module.

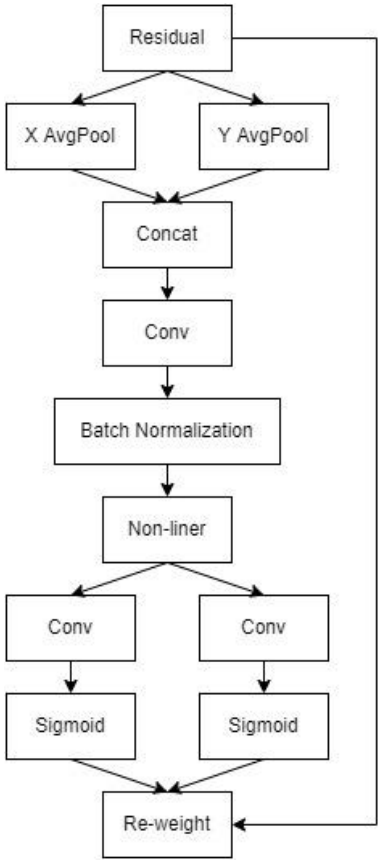

Figure 4.2 The structure of Coor attention mechanism

The SPPCSPC-COOR-ASF module firstly generates a feature map by pooling branches at a fixed scale. After that, we insert a coordinate attention mechanism, namely, the Coor module, to capture remote dependencies and enhance the representation of objects of interest (Hou, Zhou and Feng, 2021). The structure of the Coor module can be seen in Figure 4.2. Finally, the channels are compressed by using the RFA module to embed spatial information into the spatial attention graph (Yan et al., 2019). The new features containing multi-scale contextual information are generated by a weighted fusion of contextual features. This enables the improved YOLOv7 to have a more robust detection performance than the original YOLOv7.

Furthermore, we replace the Pyramid Split Attention (PSA) module (Zhang, Zu,

Lu, Zou and Meng, 2021) with 3×3 convolutions in bottleneck to get the new Efficient Pyramid Split Attention (EPSA) block to improve the model performance. Currently, attention modules introduced into CNN can bring significant improvement to the model, such as CBAM (Woo, Park, Lee and Kweon, 2018).

However, there are challenging problems with these modules. As an example, the first problem is that though these modules are considered both spatial attention and channel attention, they can only capture effective local information and cannot establish long-term dependencies. The second problem is that the spatial information of feature maps at different scales cannot be effectively utilized to enrich the feature space. The proposal of PyConv solves these two drawbacks but imposes a huge computational burden on the model. After that, the proposed PSA module not only solves the two problems described above but also has the advantage of light and high efficiency. Therefore, the PSA module is chosen in this thesis to improve YOLOv7.

The structure of PSA module is shown in Figure 4.3. The input tensor enters the SPC module and is divided into S groups, where the convolutional kernel size tends to increase, and then the convolutional layers are grouped to avoid the increase in computation. After the SPC module processing, the attention values of different scales are generated in the SE Weight module. Finally, the spliced attention weights are also subject to Softmax operation to get the output. In this way, the PSA module can handle the spatial information of the input feature maps at multiple scales that can effectively establish the long-term dependencies between the attention of the multi-scale channels.

Figure 4.3 The structure of the PSA module

In this thesis, we select Transformer to create Transformer Head as the decoupling head of YOLOv7, to improve the accuracy of visual object detection (Vaswani et al.,

2019). The Transformer has the advantage of fast-forward propagation, low structural complexity, and high efficiency of feature extraction. It is mainly based on scaled dot-product attention. The calculation formula is shown in Eq. (4.1). Because YOLOv7 is different from YOLOv5 (Zhu, Lyu, Wang and Zhao, 2021), it does not have a C3 module; we use the Transformer block.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_K}}\right) \times V \tag{4.1}$$

where Q shows query matrix, K represents key matrix, and V is value matrix. Then, d_k represents the dimension of Q and K, which is employed to control the range of dot product values. The Transformer model is based on multi-headed attention with the addition of normalization, summation, and multi-layer perceptron structure. The encoder of this structure consists of a stack of N-layer networks. Each layer contains an encoder and a decoder. The encoder contains two layers, the multi-head attention mechanism, and the feedforward neural network, while the decoder has three layers, the multi-head attention mechanism with mask, the multi-head attention mechanism, and the feedforward neural network (Vaswani et al., 2019).

The advantage of this Transformer is that it can minimize the dependence on external information and focus the arithmetic power on the correlation information of the sequence data itself. The combination of Transformer and YOLOv7 can make full use of the convolutional neural network to filter out a large amount of irrelevant information while using the extracted feature information as input to speed up the network convergence, reduce the training computation and improve the model performance.

Our experimental hardware configurations are Intel I7 and Nvidia GeForce GTX 2060 graphics card. The software includes Microsoft Windows operating system, CUDA11.1 software environment, and PyTorch deep learning framework. The initial learning rate of the training process is assigned to 0.02; the batch size is set to 8. Moreover, 300 epochs were applied to the model training process.

4.2 YOLOv8

Compared with YOLOv5, YOLOv8 has made a spat of great improvements. In the backbone network, YOLOv8 continues the CSP idea, still using the SPPF module, but replacing the C3 module with the C2F module, and application two 3×3 convolutions to reduce the resolution by a factor of 4 and achieve lightweight. After that, all the convolution machine structures in the PAN-FPN upsampling stage of YOLOv5 were removed in YOLOv8. In the neck and head stages, YOLOv8 adopts Decoupled-Head, eliminates the obj branch, and replaces the anchor-base with anchor-free. YOLOv8 then eliminates the objectness branch and adopts Binary CrossEntropy Loss (BCE Loss) as classification loss, and utilizes CIoU Loss and Distribution Focal Loss (DFL) as regression loss. Finally, the matching strategy, YOLOv8, uses a dynamic Task-Aligned Assigner, discarding the IoU matching method (Bochkovskiy, Wang and Liao, 2020). We chose the YOLOv8 model as our baseline and improved the model performance by following three aspects. Figure 4.4 shows the overall structure of our model.

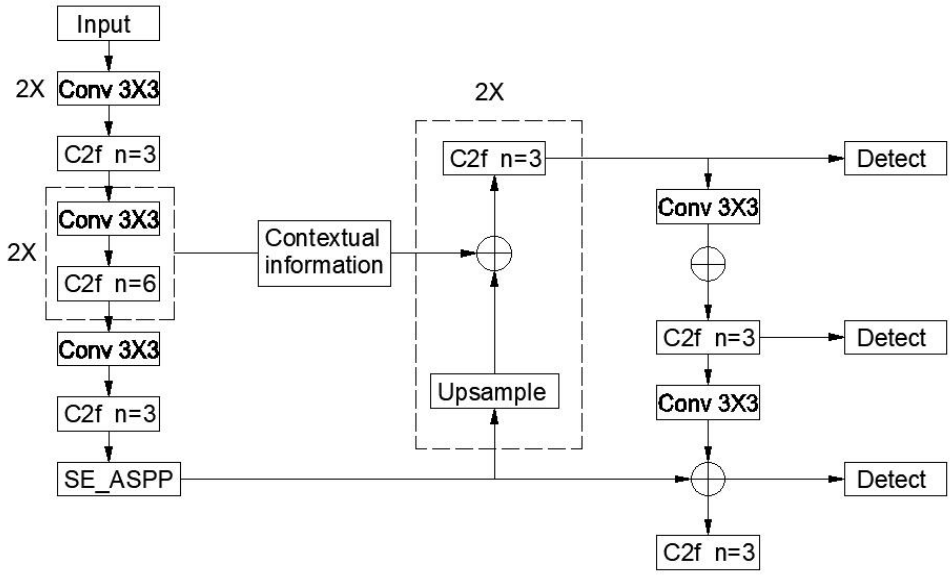

Figure 4.4 The structure of improved YOLOv8 model

After that, to further improve the performance of the model, we integrate the contextual information module into the model (Chen et al., 2020). This strategy is

implemented to effectively extract higher-level abstract features in waste images. The extraction of high-level features not only helps the model process complex visual features, but also enhances the model's ability to capture semantic information. Figure 4.5 illustrates the structure of contextual information mechanism. Feature fusion can synthesize the information of both shallow and deep features to achieve the complementary advantages of the two features and make the detection of the model more robust and accurate (Li and Zhou, 2017; Nguyen and Yan, 2021; Wang and Yan, 2023).

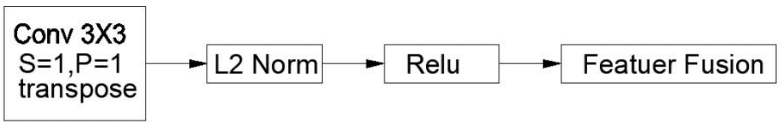

Figure 4.5 The structure of contextual information mechanism

Finally, we replace the SPPF module in YOLOv8 with the SE_ASPP module. SE_ASPP is a combination of Atrous Spatial Pyramid Pooling (ASPP) and the channel attention mechanism SENet (Hu, Shen and Sun, 2018). In general, the receptive field is closely related to the object detection performance, the larger the receptive field the better the network performance, but the receptive field should not be extremely large, it will lead to the model is difficult to converge. If the model is required to have a large receptive field while ensuring that the resolution of the feature map does not lose much (which may lose the image details), dilated convolution is essential.

Therefore, the ASPP module has the advantage of being able to balance the receptive field and resolution well. ASPP module uses multiple parallel dilated convolution layers with different sampling rates for the input features to be sampled. This allows the model to individually construct different receptive fields from branches of different scales, extract the input features, and use them to generate the final feature results. Moreover, utility of the channel attention mechanism SENet not only effectively enables the parallel transfer of key feature information, improves information reuse and

enhances useful information, but also compresses useless feature information. Figure 4.6 illustrates the specific structure of SE_ASPP.

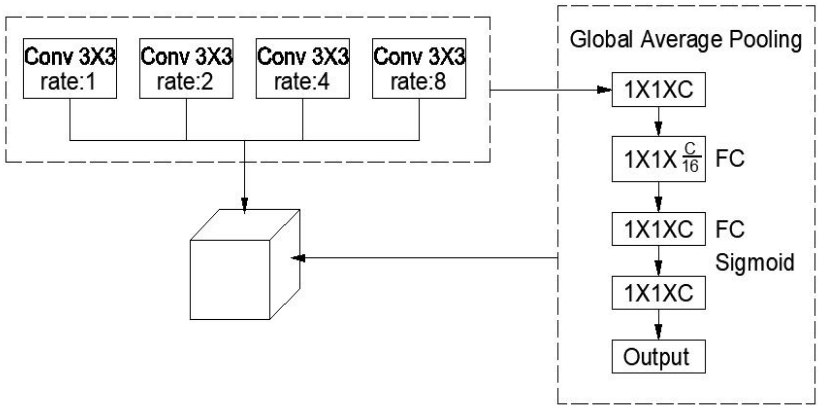

Figure 4.6 The structure of SE_ASPP module

The hardware configuration of the experiment is NVIDIA GEFORCE GTX 2060 graphics card and Intel I7. Python and Torch are used for the experiment software. The parameters of this experiment are shown in Table 4.1.

Table 4.1 The parameters of experiment.

Classes	Parameters
Initial Learning Rate	0.01
Momentum	0.9
Weight Decay	0.0005
Batch Size	16
Epoch	300
Optimizer	SGD

4.3 Swin Transformer

Targeted at resolving the problem of a large amount of calculations, Swin Transformer model takes use of the designed window, which only computes the self-attention inside

the window, dramatically reduces the amount of calculation and constructs a hierarchical Transformer that was used as a backbone network for visual tasks such as image classification, object detection, and semantic segmentation (Vaswani et al., 2017; Liu et al., 2021). This model was applied to our experiments, which will be detailed in this thesis.

Swin Transformer model is improved based on Vision Transformer, which will be applied to carry out visual object detection and semantic segmentation (Dosovitskiy, et al., 2021). Therefore, in this thesis, we apply this model to carry out waste detection, segmentation, and classification to improve accuracy and training speed.

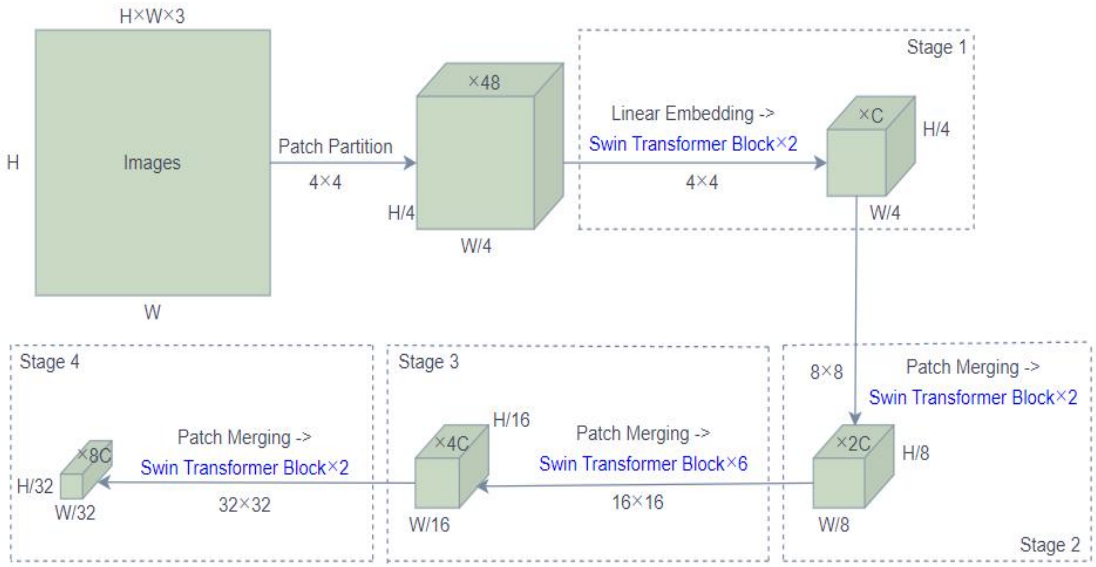

Figure 4.7 The structure of Swin Transformer

Similar to Vision Transformer, Swin Transformer also adopted the segmented blocks. If the size of each patch is 4 × 4, the characteristic dimension of the patch is 4×4×3. However, it is worth noting that Swin Transformer determines the number of patches at first, while it confirms the size of the patch first. After that, the architecture is similar to that of CNN, which constructs four stages as shown in Figure 4.7.

There are three concepts that need to be known, namely patch, token, and window. Assuming that the size of the input image is 224×224, it is firstly divided into a number

of small pieces of 4×4 pixels, so a total of 56×56 small pieces can be divided. Each small piece is called a patch, which is also called token. In addition, if an image is divided into 7×7 windows, so that each window will contain 8×8 patches.

In this model, it is worth noting the Swin Transformer block part which is an improvement on the standard Transformer, mainly take use of shifted window to improve the standard multi-head self-attention module, which are W-MSA and SW-MSA.

W-MSA indicates that multi-head self-attention inside the window. We treat the window as a global independent no calculating the attention of each token in the window, reducing the computational complexity. We should take use of SW-MSA to aggregate information between different windows. The difference between SW-MSA and W-MSA is that SW-MSA offsets the coverage of the window, and the original text is set to half of the side length of the window. After the window slides diagonally, the window in the middle can get the information of all windows on the upper layer.

Thus, according to the characteristics that Swin Transformer is utilized as the backbone network, in this thesis, we combine Swin Transformer with Mask R-CNN and take Swin Transformer as the backbone network of Mask R-CNN for waste classification. The framework of the model is shown in Figure 4.8.

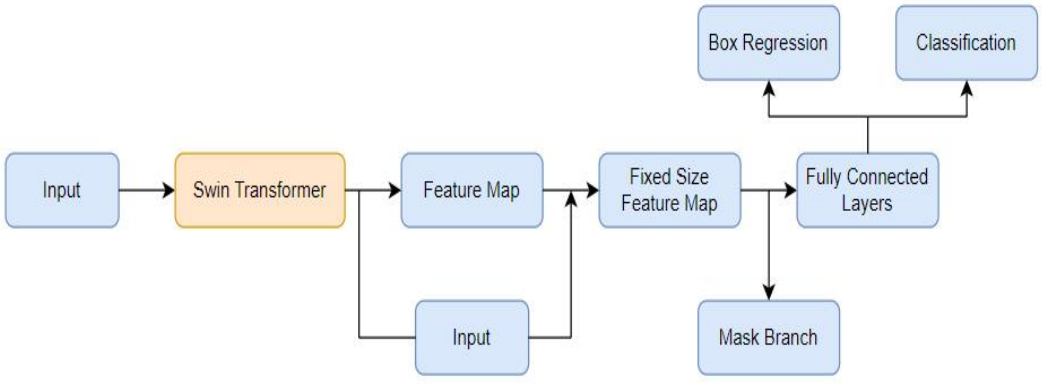

Figure 4.8 The structure of the combined model

Mask R-CNN is an improved model based on Faster R-CNN. After the primary

feature network, a fully connected split subnet is added, a mask prediction branch is added to each ROI (Wu et al., 2020). It mainly has seven steps to be implemented:

(1) Step 1: Input images

(2) Step 2: Use the backbone network to obtain the feature map

(3) Step 3: Set ROI for feature map to obtain multiple ROI

(4) Step 4: Use the ROI with the RPN network for binary classification and bounding box regression to obtain the filtered ROI

(5) Step 5: Output ROI after ROIAlign with the filtered ROI

We see that Mask R-CNN has the advantages of good segmentation and fast training speed, which completes the tasks of detection and segmentation at the same time. Therefore, we choose to combine Mask R-CNN with Swin Transformer, the Swin Transformer is the backbone network. We take hazardous waste such as batteries as an example, the feature map of Swin Transformer is shown in Figure 4.9.

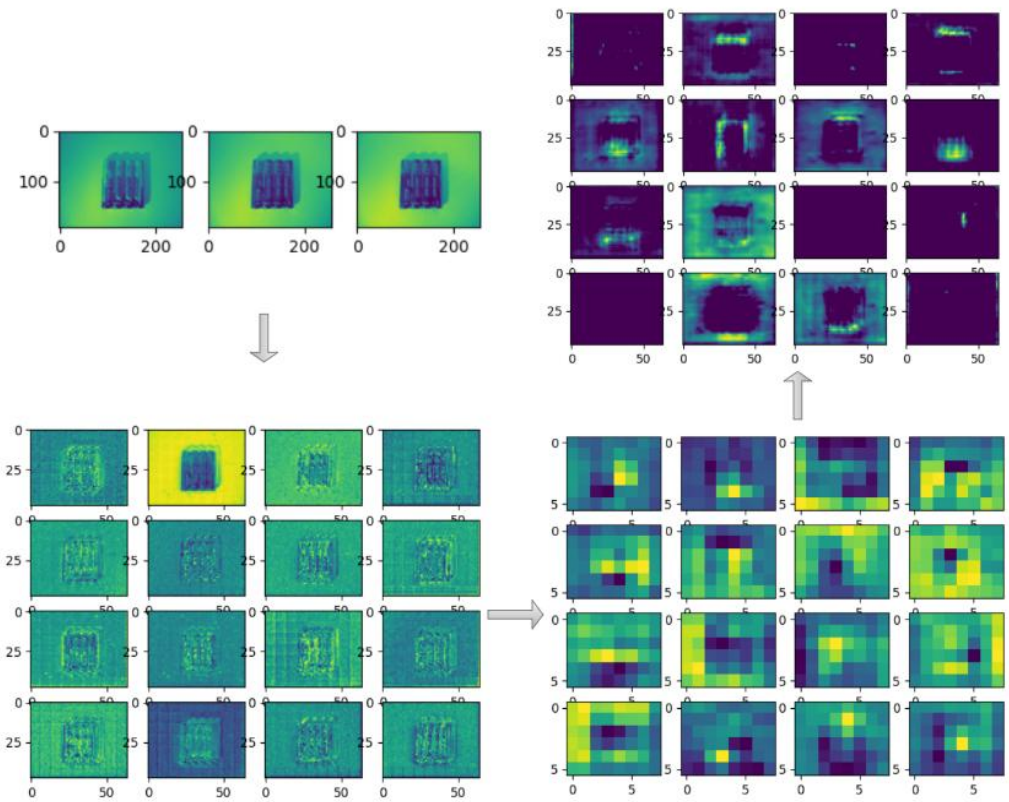

Figure 4.9 The Swin Transformer feature maps

Moreover, in this experiment, we made use of NVIDIA GEFORCE RTX 2060 GPU and Intel I7 CPU and installed code editors PyCharm, neural network framework PyTorch, CUDA, CUDNN and OpenCV.

4.4 Large Language Model

Artificial intelligence (abbreviation AI), particularly large-scale language models, has shown remarkable promise in various applications, including image classification tasks (Dai et al., 2023; Ding et al., 2023). The cutting-edge GPT-4, for instance, can perform image classification, text-to-image conversion, and image-to-text translation. Its superior generalization and zero-shot learning abilities enable the processing of complex datasets with high efficiency (ChatGPT, personal communication, 2023). Building on this potential, we introduce a pioneering approach to waste classification by harnessing the semantic capabilities of large language models. We utilize MiniGPT-4 to generate textual descriptions of waste images, which we then input into the pre-trained language model RoBERTa (Liu et al., 2019; Zhu, Chen, Shen, Li and Elhoseiny, 2023). Concurrently, we process the waste images directly through the Swin Transformer model (Liu et al., 2021).

Despite these advancements, large language models remain challenging in practical scenarios. They are composed of multilayer neural networks with hundreds of millions of parameters, necessitating substantial computational resources and extended training periods. This complexity results in considerable training expenses. Furthermore, an increase in model parameters can complicate the model's interpretability and elevate its complexity (Singla, 2023; Zan et al., 2023; Zvyagin et al., 2022).

Therefore, to solve this problem and avoid manually collecting image description information from a large language model, we introduce MiniGPT-4 into our model through an API interface, aiming to leverage the rich semantics of the large language model in a simplified way to create an efficient and highly accurate waste classification model. Our description-driven approach to image classification shows promise,

particularly when image data is scarce, making it well-suited for the task of waste classification.

The Structure of Our Framework

Deep learning models, such as CNN, have performed well on computer vision tasks over the past years. It can learn visual features directly from image data. This means that the model can predict the output directly from the input data without any human intervention at intermediate steps in the training process. Recently, it has been confirmed that large language models have excellent performance (Chen, Guo, Yi, Li and Elhoseiny, 2022; Driess et al., 2023; ChatGPT, personal communication, 2023). Therefore, we speculate that applying large language models to image classification models by introducing multimodal information so that the image classification model is not limited to learning features only from the image data, which will bring about a performance improvement. Our model framework is shown in Figure 4.10.

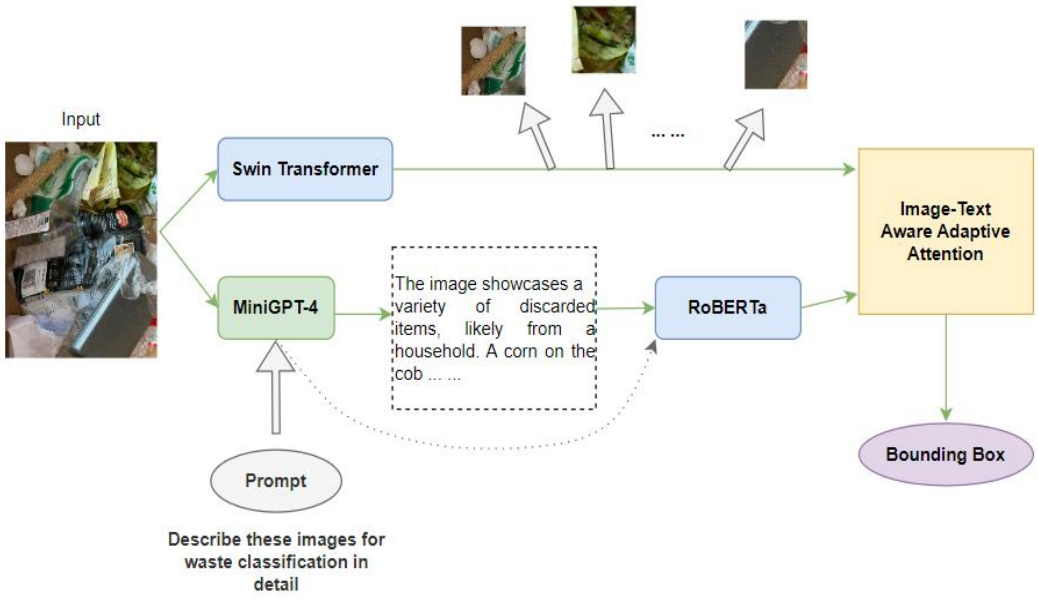

Figure 4.10 The framework of our model

Firstly, we input our waste image to MiniGPT-4 to generate a detailed description of the image. At this point, we give the prompt “Describe these images for waste

classification", which is designed to fit the prompt more closely to our task, thus increasing the accuracy of waste classification. In this step, we take the approach of describing the images by introducing the MiniGPT-4 API. The purpose of this method is to save resource allocation without affecting the classification performance. We did not choose to adopt the better-performing GPT-4 model since GPT-4 has not opened the API function for the image-to-text module. Figure 4.11 shows the description of the image produced by MiniGPT-4. After that, we feed all the image descriptions generated by MiniGPT-4 to the RoBERTa model, which converts the text descriptions into high-dimensional embedding vectors that provide semantic classification decisions for the images based on their descriptions (Liu et al., 2019). At the same time, we will also send the image data to Swin Transformer for direct image classification.

Description:

The image displays a diverse assortment of waste items. Dominating the view is a "Monster" snack packet, designed with a vibrant creature graphic, labeled as "BBQ flavor", and an empty, transparent plastic bottle with a faded orange cap, typically associated with beverages and potentially recyclable depending on its plastic code. Adjacent to these are organic waste materials: a partially visible corn cob inside a metal container and a fresh-looking pear with a small label. Scattered throughout are varied plastic waste including a plastic bag containing green lettuce leaves, a transparent clamshell container, and other less discernible packaging.

Description:

The image showcases a variety of waste materials that primarily fall under non-biodegradable and potentially recyclable categories. Predominantly, there are several pieces of white polystyrene foam, often referred to as Styrofoam, which is generally non-biodegradable and requires specialized recycling facilities. Additionally, there are plastic bottles, one of which has a visible brand label "Rawmind" and another displaying nutrition information, both potentially recyclable under the plastics category. Various plastic wrappings and packaging are also seen, which would need material verification to decide on their recycling feasibility. A torn paper or plastic sachet is also evident. To efficiently manage this waste, separation of recyclable plastics from non-recyclable items like Styrofoam is crucial.

Figure 4.11 Image description generated with MiniGPT-4

Image descriptions generated using large language model

Our proposed multimodal waste classification model can integrate textual descriptions and image processing content for waste classification. In the entire model, generating image descriptions through MiniGPT-4 is one of the most important aspects. Firstly, the waste image and the prompt related to this image are inputted into the model to get detailed description information of the waste image. During this process, we found that

even if the same image is input into MiniGPT-4, the image description generated is different every time.

We conjectured that different descriptions generated for the same image would have different effects on the model training results. Thus, we explored the effects of different lengths of image descriptions and different prompts of input on the model results in our ablation experiments. Based on the experimental results, we finally chose to define the prompt as “Describe these images for waste classification in detail”. Moreover, we selected the description information of the image as shown in Figure 4.12. As can be seen, "Description1" gives key information about the waste in the image. Whereas "Description2" lacks information about some objects such as corn cobs and plastic bags.

Figure 4.12 An example of the image description generated by MiniGPT-4 with different prompts

4.5 Summary

In this chapter, we simply applied YOLOv7, YOLOv8 and Swin Transformer for the waste classification tasks and achieved positive detection results, which shows that the deep learning method is effective. Afterwards, we combined the large language model to convert the rich semantic information of MiniGPT-4 into waste image data, improving the accuracy of waste classification.

Chapter 5
Data Augmentation for Waste Classification

In this chapter, we mainly summarize the role of non-uniform data augmentation on waste classification models. Non-uniform data augmentation includes non-uniform color data augmentation and non-uniform offset data augmentation. Not only does it significantly improve model performance, it also offers effectiveness and simplicity in solving real-world waste classification problems. In this chapter, the dataset used is ZeroWaste.

5.1 The Structure of Our Framework

Integrated with deep learning and tailored for complex real-world waste classification backgrounds, we prefer semantic segmentation to enhancing conventional object detection techniques. This preference stems from our observation of the severe clutter encountered in such scenarios, where the presence of a large and diverse amount of waste can lead to items obscuring each other. Semantic segmentation allows for the division of the entire image into distinct regions without gaps, assigning each region to a specific category, and is adept at object sizes, shapes, and conditions (Chen, Du, Zhang, Qian and Wang, 2022; He, Yang and Qi, 2021; Huo et al., 2021). Our approach not only enhances waste classification efficiency and reduces the need for comprehensive manual pixel-precise labeling but also addresses the issue of data scarcity when training model (Mahajan, 2018; Yan, 2023).

From self-training and collaborative methods to consistent regularization and the latest trend of generative adversarial networks, semi-supervised learning techniques have been evolving for a while (Chen, Yuan, Zeng and Wang, 2021; Mittal, Tatarchenko and Brox, 2019; Qiao, Shen, Zhang, Wang and Yuille, 2018; Yang, Zhuo, Qi, Shi and Gao, 2022). This thesis seeks to explore how we can overcome the challenges posed by limited waste datasets, their deficiencies, and the intricacy of scenarios for waste classification. We investigate the application of semi-supervised learning to boost the efficiency and accuracy of waste classification and to improve the severe shortage of pixel-precise annotations in semi-supervised semantic segmentation area. These are significant questions that we aim to answer through our research.

In this thesis, we direct our attention towards leveraging the consistent regularization strategy as a solution to the previously outlined challenges by crafting a non-uniform data augmentation technique specifically designed for the distinct attributes of real-world waste classification. Through detailed analysis, we observed that traditional data augmentation methods such as Mixup, Mosaic, and Cutout, although innovative, exhibit inherent shortcomings (DeVries and Taylor, 2017; Guo, Mao and

Zhang, 2019; Yun et al., 2019). Cutout, for instance, disregards segments of the image, failing to capitalize on the entirety of the available image data, whereas images processed with Mixup often appear unrealistic due to localized blurring effects. We introduce an innovative solution for waste classification through the use of a cutting-edge non-uniform data augmentation technique. This approach excels at replicating diverse environmental scenarios, including variations in lighting and object forms, thereby significantly improving the robustness of the model (Qi, Nguyen and Yan, 2024).

Initially, we adopt U-Net as the foundational framework and ResNet-50 as the backbone network to assess the efficacy of our specialized non-uniform data augmentation and adaptive weighted loss function. The architecture of the network is depicted in Figure 5.1. During the training phase with unlabeled data, comparisons are made between outputs from data processed without non-uniform data augmentation and those subjected to it, measuring the L1 loss to gauge the consistency between the predictions from these two data variants (Gao, Nguyen and Yan, 2023; Nguyen and Yan, 2023). A lower L1 loss indicates improved model performance. To further enhance the model's ability to generalize, we introduce drop perturbation techniques to both the original input channels and the feature channels (Srivastava, Geoffrey, Alex, Ilya and Ruslan, 2014). This strategic integration aims to refine the model's predictive accuracy and robustness.

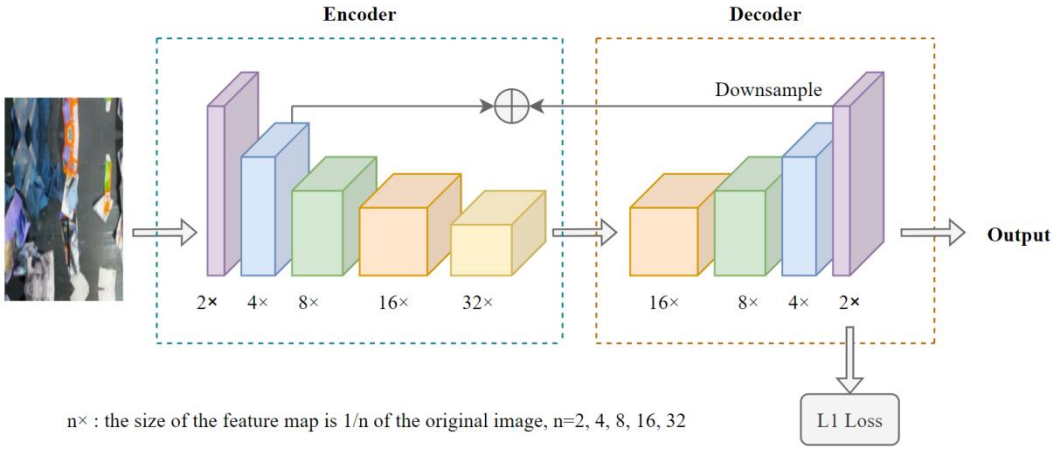

Figure 5.1 The architecture of the network

5.2 Non-uniform Color Data Augmentation

Our method of non-uniform color data augmentation intricately simulates the complexities of natural lighting found in the real world, setting it apart from the simpler approach of random brightness augmentation method (Yang, Xu, Wang and Zhang, 2022). Instead of uniformly adjusting brightness across all pixels like the random brightness augmentation method, our technique introduces a wholly random effect on each pixel point, as illustrated in Figure. 5.2.

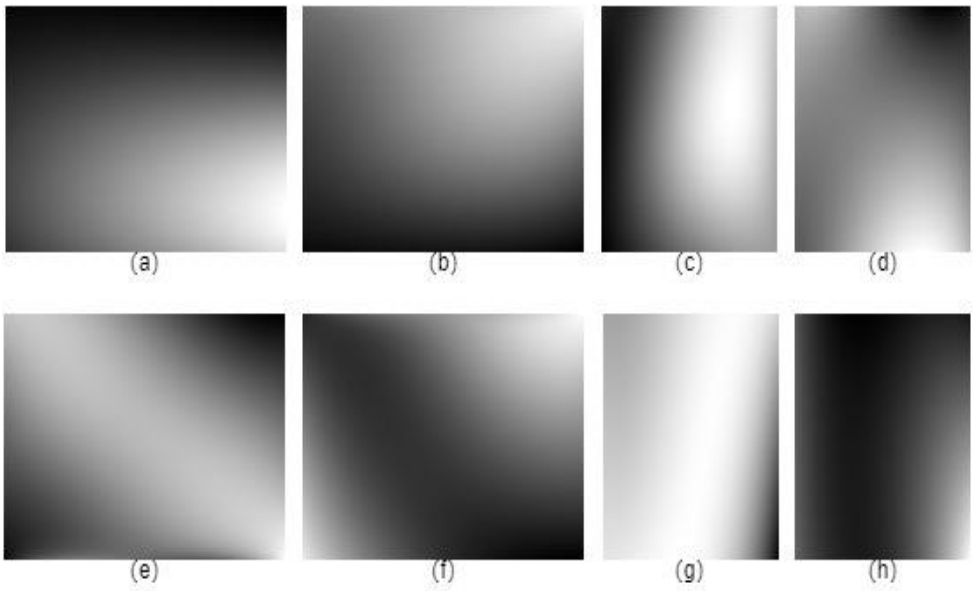

Figure 5.2 The examples of typical non-uniform color data augmentation

The images from Figure. 5.2(a) to Figure. 5.2(d) demonstrate how light and shadow vary across the bottom, top, right, and left sides in the image, adding depth and realism to the scene. Specifically, Figure. 5.2(e) presents a gradient of light that fades towards the image's edges, centering the brightness, whereas Figure. 5.2(f) displays the reverse pattern to Figure. 5.2(e). Continuing, Figure. 5.2(g) and Figure. 5.2(h) depict scenarios with significantly dark and intensely bright lighting, respectively. Although Figure. 5.2 provides just eight examples of our diverse non-uniform color data augmentation techniques, numerous similar variations are present, as Figure. 5.3 further explores.

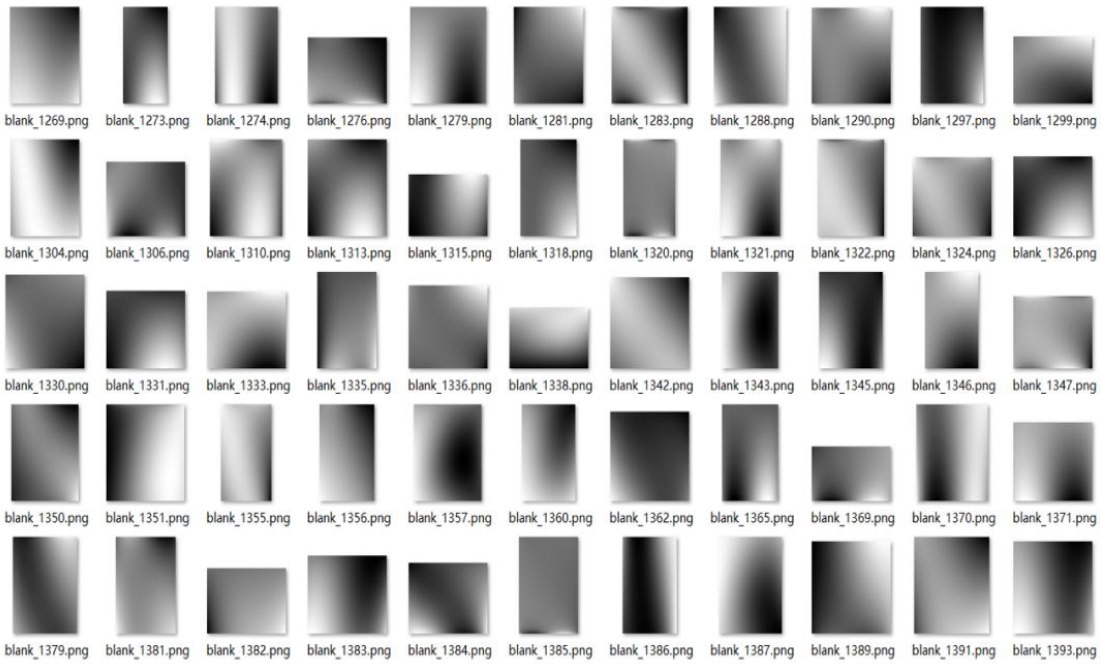

Figure 5.3 Some examples of atypical non-uniform color data augmentation

Figure 5.4 Comparison of the image after non-uniform color data augmentation and the original image from the ZeroWaste dataset. (a) is the original image. This image also incorporates natural light produced during waste classification. The light shines from the upper right corner, and the lower left corner of the image is darker. (b) is the image obtained after data augmentation processing. It can be seen that the light has been changed, the dark part has become the upper left corner, and the light part has been moved to the lower right corner

When applying this technique to the ZeroWaste dataset, we keep the pixel positions constant but vary their values randomly and smoothly, creating a dynamic

range of lighting effects. For instance, in Figure. 5.4(b), the light is dimmed from the top left corner while being bright from the bottom right, creating a nuanced contrast when compared to the original dataset image in Figure. 5.4(a). This adjustment darkens the upper left while brightening the lower right, effectively replicating the varied lighting conditions that could affect waste samples, despite them being of the same type, photographed from the same angle. This method not only enhances the dataset's diversity but also prepares the model to recognize waste under a broader spectrum of lighting scenarios, making our approach uniquely effective in simulating real-world conditions.

5.3 Adaptive Weighted Loss Function

If the amount of data images in some categories is too small, it will lead to serious imbalance in data distribution. Therefore, the model will not learn enough features of these categories, affecting model performance. During the model training process, we noticed that the number of waste images in the two categories “Metal” and “Rigid Plastic” in the dataset was low. To overcome this problem, we designed a new loss function called adaptive weighted loss, as shown in the following Eq. (5.5), Eq. (5.6), and Eq. (5.7).

$$L_{cls} = -\frac{1}{Z}\{\sum_{i=1}^{N} loss\,(p_i)\} \tag{5.5}$$

and

$$loss\,(p_i) = \begin{cases} e^{w-p_i} \bullet \ln p_i, & p_i < \eta \\ 0, & p_i \geq \eta \end{cases} \tag{5.6}$$

and

$$Z = \sum_{i=1}^{N} [p_i < \eta] \tag{5.7}$$

where i denotes the pixel point, the number of e elements in the mask is referred to Z, which is equivalent to the adjustment coefficient. p_i is the pixel prediction probability, and w represents the weight. Moreover, η is the hyper parameter, and the set value is

0.99. Finally, e_w is the overall weight assigned to each type of waste, which is based on our experimental settings.

The ZeroWaste dataset contains four waste categories and one background category. For the case where there is less data in the Rigid Plastic and Metal categories, we set *w* for these two classes to 3.0, and *w* for all other categories to 1.0. This is the training optimal value obtained after multiple rounds of ablation studies. In Eq. (5.6), e_p aims to justify the pixel weight assignment. If the *p*-value is smaller, the weight assigned to it will increase, which helps to achieve data balance. On the contrary, if e_p is stable, *p* should be larger, making the e_p value smaller.

Besides, if p_i is greater than or equal to η, we will set the mask to 0, automatically adjust the number of negative and positive samples through hard coding method, and filter out the samples that the model has learned well without limitations. Finally, the value of the loss function is calculated by multiplying the weight and cross entropy, which is different from the traditional weight loss function OhemCELoss (Shrivastava, Gupta and Girshick, 2016). OhemCELoss limits the proportion of negative and positive samples, while our method does not set this limit and extracts valuable features in a simple and effective way, thereby improving model performance.

Table 5.2 Training parameters of this experiments.

Classes	Parameters
Initial learning rate	$1e^{-4}$
Optimizer	Adamw
Weight decay	$1e^{-3}$
Batch size	26
Epoch	100

Our experiments were based on a server with RTX A5000 GPU and AMD EPYC 7543 CPU. Installed code editors VSCode, neural network framework PyTorch, CUDA, CUDNN, and, OpenCV. The specific experimental parameters are reflected in Table 5.2.

5.4 Non-uniform Offset Data Augmentation

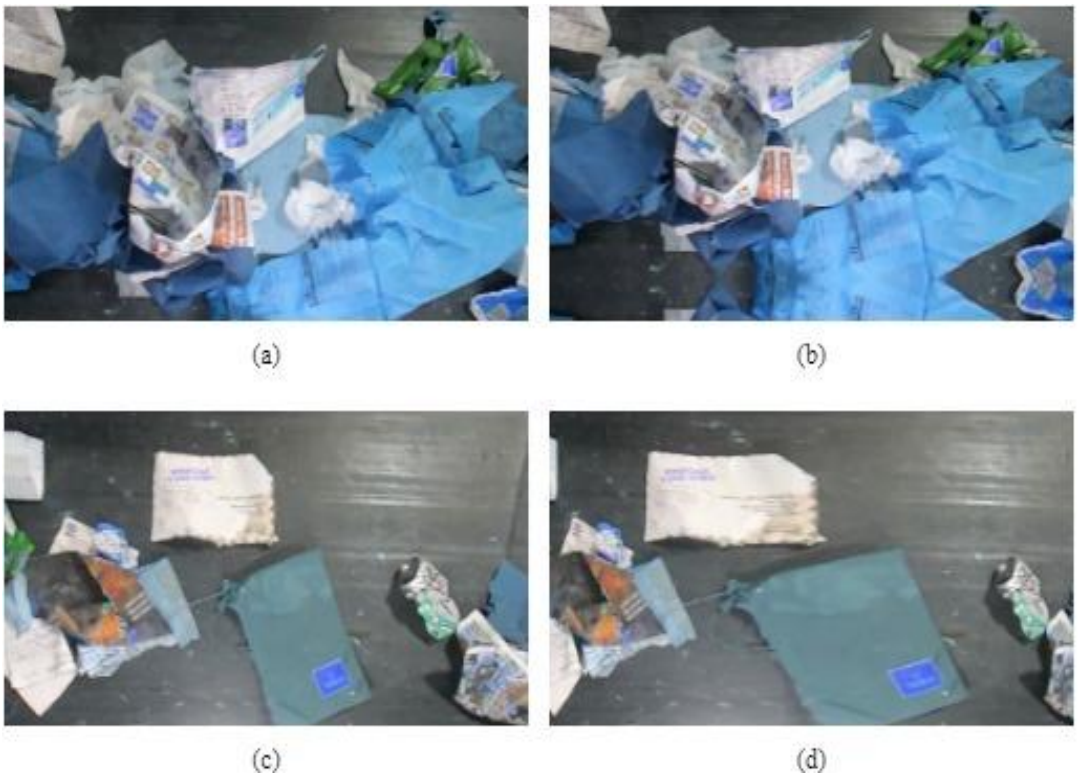

Figure 5.5 Comparison of the image after non-uniform offset data augmentation and the original image from the ZeroWaste dataset. (b) and (d) are the images obtained after data augmentation processing of the original images (a) and (c) respectively

In Figure. 5.5(b), we observe a condensed rendition of Figure. 5.5(a), where the entirety of the pixel arrangement appears to be shifted upwards, suggesting a vertical translation. Conversely, Figure. 5.5(d) represents an extended version of Figure. 5.5(c), with pixels collectively migrating to the right, indicating a horizontal pan. This manipulation forms the core of our innovative non-uniform offset data augmentation strategy, which stands in contrast to the conventional methods of scaling and aspect ratio adjustments typically employed in image augmentation practices. Our approach is meticulously designed to

assign a unique, random displacement to each pixel within the image. For instance, while one pixel might experience a shift of 1 pixel, another could be moved by 3 pixels, with these adjustments decreasing gradually until they reach a predetermined peak value. The intensity of these shifts is modulated using a sine wave function.

In this method, $i \in I^{W \times H \times C}$ symbolizes a training image including two spatial dimensions, x and y. The objective behind our non-uniform offset data augmentation strategy is to create a modified version of the training image, denoted as $\tilde{i}(x, y)$, through the specifically defined formulas as follows:

$$\tilde{i}(x, y) = \Delta i(x, y) * \sin(2\pi * r / v_c) + i(x, y), \tag{5.1}$$

and

$$v_c = 1200 + 200 * ((r - 0.5) * 2) \tag{5.2}$$

where x and y detail the pixel intensity along their respective axes. The process begins with an initial pixel offset, $\Delta i(x, y)$, randomly assigned to each pixel, subsequently fine-tuned through multiplication with a sine wave to finalize the offset. The variable r denotes a random figure complying to a uniform distribution across 0 to 1.0, ensuring the randomness of the Sine function's output. The symbol v_c represents the peak value of this sine modulation. Eq. (5.3) and Eq. (5.4) show the equation of $\Delta i(x, y)$.

$$x \sim Unif(0, W), \qquad y \sim Unif(0, H) \tag{5.3}$$

and

$$\Delta i(x, y) = \begin{cases} \Delta x = v_o + 15 * ((r - 0.5) * 2) & r < 0.5 \\ \Delta y = v_o + 15 * ((r - 0.5) * 2) & r \geq 0.5 \end{cases} \tag{5.4}$$

The mechanism for calculating $\Delta i(x, y)$ involves setting a base pixel value, here chosen as 70, and then applying random arithmetic operations to obtain the final offset

value, $\Delta i(x, y)$. Depending on r's value, the offset alternates between Δx and Δy to maintain the shift within a desirable range. This experimental setup selected an initial offset range between 55 and 85, with the baseline of 70 and this range is the most effective through our testing. It is worth mentioning that the initial offset range (from 55 to 85) is applicable to the ZeroWaste dataset. If different datasets are applied, the initial offset range can be changed after experimental testing. Eq. (5.1) illustrates how we compute the final pixel shift by combining the base offset value with a sine function. In fact, cubic, cosine, quadratic, and other mathematical functions can also be utilized instead of sine function, but this also depends on different dataset environments. The sine wave's peak is similarly defined, with values ranging between 1,000 and 1,400.

Table 5.1 The examples of offset values for x-axis.

Pixels	Different offset values
000	0.00000000000000
001	0.47157373246046
002	0.94312902362002
… …	… …
255	75.403237628632
256	75.401564240098
257	75.400021858127
258	75.393496720912
… …	… …

Table 5.1 illustrates the offset values along the x-axis. Our method ensures a smooth and continuous transition in pixel changes, which is significantly different from mere noise addition, which can abruptly blur pixel details. By combining this method with the ZeroWaste dataset, which contains around 6,212 unlabeled images, we significantly mitigate the challenges posed by insufficient data. Our method not only preserves the integrity of the image, but also prevents the misassignment of negative

and positive labels that often occurs when image resizing, thereby enhancing the utility of the dataset and the model's ability to effectively generalize to real-world waste classification. Moreover, our non-uniform data augmentation also promotes smooth transitions between pixel adjustments, giving our model enhanced generalization capabilities and making it more adept at handling the complexities of real-world waste classification.

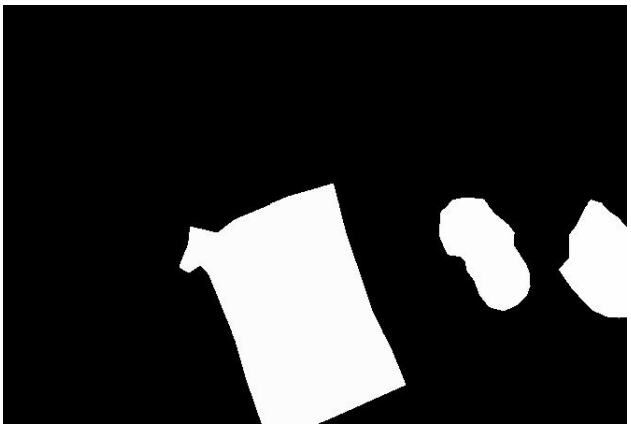

Figure 5.6 Example of a masks of the original image

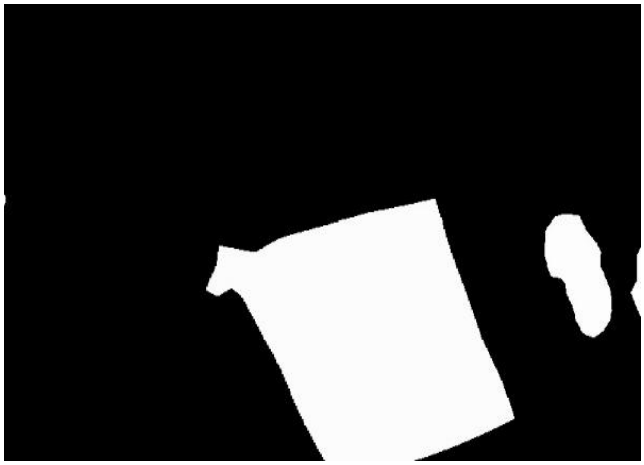

Figure 5.7 Example of a masks of the original image after applying non-uniform offset data augmentation

In addition, we have attached Figure 5.6 and Figure 5.7, which represent the masks of Figure. 5.5(c) and Figure. 5.5(d) respectively. It can be seen from these two images that the model enhanced by non-uniform offset data augmentation is stable.

5.5 Summary

Our innovative related to non-uniform data augmentation is characterized by its ability to ensure that image information is not artificially added or subtracted. By uniformly and continuously adjusting the entire image, changes in each pixel maintain its intrinsic relationship and avoid unnecessary noise interference. The complex image processing greatly improves the generalization performance of the model. Furthermore, during an in-depth analysis of the ZeroWaste dataset, we noticed significant imbalances in the data, especially in the two categories with less data, which severely biased the sample size. This finding highlights the importance of our adaptive augmentation strategies aimed at reducing these differences, improving the efficiency of model training, and improving the accuracy of waste classification, thereby advancing the development of waste management technology.

Faced with the increasing cost of misclassification, we design an adaptive weighted loss function that allocates weights according to the data volume characteristics of different categories. This method dynamically adjusts the ratio between negative and positive samples by introducing a mask function, prompting the model to exclude those features that have been well learned without setting fixed bounds. When this technology is combined with the U-Net architecture, it can significantly improve the model's mean Intersection over Union (IoU) to 3.74%. This result demonstrates the effectiveness and simplicity of our approach in solving real-world waste classification problems and is applicable to various datasets dealing with data imbalance. Our main contributions include:

(1) Constructing a semi-supervised semantic segmentation waste detection model for actual waste classification scenarios, and test results show its advanced performance.
(2) A new non-uniform data augmentation method is proposed to make the model more suitable for waste classification by simulating natural lighting conditions. This not only expands the waste dataset, but also helps reduce the risk of the

network overfitting the training data. Our method ensures that the network has good generalization ability to new data.

(3) An adaptive weighted loss function is designed to specifically solve the problem of model vulnerability caused by imbalance data distribution.

Although our method is specifically designed for waste classification, it may also be applied to other fields such as traffic monitoring (applying datasets such as PASCAL VOC and MS-COCO). Our data augmentation method is based on the sine function operation and has a wide range of application potential that can be adjusted by introducing other mathematical models such as cosine or quadratic to expand its scope of application.

Going forward, we are committed to applying our approach to a wider range of deep learning challenges and continually refining our methodology. Furthermore, we believe that there is room for further improvement in non-uniform color data augmentation strategy. We will explore how to further improve the performance of the model by adjusting the color configuration of different objects, opening new avenues for future research. At the same time, we also realize that although the simplicity of the adaptive weighted loss function helps improve the accuracy of the model, there is still room for improvement. Therefore, we plan to conduct more in-depth research focusing on optimizing adaptive weighted loss functions, with the goal of developing more complex and effective solutions. This ambitious research direction underscores our commitment to transcending existing methodological limitations and pursuing excellence not only in the field of waste classification, but also in handling the complexities of deep learning applications.

Chapter 6
Semi-Supervised Learning for Waste Classification

In this chapter, we discuss the effectiveness of semi-supervised learning for waste classification. How to maximize the utilization of pseudo-label and unlabeled data is mainly considered. In addition, the Mean Iteration strategy was also proposed. In this chapter, the dataset used is WasteNet.

6.1 The Structure of Our Model

We propose an iterative and collaborative semi-supervised object detection (SSOD) framework that can utilize large amounts of unlabeled data, called CISO. Afterwards, we introduced the Mean Iteration strategy (a pseudo-label selection mechanism based on mean IoU), with the purpose of preventing model overfitting caused by pseudo-labels not being updated and reducing the generation of incorrect pseudo-labels (Qi, Nguyen and Yan, 2024). Finally, our framework also adopts weak-strong data augmentation techniques and knowledge distillation techniques to improve the efficiency and accuracy of the model (Heo, 2019; Park, Kim, Lu and Cho, 2019; Romero et al., 2014; Yim, Joo, Bae and Kim, 2017). Through extensive testing on the WasteNet dataset, we verified the effect of CISO, and the results showed that our method reached the superior performance. Simultaneously, we also conducted ablation studies to provide in-depth analysis of our strategy.

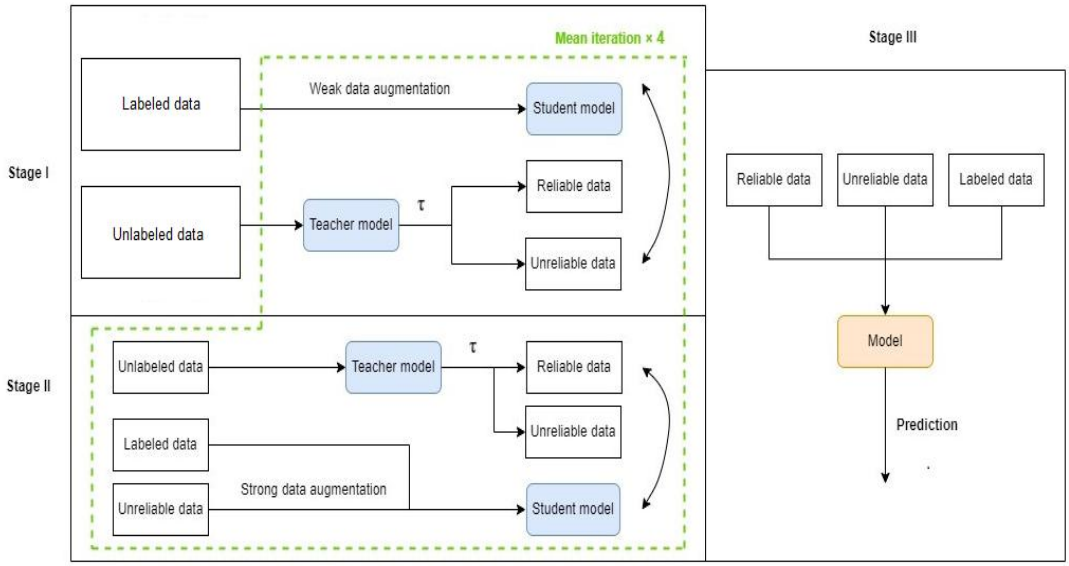

Figure 6.1 The CISO framework

The structure of CISO framework can be divided into three main stages, as shown in Figure 6.1. In the first stage, pseudo-labels of unlabeled data are generated by the teacher model, and the student model is trained on a small batch of randomly selected labeled data. For unreliable data and reliable data, their screening is based on threshold

$\tau \geqslant$ Mean (IoU). Afterwards, the unreliable data is returned to the unlabeled data pool in the second stage to generate pseudo-labels again on the complete unlabeled dataset. Next, the reliable data and the labeled data will be input into the student model at the same time for training. Then, repeat the steps of selecting the reliable data. It is worth mentioning that we perform four Mean Iteration iterations and apply a weak-strong data augmentation strategy in each iteration. In the third stage, the labeled data, the unreliable data, and the reliable data will be input into the model together for final training to complete the construction of the model.

6.2 CISO: Co-Iteration SSL for Object Detection

6.2.1 Pseudo Labeling

Numerous experiments illustrate that strategic utilization of pseudo-labeled data can significantly improve the accuracy of various algorithms, a finding that highlights the potential of integrating pseudo-labeled data to improve model effectiveness (Bar et al., 2022; Lee, 2013; Li, Liu, Zhao, Zhang and Fu, 2021; Liu et al., 2021; Zhu, Peng and Yan, 2024). This innovative strategy differs from traditional methods such as Instant-Teaching and STAC, which have each contributed to the field in their own way (Yu, Jiang, Wang, Cao and Huang, 2016; Sohn et al., 2020). For example, STAC is a pioneer in applying semi-supervised learning (SSL) to object detection, employing pseudo-label self-training and data augmentation techniques characterized by consistent regularization. This requires preliminary training of the teacher model before proceeding to train the student model, whereas our CISO approach circumvents this step.

In contrast to these approaches, CISO facilitates the end-to-end transfer of parameters between models through the application of knowledge distillation, thereby simplifying the semi-supervised learning process. While Instant-Teaching also provides an end-to-end solution, and shares the self-training aspect with CISOs, our approach stands out by retaining all unlabeled data, thus maximizing the use of pseudo-labels

with high confidence. Additionally, we introduce the concept of Mean Iteration, where the threshold τ is dynamically adjusted to optimize the use of pseudo labels and enhance the overall model performance.

Looking into the details of CISO, the process starts by generating pseudo-labels for unlabeled data and leveraging the combination of these pseudo-labels with a limited amount of labeled data to train each iteration. Specifically, in data batches, the unlabeled and labeled data are randomly sampled according to a set ratio, usually 1:10. Following that, we employ two models during the training process, namely, the student model for knowledge distillation and the teacher model. The teacher model is responsible for generating a pseudo-label for the unlabeled data, while the student model is responsible for conducting the training. Notably, the teacher model is based on the student model updated with the Exponential Moving Average (EMA) (Tarvainen and Valpola, 2017). This end-to-end framework eliminates the need for complex multi-stage training schemes.

Furthermore, implementing Mean Iteration within CISO enables cooperative improvement of the detection training and pseudo-labeling, making the training results progressively more robust and effective. This technique, along with the combined training of all unlabeled and labeled data in the network, ultimately develops a comprehensive final detection model. To compare with Instant-Teaching and STAC, CISO adopts a weak-strong data augmentation strategy that leverages unlabeled data. Initially, the weakly augmented data is subjected to inference to generate a prediction score that determines the pseudo-label based on a threshold τ. Subsequently, the strongly augmented data is processed to refine the prediction scores and calculate pseudo-label related losses.

In summary, CISO follows the same loss function in Instant-Teaching and STAC, incorporating cross-entropy loss and consistency regularization loss into its framework (Sohn et al., 2021; Yu, Jiang, Wang, Cao and Huang, 2016). Eq. (6.1) details the supervised loss consists of the L1 (bounding box regression loss function) and the Lce

(classification loss function).

$$L_s = \sum_s [\frac{1}{n} \sum_i L_{ce}(P(c_i \mid \alpha(Xs)), G(c_i)) + \frac{\lambda}{n} \sum_i G(c_i) L_1(P(r_i \mid \alpha(Xs)), G(r_i))] \quad (6.1)$$

where n represents the bounding box number, i is the anchor index of images, s represents the index of the labeled image in the dataset. For each anchor point i in image, $G(c_i)$ provides the actual label assigned to this anchor, while $G(r_i)$ is the true coordinates of the label. Next, $P(r_i)$ refers to the coordinates of the bbox predicted by our model, and $P(c_i)$ computes the probability that this anchor is classified as an object.

For the unsupervised loss component, we initiate this process by obtaining a small batch of unlabeled data that has been weakly augmented. For this batch, we use Eq. (6.2) to determine the corresponding coordinates for each frame and its predicted probability distribution. Finally, by applying Eq. (6.3), the final labels output by our model are hard labels converted from pseudo-label.

$$G(c_i^u), G(r_i^u) = P(c_i, t_i \mid \alpha(Xu)) \quad (6.2)$$

$$\hat{G}(c_i^u) = argmax(c_i^u) \quad (6.3)$$

Therefore, we formalize the unsupervised loss function by Eq. (6.4), which is displayed as

$$L_u = \sum_u [\frac{1}{n} \sum_i L_{ce}(P(c_i \mid A(Xu)), \hat{G}(c_i^u)) + \frac{\lambda}{n} \sum_i (M(c_i^u) \geq \tau) L_1(P(r_i \mid A(Xs)), G(r_i^u))] \quad (6.4)$$

where u refers the index of the unlabeled image, $G(r_i^u)$ and $\hat{G}(c_i^u)$ denote the pseudo-label. Then, $M(c_i^u)$ captures the highest predicted value among these

pseudo-labels, and τ represents the threshold of the confidence level that determines which pseudo-labels are considered reliable enough to be included.

By combining Eq. (6.1) and Eq. (6.4), we arrive at the combined final loss function, shown in Eq. (6.5), λ_u is introduced as a weight parameter to adjust the unsupervised loss in the overall loss function.

$$L_{total} = \lambda_u L_u + L_s \tag{6.5}$$

6.2.2 Mean Iteration Strategy

The student model is strategically trained by the CISO using a part of labeled data, while at the same time, the teacher model begins creating pseudo-label for the unlabeled data. At this point, we carefully calculate the IoU of all pseudo-label instances (IoU) and subsequently average these IoU metrics to establish the pseudo-label generation threshold τ. Furthermore, by adopting the mean IoU value as the threshold τ, we classify the pseudo-label data into two different groups: a low-confidence group and a high-confidence group. Pseudo-label below the average threshold τ is considered unreliable, while pseudo-label above the average threshold τ is considered reliable. Next, the student model utilizes the newly generated reliable data and the original labeled data for a second round of training. After training, the teacher model steps in again to evaluate the unlabeled dataset, regenerating unreliable data and reliable data in the process. To facilitate iterative learning, CISO retains every piece of unlabeled data at each training stage of the student model, avoiding excluding any data previously classified in the pseudo-labeled dataset. A noteworthy aspect of this approach is the randomness of pseudo-label generation in each iteration, ensuring dynamic differentiation between unreliable and reliable data at each iteration.

This process allows the threshold τ to be dynamically adjusted in successive iterations. Different from traditional semi-supervised learning models, which usually tend to use pseudo-label at high thresholds such as thresholds τ equal to 0.9 and

inadvertently cause imbalance in data, CISO carefully optimizes pseudo-label data use. This not only ensures the accuracy of pseudo-label through collaborative iterations, but also greatly improves the learning efficiency and accuracy of the model. Our experiments were limited to four iterations, and the results showed that extending to the fifth iteration did not yield any new advances from the model. We will elaborate on this observation in ablation experiments. Experimental results show that our model significantly improves model performance, highlighting the efficacy of iteration strategy and pseudo-label utilization in improving model reliability.

6.2.3 Weak-strong Data Augmentation

Data augmentation methods not only expand the dataset, but also greatly enrich the amount of information learned by the model in the pseudo-label data (Kisantal, Wojna, Murawski, Naruniec and Cho, 2019; Lin, 2019). Therefore, data augmentation is also inseparable from SSL techniques using consistent regularization. When dealing with soft augmentation, to deal with possible quality problems of pseudo-label data, we adopt pre-training methods such as flipping, rotation, translation, and cropping to improve the effectiveness of labeled data. In addition, for the consistent learning strategy, we utilized the Cutmix method (Yun et al., 2019). This is the reason why Cutmix can implement soft fusion and hard fusion between two images, so that the information of the entire image can be fully utilized, and there is no need to make any modifications to the dataset during the image mixing process.

In contrast, Mixup will introduce pseudo-pixel information that may mislead model training, while Cutout will reduce training efficiency by discarding image region information. By combining the application of weak and strong data augmentation techniques, we not only expanded the size of the dataset, increased the diversity and complexity of the data, but also effectively improved model robustness and generalization ability to various perturbations, further prevents the occurrence of model overfitting (Inoue, 2018).

Taken Cutmix as an example, this method randomly selects two different samples in the dataset and fuses their partial areas to create new training images. Using U_i to represent the selected unlabeled data, the two images randomly selected by Cutmix can be expressed as $U_1 = (X_{U_1}, Y_{U_1})$ and $U_2 = (X_{U_2}, Y_{U_2})$ respectively. Finally, the newly generated image is defined as $N = (X_n, Y_n)$. This process involves removing a specific region from the first image U_1 and filling this region with the corresponding region in the second image U_2, thus achieving an efficient replacement of the selected region in U_1. In this way, the features of the two images are effectively combined, and a new training sample containing the attributes of the two original images is generated, as seen in the Eq. (6.6) and Eq. (6.7).

$$X = \boldsymbol{M} \odot X_{U1} + (\mathbf{1} - \boldsymbol{M}) \odot X_{U2} \tag{6.6}$$

$$Y = \lambda Y_{U1} + (\mathbf{1} - \lambda) Y_{U2} \tag{6.7}$$

In this section, we discuss the image samples represented by X and the image labels referred to Y, while introducing λ as a parameter for adjusting the proportion of the combined regions U_1 and U_2 in the image. Similar to the Cutmix strategy, the value of λ is limited to the range (0, 1). Additionally, we use a binary mask M to identify specific regions selected from images U_1 and U_2. In this setting, the element value in the mask matrix is set to 1, indicating that the area is selected to participate in the combination. Finally, in the equation operation, we use element-wise multiplication $\odot$ to complete the fusion of the Images.

$$r_X \sim \text{Unif}(0, W), \qquad r_W = W\sqrt{1-\lambda} \tag{6.8}$$

$$r_Y \sim \text{Unif}(0, H), \qquad r_H = H\sqrt{1-\lambda} \tag{6.9}$$

Subsequently, Eq. (6.8) and Eq. (6.9) illustrate the calculation method of the extracted mask region. We borrowed the random strategy used by Cutmix to determine the specific coordinates of the mask region $C = (r_{X,} r_Y, r_W, r_H)$, where H and W are the length and width of the image U_i respectively, and r_X and r_Y are in (0, W) and (0, H) randomly selected within the range.

In this thesis, we introduce three key hyperparameters λ, τ, and λ_u, where λ and λ_u are both set to 1.0, while τ is dynamically adjusted based on the mean IoU value. We combined Swin Transformer with CISO. The initial network weights are obtained using the pre-training model of ImageNet. We conducted a series of experiment protocols based on the 1%, 5% and 10% criteria of MS-COCO and adopted a quick learning strategy (Lin, 2014). In addition, our training settings are consistent with the methods of Instant-Teaching and STAC, and the specific details are described in Table 6.1.

Although we chose Swin Transformer as the subject of feature extraction, to ensure that the experimental results with other models can be compared fairly, we adopted Faster R-CNN as the detector. At the same time, to further verify the effectiveness of our model, we also conducted experiments using the same ResNet-50 as the backbone network as other models. These steps ensure that our findings are both innovative and easy to compare and validate with existing technologies.

Table 6.1 Training parameters of our framework.

Classes	Parameters
Initial learning rate	0.01
Momentum	0.9
Weight decay	$1e^{-4}$
Training step	180K
Learning rate decays (120K, 165K)	10

6.3 Summary

In this thesis, we introduce an innovative semi-supervised object detection (SSOD) learning method named CISO, which combines the weak-strong data augmentation strategy and the knowledge distillation technology on pseudo-label data. The core of this strategy is to fully exploit and utilize the potential of unlabeled data to significantly improve the performance of the model through an iterative learning process. Faced with the problem that the model may fall into overfitting because it cannot update pseudo-label in time, we have carefully designed a new strategy called "Mean Iteration" to reduce the risk of overfitting by continuously optimizing the generation process of pseudo-label, allowing the model to learn and adapt to information from unlabeled data more effectively.

Although our experiments are mainly based on Swin Transformer and self-attention mechanism to evaluate the performance of CISO, the method we designed has good generality and can be applied to other types of detection models. Through a series of experimental verifications on our datasets, the CISO method performs stably on the waste classification task.

It is worth mentioning that our research has not involved the precise selection of training samples, but only randomly selected training samples from the dataset. However, in actual application scenarios, there are often a variety of distribution differences between unlabeled data and labeled data, because unlabeled data may come from completely different environments from labeled data. This distribution inconsistency may affect the learning effect and generalization ability of the model. Therefore, our future research will focus on developing new training sample selection strategies, especially those take into account of the differences in data distribution, to further improve the adaptability and performance of SSOD models in various complex environments.

Finally, we speculate that our method can also be applied to other dataset, such as

MS-COCO, so we also conducted some additional experiments based on these two datasets to verify the stability and effectiveness of the model (visible in the Chapter 7.3.2 and Chapter 7.3.3). Experimental results show that CISO exhibits excellent performance, with model accuracy significantly better than other existing state-of-the-art techniques.

Chapter 7
Results and Analysis

This chapter summarizes in detail the basic experimental and ablation experimental results related to all methods proposed in this thesis. These results are presented in tables and figures.

7.1 Basic Method Results of Waste Classification

7.1.1 YOLOv7

We use Average Precision (AP), Mean Average Precision (mAP), Precision-Recall curve (PR curve), and F1 score to measure and evaluate the performance of the proposed YOLOv7 model.

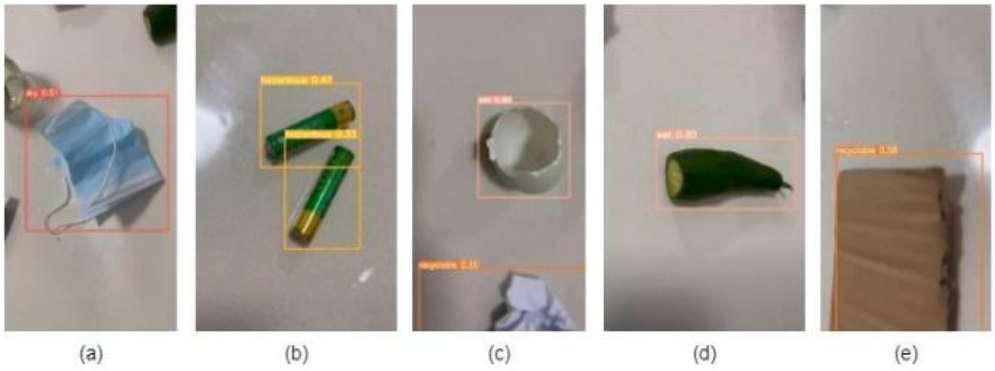

Figure 7.1 Waste detection results. (a) The result of classifying mask, classified into class “Dry”. (b) The result of classifying battery, classified into “Hazardous”. (c) The result of classifying egg shell and paper, classified to “Wet” and “Recyclable” respectively. (d) The result of classifying cucumber, classified into “Wet”. (e) The result of classifying cardboard, classified into “Recyclable”

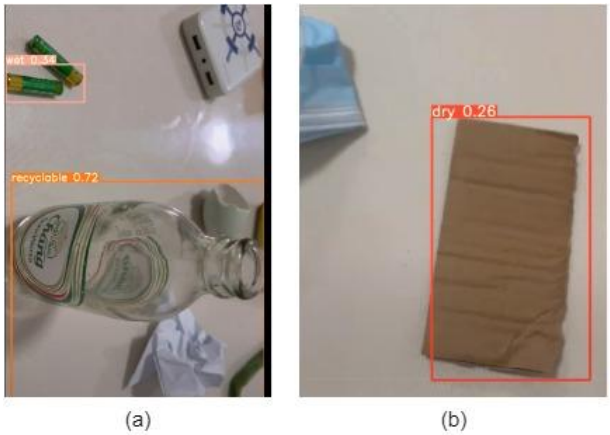

Figure 7.2 The incorrect classification results. (a) incorrect labels for classifying battery and glass bottles. (b) incorrect classification results for cardboards

We illustrate the waste detection results in Figure 7.1 as examples. All the images are taken from our waste detection videos; from them, we see all waste classes with color bounding boxes, including the class labels “Recyclable”, “Hazardous”, “Wet”, and “Dry”. Figure 7.1 depicts the waste detection with correct results. Figure 7.2 shows visual objects with wrong labels of classes. Batteries, for example, should belong to the “Hazardous” class, but the results under multiple angles show that they have been classified into the class “Wet”.

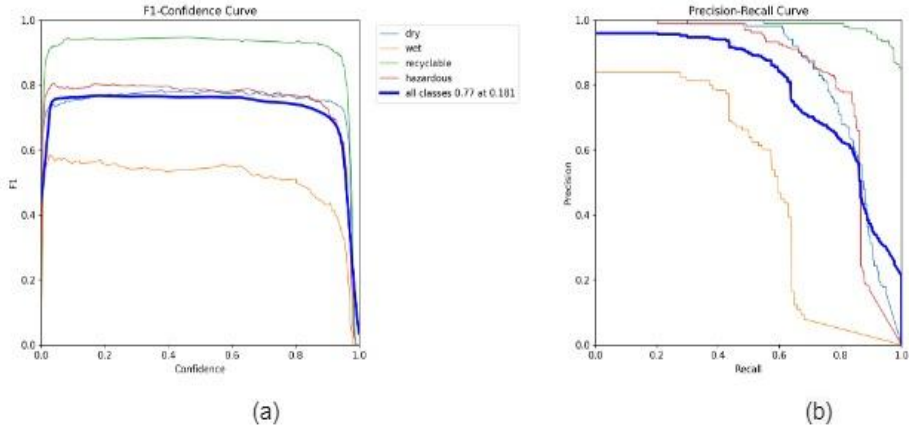

Figure 7.3 F1 score and PR curve of the waste classification

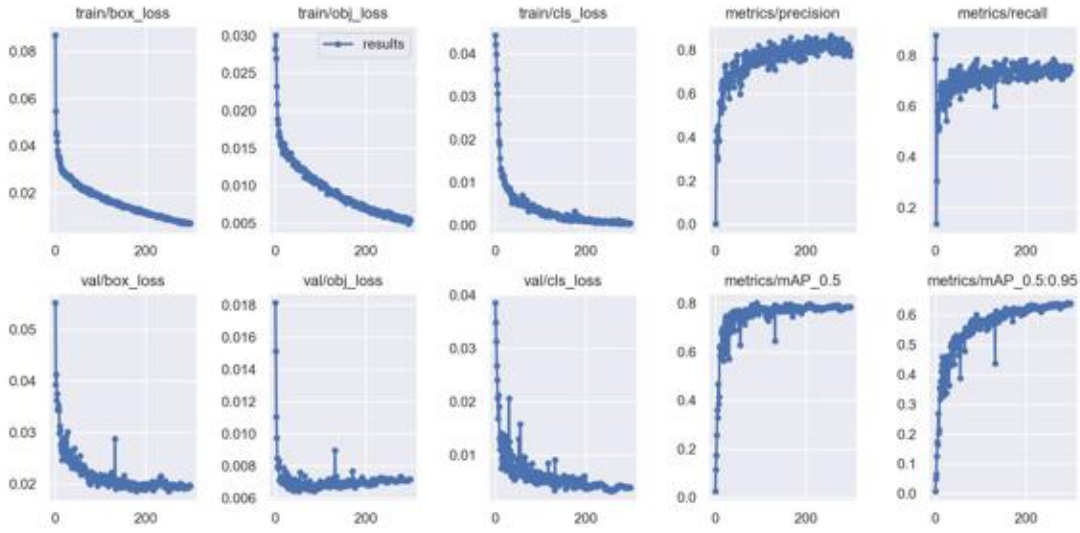

Figure 7.4 The mean average precision and loss of the waste classification

In Figure 7.3 and Figure 7.4, the model training is stable; after 300 epochs of

training, the F1 score of the model is about 0.770. Additionally, AP values for the classifications of dry, wet, recyclable, and hazardous waste are 0.841, 0.496, 0.983, and 0.829, respectively. The overall mAP of the proposed model is 0.787.

To further explore the performance of the model, we compared our proposed model with other models based on our dataset. Table 7.1 presents the comparisons of the six state-of-the-art models. The mAP of Faster R-CNN is the lowest one among the six models. The models SSD, YOLOv3, and YOLOv5 have mAPs not higher than 0.700, and the mAPs are 0.622, 0.675, and 0.691, respectively. Finally, we trained the original YOLOv7 model and obtained a mAP of 0.066 higher than that of YOLOv5. However, after we improved the YOLOv7 model, the mAP improved by 0.03 to 0.787, which was higher than the mAP of the other five models.

Table 7.1 Mean average precision results between four models.

Models	mAP	F1 score
Faster R-CNN	0.613	0.589
SSD	0.622	0.602
YOLOv3	0.675	0.656
YOLOv5	0.691	0.673
YOLOv7	0.757	0.731
Ours	0.787	0.770

To gain a deeper understanding of our proposed model, we carry out a number of ablation studies. The ablation experiment works similarly to a controlled variable method with the impact of a particular feature on the model. In deep learning, an ablation study is generally performed based on the proposed model by reducing the features to verify the necessity of the corresponding improved features.

Through ablation experiments, we verified the feasibility and validity of the improved YOLOv7 model proposed in this paper. In Table 7.2, we conducted eight

experiments, the original YOLOv7 model had the lowest mAP before the improvement. Then, we added the SPPCSPC-COOR-ASF module, which improved the experimental mAP by 0.002, but the result was not significant. Next, we introduced the EPSA module into the net, which improved the mAP by 0.015. The performance of the model was improved even more by using the Transformer block, which increased the mAP from 0.757 to 0773. However, the difference in mAP between the EPSA module and the Transformer block was only 0.001. This shows that the insertion of EPSA module and Transformer block affects the model performance.

Table 7.2 The mAP of the model in the ablation experiments.

Models	SPPCSPC module	Transformer block	EPSA	mAPs
YOLOv7				0.757
	√			0.759
		√		0.773
			√	0.772
	√		√	0.775
	√	√		0.777
		√	√	0.784
	√	√	√	0.787

In this thesis, we also conducted experiments with two-by-two combinations of the three improvement features. The mAPs of applying SPPCSPC-COOR-ASF and EPSA, SPPCSPC-COOR-ASF and Transformer block, and EPSA and Transformer block simultaneously were 0.775, 0.777, and 0.784, respectively. We see that the three improvement features are applied to the YOLOv7 model at the same time with the best results, resulting in a mAP of 0.787. Moreover, the stability of the model is also verified.

7.1.2 YOLOv8

In this experiment, Average Precision (AP), Mean Average Precision (mAP), F1 score, and Precision-Recall curve (PR curve) are the metrics we use to evaluate the accuracy and performance of our model. To make the detection results more visual, we recorded a waste detection video using the available waste. The waste detection results shown in Figure 7.5. In the detection results, we see that different categories of waste have different color bounding boxes, including “wet”, “dry”, “recyclable”, and “hazardous”. Most of the detection results are correct, but the tape in Figure 7.5(c) is missed detection, it should belong to the “dry” category.

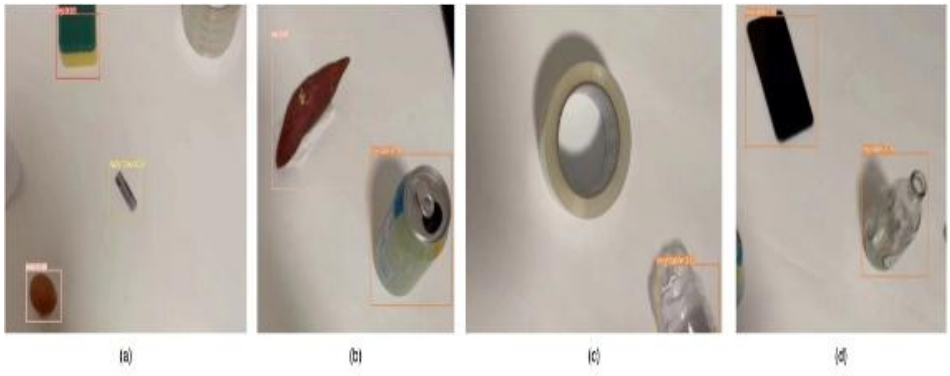

Figure 7.5 Waste detection results. (a) The result of classifying egg shell, battery, and sponge dishcloth, classified to “Wet”, “Hazardous”, and “Dry” respectively. (b) The result of classifying sweet potato and can, classified to “Wet” and “Recyclable” respectively. (c) The result of classifying plastic bottle, classified into class “Recyclable”. (d) The result of classifying phone and glass bottle, classified to “Recyclable” and “Recyclable” respectively

Furthermore, the parameters such as mAP and R for each waste class are described in Table 7.3. After 300 training epochs, the total mAP of the model is 0.856, where the mAP values of hazardous, recyclable, wet, and dry correspond to 0.927, 0.955, 0.820, and 0.720, respectively. This is consistent with the mAP values shown in Figure 7.6, reflecting the stability of our model.

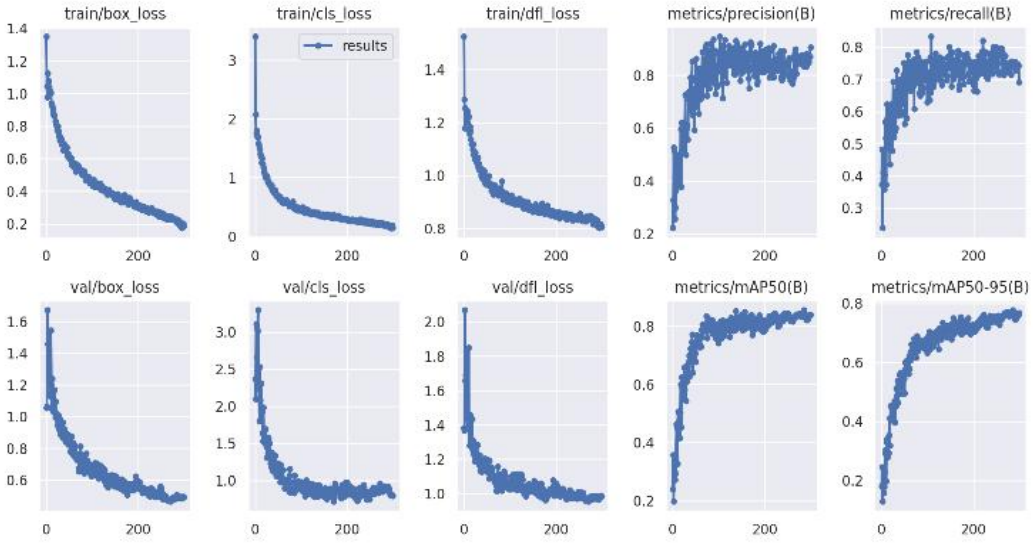

Figure 7.6 The mean average precision and loss of the waste classification

Table 7.3 The results of the model.

Class	Box	R	mAP50	mAP50:90
Hazardous	0.874	0.848	0.927	0.884
Recyclable	0.912	0.900	0.955	0.905
Wet	0.866	0.761	0.820	0.674
Dry	0.717	0.467	0.720	0.650
TOTAL	0.842	0.744	0.856	0.778

To further explore the performance of the model, we make use of our dataset to compare our method with the other five advanced models in Table 7.4. Our model has the highest mAP value of 0.856, which is higher than the original model of YOLOv8 at 0.054. while the mAP values of YOLOv5 and YOLOv7 are 0.717 and 0.759, respectively. Moreover, the mAP values of SSD and Faster R-CNN are relatively low, not higher than 0.700, with the lowest mAP of Faster R-CNN at 0.639, which is smaller than the mAP value of SSD 0.026.

Table 7.4 Mean average precision results between six models.

Models	mAPs	F1 score

Faster R-CNN	0.639	0.611
SSD	0.665	0.616
YOLOv5	0.717	0.676
YOLOv7	0.759	0.720
YOLOv8	0.802	0.768
Ours	0.856	0.790

In this thesis, we conducted a large number of ablation experiments to reduce some improved features on the proposed model in order to verify the validity and necessity of our proposed improved features (similar to the control variables). The results of the ablation studies illustrate that our model is feasible and valid. The results of our eight ablation experiments are shown in Table 7.5. The original YOLOv8 model without any changes has the lowest mAP value of 0.802.

Then, we added the data augmentation, feature fusion, and SE_ASPP modules separately. Comparing the results of these three modules, we found that adding the SE_ASPP module resulted in the largest performance improvement, with an increase in mAP of 0.022. In feature fusion, the presence increased the mAP of YOLOv8 by 0.015. Finally, data augmentation, though its introduction into the model resulted in the smallest performance improvement, it was only 0.003 lower than that of feature fusion. This indicates that data augmentation, feature fusion, and SE_ASPP are all important for the performance improvement of the model.

We also combined the three introduced features two by two to obtain three new combined features. The mAP value using both data augmentation and feature fusion strategies is 0.834. Mean AP value obtained by discarding feature fusion alone is 0.845. Then, mAP value reaches 0.851 if feature fusion and SE_ASPP are applied simultaneously. These results are lower than the mAP value of 0.856 for the model with all three features utilized, which shows that optimal results can be obtained by applying all three features.

Table 7.5 The mAP of the model in the ablation studies.

Models	Data augmentation	Feature fusion	SE_ASPP	mAPs
YOLOv8				0.802
	√			0.814
		√		0.817
			√	0.824
	√		√	0.854
	√	√		0.834
		√	√	0.851
	√	√	√	0.856

7.1.3 Swin Transformer

After the model has been trained, the waste images and waste videos are provided for the test. For a test video, we selected 20 different waste samples for observation and detected these waste samples in the video through different angles, heights, distances, lighting conditions, speeds, and quantities. The waste detection results are shown in Figure 7.7.

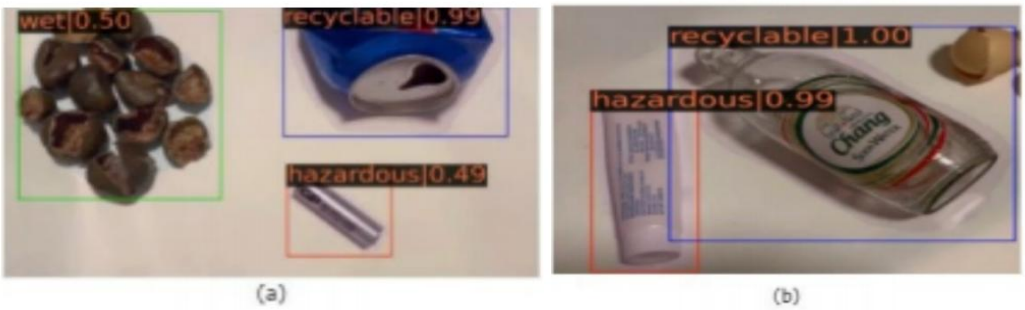

Figure 7.7 Transformer-based classification results from videos (a) the results of classifying battery, can, and chestnut, which also belong to hazardous waste, recyclable waste, and wet waste respectively (b) the classification results of glass bottle and ointment, which also belong to hazardous waste and recyclable waste

Table 7.6 The results between four models.

Class	Swin Transformer	DETR	YOLOv5	Faster R-CNN
Dry	1.00	0.83	0.70	0.89
Wet	0.93	0.91	0.71	0.85
Hazardous	0.96	0.92	0.73	0.81
Recyclable	0.97	0.98	0.79	0.93
TOTAL	0.96	0.90	0.73	0.87

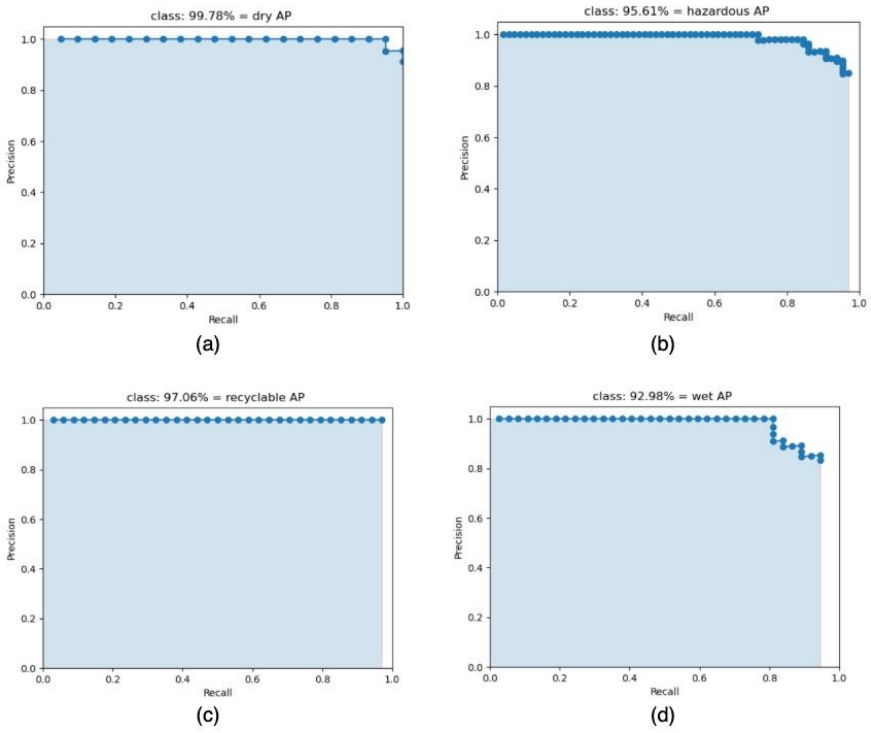

Figure 7.8 Average precisions of the four classes classification (a) the average precision of dry waste (b) the average precision of hazardous waste (c) the average precision of recyclable waste (d) the average precision of wet waste

Then, to evaluate the performance of the Swin Transformer, we quantitatively compare it with other advanced models, the comparison results are shown in Table 7.6. Regarding DETR, the mAP is 90.38%, only 6.00% lower than the Swin Transformer. Afterwards, the mAP values of YOLOv5 and Faster R-CNN are 73.25% and 86.55%, respectively. Thus, the results of Transformer architecture and attention mechanism are

better, after compared with other models, Swin Transformer has higher accuracy and better performance.

Figure 7.8 illustrates the AP rates of the four classes: “Dry”, “Recyclable”, “Hazardous”, and “Wet waste” are 99.78%, 97.06%, 95.61%, and 92.98%, respectively. Therefore, the mAP is 96.36%.

We also evaluate our model by keeping the backbone of Swin Transformer unchanged by using different algorithms. As shown in Table 7.7, after replaced Mask R- CNN with Faster R-CNN, the mAP value is decreased. Besides, for the combination with the one-stage model YOLOv5, the mAP value is 87.33%.

Table 7.7 The results between three algorithms.

Swin Transformer Models	mAPs
Swin Transformer+Mask R-CNN	0.96
Swin Transformer+Faster R-CNN	0.91
Swin Transformer+YOLOv5	0.87

Moreover, as shown in Table 7.8, if the backbone of Swin Transformer is replaced by using ResNet-50, ResNet-101, and ResNeXt-101 in the Mask R-CNN model, the accuracy is 73.93%, 71.06%, and 83.49% respectively, which indicates that using Swin Transformer as the backbone is a good choice, compared with other backbone networks, such as ResNet-50.

Table 7.8 The results of Mask R-CNN using four different backbone networks.

Deep Learning Models	mAPs
Mask R-CNN+ResNet-50	0.74
Mask R-CNN+ResNet-101	0.71
Mask R-CNN+ResNeXt-101	0.83
Mask R-CNN+Swin Transformer	0.96

To further understand the Swin Transformer model, we carried out our ablation experiments. We performed a series of comprehensive ablation studies, which cover multiple parameters such as the calculation of self- attention in window, the calculation of masked self-attention, and the position bias parameter.

Self-attention in non-overlapped windows. In the Swin Transformer model, the calculation of self-attention in window plays a decisive role in the performance of this model. In Table 7.9, we consider the impact of only calculating the self-attention within the local window on model performance. MAS denotes that if the image resolution is $h \times w$, the calculated amount of self-attention is quadratic with $h \times w$. W-MAS shows that if the image resolution is $h \times w$, the calculated amount of self-attention is linear with $h \times w$. We find that there is a significant improvement with regard to average which indicates that W-MAS is an important parameter in ensuring the performance of Swin Transformer.

Table 7.9 Influence of self-attention inside the window on model results.

MAS	W-MAS	AP	AP50
√	×	0.63	0.69
×	√	0.71	0.77

Table 7.10 Influence of masked self-attention on model results.

Masked self-attention	AP	AP50
×	0.69	0.76
√	0.71	0.77

The next experiment is related to the calculation of masked self-attention. There is a problem caused by blocking the shifted window, that is, a lot of windows will be generated, which will increase the amount of calculation of the model. By using a mask to assist in the calculation of self-attention, we reduce the amount of calculation. We investigate the validation of self-attention on the performance of the model without using a mask. As shown in Table 7.10, we see that the use of the mask can reduce the

amount of calculation, but it has limited benefit for the AP rates, the AP rate is only raised around 0.02.

Finally, the position bias parameter also affects the model performance. Whilst calculating self-attention, relative position bias is taken. We investigate the validation of self-attention on the performance of the proposed model without using the position bias. In Table 7.11, *B* is the position bias parameter. We find that the influence of position bias on the accuracy of the model is significant.

Table 7.11 The impact of parameter B on model results.

B	AP	AP50
×	0.65	0.72
√	0.71	0.77

7.1.4 Large Language Model

Utilizing the WasteNet dataset, we evaluated our proposed method against other leading-edge image classification models, primarily focusing on Average Precision (AP) as our primary metric for comparison. The outcomes, detailed in Table 7.12, demonstrate that our integrated approach—combining MiniGPT-4 with Swin Transformer—achieves an AP value of 62.20%, outperforming all other models listed. To illustrate, utilizing the Swin Transformer alone for waste image classification yields an AP of 60.70%, which is 1.50% lower than that of our combined method. The Vision Transformer model and the ResNet model follow suit, recording AP values of 59.10% and 55.30%, respectively, both trailing our method. Other models, such as ConvNeXt (Liu et al., 2022), DenseNet (Huang, Liu, Van Der Maaten and Weinberger, 2017), and EfficientNet (Tan and Le, 2019), show average precision values of 55.20%, 51.20%, and 50.50%, in that order. The VGG model (Simonyan and Zisserman, 2014), with the lowest AP at 49.90%, is positioned at the end of the ranking. Additionally, Figure 7.9 presents the loss values associated with our model.

Table 7.12 Comparisons of AP values with different models.

Models	AP
Ours	62.20
Swin Transformer	60.70
Vision Transformer	59.10
ResNet	55.30
ConvNeXt	55.20
DenseNet	51.20

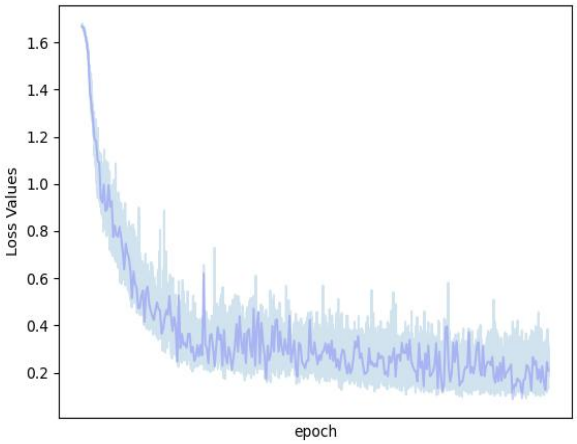

Figure 7.9 The Loss values of the model

These experimental results show that our method is effective and can significantly improve the accuracy of waste classification. It can also maintain a high level of accuracy and reliability when dealing with diverse and complex waste datasets.

To further validate the effectiveness of MiniGPT-4, we conducted ablation experiments on large language models, as shown in Table 7.13. The value of AP of our model reaches 62.20%, which is the best effect. After that, we kept the model structure unchanged and replaced MiniGPT-4 with other large language models, such as Blip (Li, Li, Xiong and Hoi, 2022) and Clip (Radford et al., 2021). The AP values are very close to the AP values of our model, but still lower than our model by 0.90% and 2.10%,

respectively. Furthermore, InstructBLIP (Dai, 2023) and LLaVA (Liu, Li, Wu and Lee, 2024) were also tested and obtained AP values of 58.80% and 56.70%, respectively. Finally, the model with the lowest AP value was Otter (Li, 2023).

Table 7.13 Comparisons of AP values with different large language models.

Models	AP
Ours	62.20
Blip	61.30
Clip	60.10
InstructBLIP	58.80
LLaVA	56.70
Otter	53.30

Afterwards, we experimented with the selection of pre-trained language models as well, and the results are shown in Table 7.14. If RoBERTa is replaced by BERT (Devlin Chang, Lee and Toutanova, 2023), XLNet (Yang, 2019), and ELECTRA (Clark, Luong, Le and Manning, 2020), AP values of 60.10%, 56.90%, and 54.20% are obtained, respectively, which are lower than those of our model.

Table 7.14 Comparisons of AP values with different pre-trained language models.

Models	AP
Ours	62.20
MiniGPT-4 + Swin Transformer +BERT	60.10
MiniGPT-4 + Swin Transformer +XLNet	56.90
MiniGPT-4 + Swin Transformer +ELECTRA	54.20

We conjecture that in the context of our task, pre-trained language models are required to more accurately understand the text of image descriptions generated by large language models, while RoBERTa and BERT perform well on the task of processing and understanding the text that can easily be used for new classification tasks. The

experiments verify that RoBERTa is the best choice. In summary, MiniGPT-4 has the potential to lead the innovation in waste classification, and we will continue to explore the application of large language models, such as GPT-4, to waste classification in our future work.

In this section, we finally choose “Describe these images for waste classification in detail” as the prompt to input MiniGPT-4. To verify whether this prompt is suitable for our waste classification task, we also tested another prompt, “Describe these images in detail”, and the results can be seen in Table 7.15. If we utilize “Describe these images in detail” in the prompt, the AP value will be reduced by 61.30%. This may be relevant to our dataset and classification task.

Table 7.15 Comparisons of AP values with different prompts.

Prompts	AP
Describe these images in detail	61.30
Describe these images for waste classification in detail	62.20

Table 7.16 Comparisons of AP values of prompts with different lengths.

Prompts	AP
Describe these images for waste in one sentence	57.70
Describe these images for waste classification in detail	62.20

From Figure 4.12, we also intuitively see that the descriptions obtained by inputting the two prompts are very different. The description given by the “Describe these images in detail” lacks object information such as corn cobs and plastic bags, which may have a negative impact on the accuracy of the model. We also expanded our investigation to include the effect of prompt length on model performance. When limiting the description of a waste image to a single sentence, we found that essential

information was often omitted—examples include specific items like lettuce, kiwifruit, and cans. This omission can lead to a reduction in model accuracy. As indicated in Table 7.16, the Average Precision (AP) value decreases by approximately 4.50% when using a single-sentence prompt compared to the more detailed prompt, "Describe these images for waste classification in detail." Figure 7.10 illustrates these experimental outcomes.

Figure 7.10 Examples of the detailed image description generated by MiniGPT-4 with short prompts

7.2 Data Augmentation Results for Waste Classification

7.2.1 Basic Results

We adopted a method based on ZeroWaste dataset, aiming to optimize the model in real waste classification scenarios. In order to ensure the validity and fairness of the experimental results, we strictly followed ZeroWaste's baseline comparison rules. As shown in Table 7.17, using ResNet-50 as the backbone network of the U-Net method (semi-supervised learning mode), we achieved a mean IoU value of 55.37% on the test set. This result is better than ZeroWaste baseline experiment results (the result is about 3.74% higher). In contrast, if we abandon ResNet-50 and switch to EfficientNet, we find that the IoU value decreases by 0.39%. In addition, for a visual comparison, we also utilized our method to DeepLabv3+ (Chen, Zhu, Papandreou, Schroff and Adam, 2018), and successfully achieved a mean IoU value of 54.77% (DeepLabv3+ is the baseline method as ZeroWaste, which uses ResNet-101 as the backbone network). This

is an increase of 3.14% from the original baseline. This result fully demonstrates the effectiveness of our method in improving the mean IoU value.

Furthermore, we compared our results with some other advanced methods, including CCT (Ouali, Hudelot and Tami, 2020), AugSeg (Zhao et al., 2023), UniMatch (Yang, Qi, Feng, Zhang and Shi, 2023), EPS (Lee, Lee, Lee and Shim, 2021), and ReCo (Liu, Zhi, Johns and Davison, 2021), according to ZeroWaste's experimental framework. In these comparisons, the mean IoU values of ReCo, AugSeg and UniMatch on the semi test sets are 44.12%, 53.88%, and 54.65% respectively, which are all lower than our results. For EPS and CCT, their mean IoU value is less than 33%, which is far less than the high level of 55.37% we achieved. In addition, considering the performance improvement potential of the Transformer model in classification and segmentation tasks, we further tested the Swin Transformer and CLUSTERFORMER method (Liang et al., 2023), which achieved IoU values of 53.21% and 52.76% respectively. These test results show that our strategy is also applicable to Transformer-based models.

Overall, our research project not only achieved remarkable results in improving model performance in waste classification tasks, but also demonstrated the superiority of our method through comparison with a series of other advanced methods. Furthermore, we further validate the broad applicability and flexibility of our approach by exploring the impact of different network architectures on model performance.

Table 7.17 The results of different models.

Method	Supervision	Validation	Test
Ours (U-Net+Resnet-50)	semi	49.27	55.37
Ours (U-Net+EfficientNet)	semi	49.01	54.98
Ours (DeepLabv3+)	semi	48.89	54.77
CLUSTERFORMER	semi	47.95	53.21
Swin Transformer	semi	46.98	52.76
U-Net	full	46.02	51.88

DeepLabv3+	full	45.61	52.30
DeepLabv3+	semi	46.13	51.63
UniMatch	semi	48.53	54.65
AugSeg	semi	47.12	53.88
Reco	full	51.30	52.28
Reco	semi	49.49	44.12
CCT	full	30.79	29.32
CCT	semi	28.70	32.49
EPS	weak-f	13.75	13.91

After adopting the DeepLabv3+ method, we observe that the mean IoU value is only 0.6% lower than that using U-Net method. Although this performance improvement is not significant, we decided to choose the U-Net model as the core architecture of our semantic segmentation network. The main reason for choosing U-Net is its unique encoder-decoder structure. This design can not only restore the spatial details of the image effectively, but also capture the global features of the image, making it suitable for processing smaller datasets (DeepLabv3+ is more suitable for processing larger datasets).

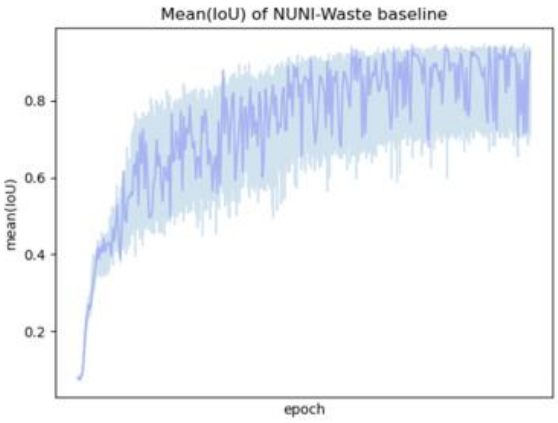

Figure 7.11 The mean(IoU) values of NUNI-Waste model

Through our experimental verification, U-Net demonstrates excellent performance

and advantages in our application scenarios. In terms of result display, Figure 7.11 and Figure 7.12 show the performance of our method in terms of mean IoU value and loss plots in detail. These figures record every change in IoU and loss values. Through the curves, we can clearly observe the overall trend of model performance. In order to better highlight the key changes in these curves, we also added a curve with a darker color and less undulating fold to approximately represent the changing trends of the mean IoU and loss values, making the changes easy to understand.

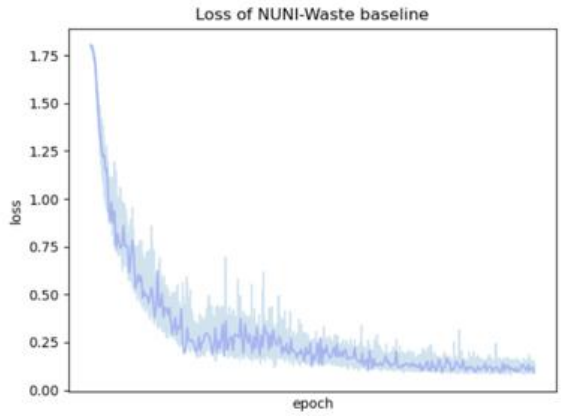

Figure 7.12 The loss values of NUNI-Waste model

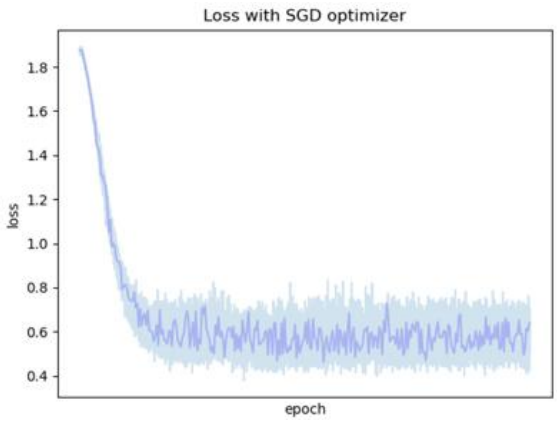

Figure 7.13 The loss values of NUNI-Waste model with SGD optimizer

In terms of optimization strategy, we also experimented with using the SGD optimizer. Although in theory, the Adamw optimizer performs better than the SGD optimizer in terms of efficient, the SGD optimizer can usually provide more stable training results in practical applications. However, in our model experiment, the performance of the SGD optimizer unexpectedly did not meet expectations. Compared

with using the Adamw optimizer, the loss value caused by SGD is higher on average. This phenomenon is reflected in Figure 7.13. This finding prompted us to re-evaluate the strategy of optimizer selection in future work in order to find a solution more suitable for our model needs.

7.2.2 Ablation Studies

Analysis of the Adaptive Weighted Loss Function

The loss function plays a key role in evaluating the difference between the model's prediction results and the true value in deep learning, and its selection has a direct impact on the success or failure of model training. Therefore, we innovatively designed an adaptive weighted loss function and verified its effectiveness through a series of experiments. The results are shown in Table 7.18. After applying the loss function we designed, the mean(IoU) of the model is improved by 1.25% compared to the case without this loss function. This result proves that by introducing a simple weighting strategy, we not only improve the performance of the model, but also enhance the model's ability and efficiency to process data.

Table 7.18 Experimental results related to adaptive weighted loss function.

Adaptive weighted loss function	Mean (IoU)
w/o	54.12
w/	55.37

Analysis of the *w* parameter

In this thesis, we adopt the ZeroWaste dataset, which covers four waste classifications, including “Metal”, “Rigid Plastic”, “Cardboard”, and “Soft Plastic”. To adapt to the requirements of semantic segmentation tasks, we additionally consider background as an independent category, bringing the total number of categories to five. However, after our analysis of the dataset, it was found that the two classes of Rigid Plastic and Metal have relatively small amounts of data, accounting for only 16.6% and 3.6% of the total

data volume respectively. This finding points to a significant imbalance in the dataset. In the initial training stage, we observed that the mean(IoU) of these two categories was only about 20.00%. This performance was far lower than other categories, which significantly pulled down the overall training effect of the model.

To address this challenge, we design and implement an adaptive weighted loss function that solves the data imbalance problem by assigning higher weights to these two categories. We conduct ablation studies on the weight value *w* in the adaptive weighted loss function, initially setting *w* value for all categories to 1.0. After the model was trained under this setting, it achieved a mean IoU value of 52.69%. We expect that the mean IoU value will not increase linearly as *w* value increases. In particular, if only Rigid Plastic and Metal are given too high a weight, it may lead to overfitting of the model. Therefore, we avoid setting too high weight values for these two classes, and the specific values are shown in Table 7.19. Our experimental results show that when the w value is set to 3, the model achieves the best training effect, with a mean IoU value as high as 55.37%. However, when the w value increases to 5, the mean IoU value drops to the lowest, 52.60%, which is only 0.09% higher than the mean IoU value when the w value is 1. Then, the mean IoU value when the w value is 3.5 is 54.02%, which is 1.35% lower than the highest value, which further confirms our hypothesis that the w value cannot be too high.

Table 7.19 Experimental results related to *w* parameter.

w value	Mean(IoU)
1	52.69
2	53.07
2.5	53.55
3	55.37
3.5	54.02
4	53.18

5	52.60

Considering that there is a difference in the amount of data between “Rigid Plastic” and “Metal”, we assign different weight values to them, such as setting the w value to 2 and 3 or 2 and 4 respectively. However, this more detailed weight adjustment did not bring about the expected increase in mean IoU value, but stabilized at about 53%, which was contrary to our initial expectations. We speculate that this may be because the difference in data volume between the two categories is not significant, and when the difference in data volume is small, the simplified weight value allocation strategy is not enough to significantly improve the mean IoU value.

In future research, we plan to deeply explore the impact of different weight value settings on model performance and further optimize our adaptive weighted loss function. Figure 7.14 represents how the model loss changes under some different weight settings.

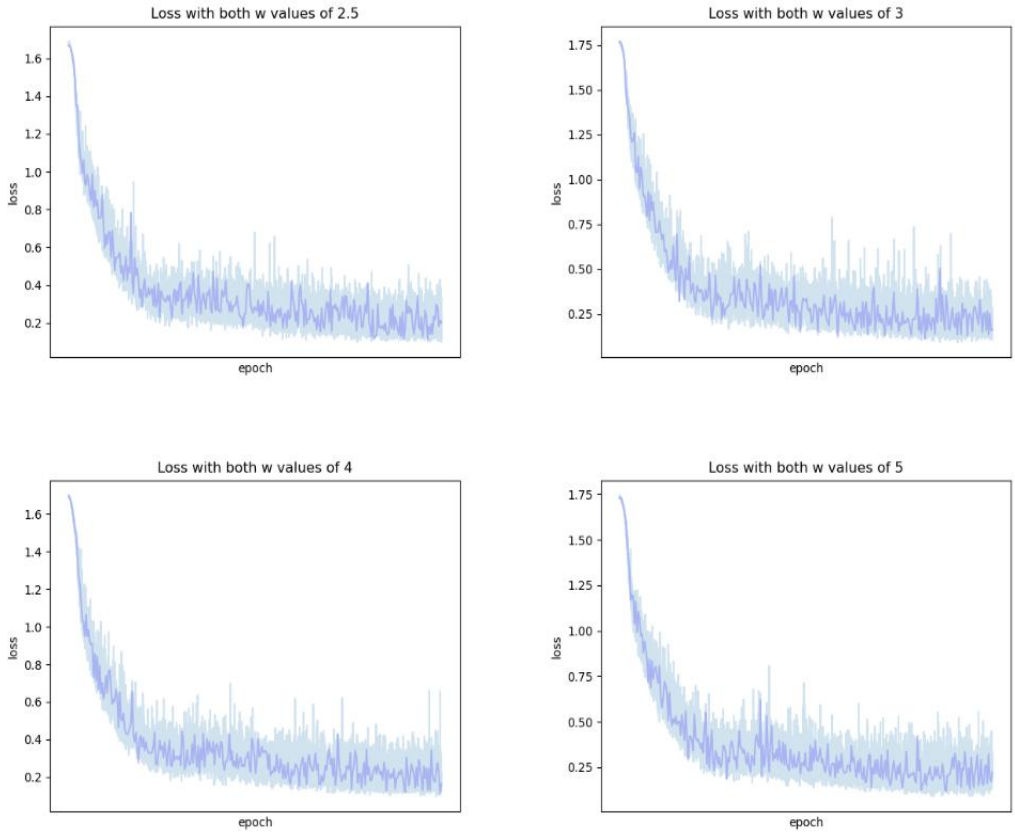

Figure 7.14 The loss values with *w* values of 2.5, 3, 4, and 5 respectively

Analysis of the Adoption of Different Data Augmentation Strategies

Furthermore, we conduct a comprehensive discussion of non-uniform data augmentation technology and carry out a detailed comparative analysis with several other data augmentation strategies. In Table 7.20, we summary the results of each experiment in detail. Among them, the non-uniform data augmentation technology we implemented showed the best effect, and its mean(IoU) value reached a high of 55.37%. Thereafter, to further verify the effectiveness of our method, we chose to keep using the U-Net model as the baseline, and tried to conduct experiments using different data enhancement methods such as Mixup, Cutmix, and Cutout. The mean IoU values obtained by these methods were 54.08%, 53.86%, and 52.21%, respectively, which cannot surpass our method.

Table 7.20 Experimental results related to the adoption of different data augmentation strategies.

Methods	Data augmentation strategies					Mean(IoU)
	ZeroWaste baseline	Non-uniform data augmentation	Cutout	Cutmix	Mixup	
DeepLabv3+	√					51.63
		√				52.88
Ours (DeepLabv3+)	√					53.52
		√				54.77
Ours (Unet)	√					52.88
		√				55.37
			√			52.21
				√		53.86
					√	54.08

To ensure a fair comparison, we also conducted a special experiment and selected DeepLabv3+ as the segmentation network to perform the task. It should be pointed out that the main purpose of this experiment is to directly compare our non-uniform data augmentation method with the strategy adopted in ZeroWaste, so we did not test Mixup, Cutmix, and Cutout in this experiment. Experimental results confirm the significant advantages of our non-uniform data augmentation method, which can achieve an overall improvement of approximately 2.49% in the mean IoU value of the model. In addition, it is also verified that choosing a suitable data augmentation method has a crucial impact on optimizing model performance. Our research provides a valuable reference for future applications of data augmentation techniques in semantic segmentation and other related deep learning fields.

Analysis of the Initial Offset Value

In this section, we detail the determination of the optimal initial values of 70 for the non-uniform offset data augmentation technique through ablation study. This result is based on the evaluation of model performance under different initial value settings. As shown in Table 7.21, when the initial value is established at 30, the mean (IoU) of the model will drop to the lowest, only 51.97%. At the same time, increasing the initial value to 100 will also cause the mean(IoU) value to decline. Figure 7.15 visually compares the difference in mean(IoU) value when the initial value is set to 30 and 100, showing the performance changes under different settings.

Table 7.21 Experimental results related to initial offset value.

Initial offset value	Mean(IoU)
30	51.97
50	53.12
70	55.37
100	53.93

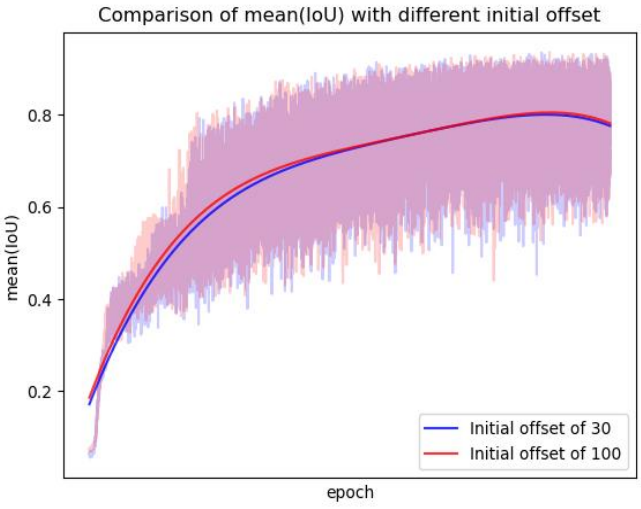

Figure 7.15 The IoU values with different initial offset values

We speculate that if the initial value of our method is too small or too large, it may have a negative impact on model performance. In the case where the pixel value of an image usually ranges from 0 to 255, a small or a large initial value may prevent the model from capturing the global information of the image, making the network unable to fully utilize the image features, thereby reducing the model's recognition ability. Thus, in our study, after considering various factors, we decided to use 70 as the best initial value for non-uniform offset data augmentation. By choosing appropriate initial values, we aim to maximize the positive impact of data augmentation techniques on model performance, thereby ensuring model accuracy.

Analysis of the Nonuniform Offset Augmentation Technology

The nonuniform data augmentation technology introduced in our research includes both of changes pixel position and adjustment in pixel color. Based on this, we conducted a detailed comparative analysis of these two different augmentation methods. When using the non-uniform offset augmentation technology, compared with the baseline, the mean IoU value of the model increased by 1.08%. When using non-uniform color data augmentation alone, we observed that the mean IoU value of the model can reach 54.39%, which is an improvement of 1.41% compared to the ZeroWaste baseline.

According to the data shown in Table 7.22, our method is indeed effective, and in terms of improving model performance, the method of simulating lighting effects is more effective than the method of simulating polymorphism changes. The cause of the performance difference between these two data augmentation strategies is a question worthy of further investigation, and we suspect that it may be related to the difference in hyperparameter configuration between these two methods. Therefore, in future work, we plan to conduct more in-depth optimization and adjustment of the model to further improve the accuracy and generalization ability of the model through more refined hyperparameter settings.

Table 7.22 Experimental results related to non-uniform offset augmentation technology.

Non-uniform data augmentation	Mean(IoU)
Non-uniform color data augmentation	54.29
Non-uniform offset data augmentation	53.96
Both	55.37

7.3 Semi-supervised Learning Results of Waste Classification

7.3.1 Basic Results

For this experiment, the dataset we applied is WasteNet. While evaluating model performance, we adopted mAP as the main metric. Experimental results show that the model proposed in this study achieved a mAP value of 58.86%. Figure 7.16 shows the mAP value obtained during the training process. In addition, we also compared our results with existing research. It can be seen that the semi-supervised model is effective, but it still needs further optimization and adjustment to reach the industry-leading level.

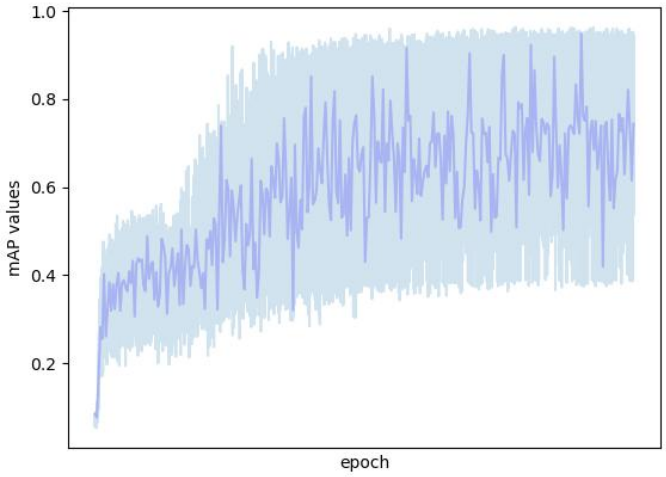

Figure 7.16 Mean average precision results of our model

In recent years, the field of semi-supervised object detection has gradually become a research hotspot. To further explore the effectiveness and stability of the CISO model, we also conducted our experiments based on the MS-COCO dataset. Besides, we have carefully compared our method CISO with some other leading semi-supervised object detection technologies, and detailed listed AP and mAP performance indicators under different experimental settings, the results are shown in Table 7.23. Ours (CISO) indicates that we applied Swin Transformer as the backbone network. Then, it shows that the ResNet-50 was selected as the backbone network. Through comparative analysis, it can be clearly seen that our CISO method performs best among all compared semi-supervised object detection methods, achieving remarkable performance, which fully proves that mean threshold and collaborative iteration strategies are effective in improving semi-supervised object detection performance.

Table 7.23 Experimental results related to different models using MS-COCO.

Method	1%	5%	10%
Supervised	9.05±0.16	18.47±0.22	23.86±0.81
CSD	10.20±0.15	18.90±0.10	24.50±0.15
STAC	13.97±0.35	24.38±0.12	28.64±0.21
DETReg (Bar et al.,	14.58±0.30	24.80±0.20	29.12±0.20

2022)			
Instant Teaching	18.05±0.15	26.75±0.05	30.40±0.05
ISMT (Yang et al., 2022)	18.88±0.38	26.37±0.24	30.53±0.52
Unbiased Teacher	20.75±0.12	28.27±0.11	31.50±0.10
Soft Teacher	20.46±0.39	30.74±0.08	34.04±0.14
LabelMatch	25.81±0.28	32.70±0.18	35.49±0.17
HT (Tarvainen and Valpola, 2017)	16.96±0.36	27.70±0.15	31.61±0.28
Ours (CISO*)	21.04±0.18	29.50±0.21	34.20±0.12
Ours (CISO)	22.00±0.17	30.90±0.15	36.20±0.26

More specifically, as shown in Table 7.23, the mAP value of our method reaches 36.20, which is 2.16 higher than mAP value of 34.04 of the Soft Teacher result (Xu et al., 2021), under 10% protocol. After that, the mAP value of our method reached 30.90 under 5% protocol. Finally, CISO was proposed to increase the mAP value of the Soft Teacher method by 1.54 to 22.00. At the same time, compared to the latest semi-supervised learning method LabelMatch (Chen et al., 2022), our mAP value under 10% protocol is 0.71 higher. Even when using ResNet-50, CISO outperforms all other compared models, with mAP values of 34.20, 29.50, and 21.04 at 10%, 5% and 1% data protocols respectively. In addition, when using Swin Transformer with self-attention mechanism as the backbone of the model, our CISO method also shows good adaptability and superiority.

It is worth noting that as the amount of available labeled data increases, CISO shows increasingly obvious performance improvements, especially in the transition from the 1% protocol to the 10% protocol, where the mAP improvement increases from 1.54 to 2.16. This result shows that releasing pseudo-label data into unlabeled data has a positive impact on model performance, possibly because this strategy increases the chance that the model will reuse valid pseudo-label data in iterations. This hypothesis

will be further explored in future research. In addition, Figure 7.17 shows our prediction results. Our research provides valuable experience for future research in this field.

Figure 7.17 The prediction results

7.3.2 Ablation Studies

Analysis of the Number of Iterations

In this section, we will explore the specific impact of different iteration numbers on model performance. To this end, we choose to test the model under the MS-COCO dataset protocol containing 10% labeled data and 90% unlabeled data. The detailed results of the experiments are recorded in Table 7.24. In our experiments, a total of six rounds of testing were conducted, with the number of iterations set to 1, 2, 3, 4, 5 and 6 times respectively. We found that as the number of iterations increased from 1 to 6, the overall performance of the model improved significantly. However, it is worth noting that when the number of iterations reaches 5, the model performance begins to stabilize, and the performance improvement brought by increasing the number of iterations becomes very limited.

Specifically, after the 6th iteration, the mAP value of the model only increased by 0.06. Based on this, we concluded that setting the number of iterations to 4 can enable the model to maintain optimal performance while ensuring high efficiency. We not only

revealed the positive impact of increasing the iteration number on model performance, but also found that when the iteration number increases to a certain level, the effect of performance improvement gradually weakens. Future work will further explore why the mAP value does not change when the number of iterations reaches a certain level, and how to adjust the iteration strategy according to specific task requirements and data characteristics, to ensure the accuracy and practicality of the model.

Table 7.24 Experimental results related to the number of mean iterations.

The number of mean iterations	mAP
1	27.40
2	29.80
3	33.60
4	36.20
5	36.40
6	36.46

Analysis of the Strong Data Augmentation strategy

In the process of exploring the optimization of semi-supervised object detection models, we recognized the critical impact of data augmentation strategies on model performance. Therefore, in the CISO model, we adopted the weak-strong data augmentation method. Some data augmentation strategies can be seen in Figure 7.18. In particular, solid data augmentation has a positive effect that cannot be ignored in improving the performance of the model. In order to achieve a fair comparison baseline, we introduced the Cutmix and also considered the combined strategy of Color+Cutout.

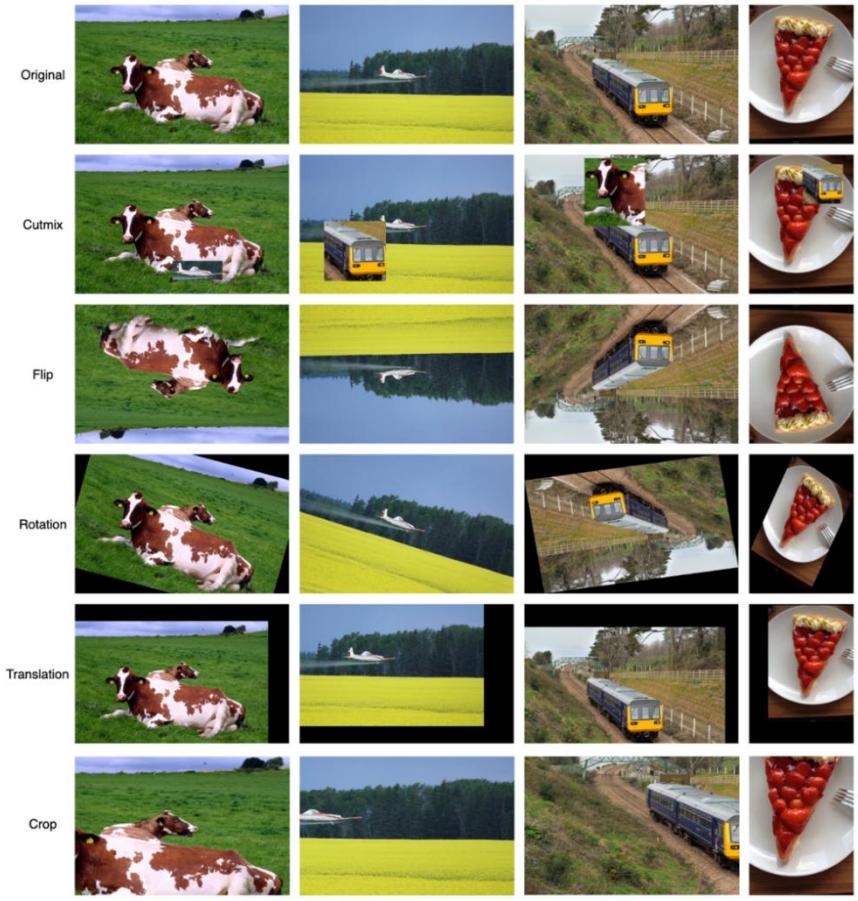

Figure 7.18 Visualization of weak-strong data augmentation strategies

As summarized in Table 7.25, we compare the mAP performance of our models using different robust data augmentation methods in detail, which is based on 5% protocol. When applying Geometric and Color+Cutout methods, our CISO method achieves a limited mAP improvement of only 1.26. However, after the Cutmix strategy was introduced, the performance of the model was significantly improved. Compared with utilizing the Mosaic and Mixup strategies, the mAP value increased by 0.50. Then, the CISO model taking use of the Cutmix data augmentation method finally achieved the highest mAP value of 29.70. This finding confirms the effectiveness of the Cutmix strategy in improving the performance of models. Through this series of experiments, we not only gained a deep understanding of the impact of different data augmentation strategies on the performance of semi-supervised object detection, but also explored effective ways to improve model accuracy.

Table 7.25 Experimental results related to strong data augmentation strategy.

Methods	Strong data augmentation strategies					mAP
	Color+Cutout	Geometric	Mixup	Mosaic	Cutmix	
STAC	√	√				23.14
Instant Teaching	√		√	√		25.60
CISO	√	√				24.40
	√		√	√		29.20
	√				√	29.70

Analysis of parameter τ

The parameter τ plays a crucial role in the semi-supervised object detection task, and its setting method directly determines the performance of the model. Different from the fixed τ value used in other semi-supervised object detection models, we propose to set τ as a dynamically changing parameter, and generate pseudo-label by setting the τ value to be greater than or equal to the mean value of the data. This experiment is based on 10% MS-COCO protocol. Moreover, our method takes into account the dynamic changes of unreliable data and reliable data in each iteration process of the model, allowing us to adjust the τ value according to the actual situation, thereby optimizing the generation process of pseudo-label.

Through the data analysis in Table 7.26, we found that when the τ value is dynamically averaged, the model can achieve optimal performance, with a mAP value as high as 36.20. In addition, we also observed that as the τ value decreases, the mAP value of the model also shows a downward trend. This result supports our initial hypothesis: dynamically adjusting τ that can effectively improve the quality of pseudo-label, thereby enhancing the performance of the model. Finally, besides the mean value, exploring more appropriate dynamic τ value setting methods to further improve the performance of model will become the focus of our future research.

Table 7.26 Experimental results related to different parameter τ.

τ	mAP
0.30	29.4
0.50	31.60
0.70	33.60
0.90	34.80
Mean (IoU)	36.20

Analysis of parameter λ_u

Furthermore, we explore the impact of introducing the parameter λ_u into the loss function on model performance. The specific experiments were performed under the 10% MS-COCO protocol. By dynamically adjusting the mean value of the confidence threshold τ, the impact of λ_u (including 0.25, 0.50, 1.00, 2.00, 3.00, and 4.00) on model performance was tested. According to our experimental results, as presented in Table 7.27, when λ_u is set to 1.0, the best performance of the model can be obtained. However, if λ_u increases to 2.0, although the model performance decreases, the mAP value still reaches 35.80, which is only a decrease of 0.40 compared with the highest value of 36.20. It is worth noting that if λ_u is set to other values, the performance of the model generally decreases, especially if λ_u is 0.25, the mAP value decreases most significantly, with a decrease of 5. This finding shows that our proposed model is relatively robust to changes in λ_u.

Table 7.27 Experimental results related to different parameter λ_u.

λ_u	mAP
0.25	30.20

0.50	32.50
1.00	36.20
2.00	35.60
3.00	32.90
4.00	31.40

Analysis of Mean Iteration Strategy

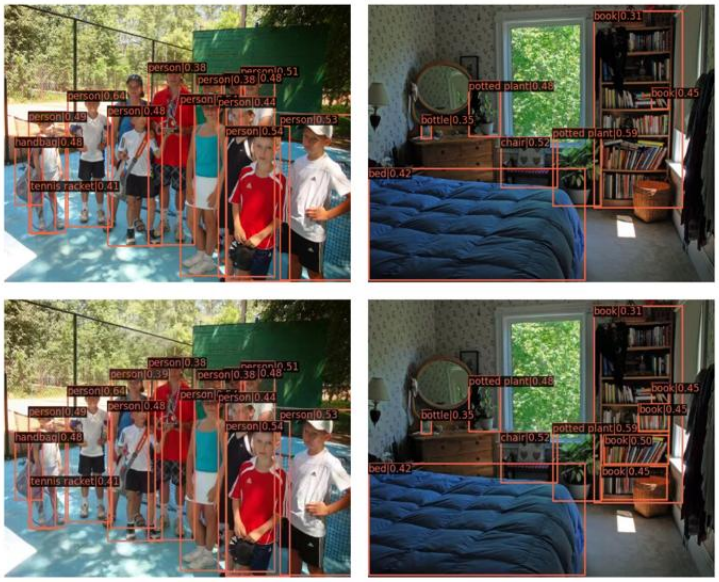

Figure 7.19 The pseudo-label visualization effect of unlabeled data

Table 7.28 Experimental results related to mean iteration.

Mean iteration	mAP
	33.10
√	36.20

Additionally, we also introduce a Mean Iteration strategy, which is executed based on the dynamically adjusted mean τ value (under 10% MS-COCO protocol). It focuses on reintegrating the extracted pseudo-label into the unlabeled data in each iteration, aiming to maximize the utilization of unlabeled data to improve pseudo-label quality.

In Table 7.28, if the Mean Iteration strategy is not adopted, the mAP value of the

model is 33.10, which is lower than the result after applying the Mean Iteration. Figure 7.19 further shows the pseudo-label visualization effect of unlabeled data depending on whether the Mean Iteration strategy is adopted or not. We observe that adopting the Mean Iteration can effectively generate more accurate pseudo-label, thereby significantly improving the overall performance of the model. In summary, by in-depth analysis of the impact of Mean Iteration strategy, our research demonstrates the effectiveness of dynamic adjustment strategies in improving model performance.

Chapter 8
Conclusion and Future Work

In this chapter, we summarize the contribution of this thesis to the field of waste classification and details the limitations of this research. Afterwards, directions for future work are also proposed, including adjustment of the dataset and further optimization of the model structure.

8.1 Conclusion

In this thesis, we conduct research and innovation from three aspects on the issue of waste classification. These three aspects are waste datasets, data augmentation strategy, and semi-supervised learning models. We know that complex and changing environmental conditions, mutual obstruction of waste objects, and various types of waste are the current difficulties in waste classification. This makes collecting and annotating a large number of waste images a time-consuming and labor-intensive process, as the dataset needs to be diverse enough to cover different waste types and environmental conditions.

Therefore, we collected two domestic waste datasets, namely WasteData and WasteNet. Both datasets are collected according to waste classification standards and have four waste categories, namely dry waste, wet waste, recyclable waste, and hazardous waste. Each waste category covers several different waste objects. Taken recyclable waste as an example, cardboard, paper, metal, plastic, and glass are all included in the recyclable waste category, which ensures the diversity, comprehensiveness, and richness of the dataset. Among them, the WasteData dataset contains 1,560 waste images, and its characteristic is that each image contains only one waste object. The dataset in this case is easier to process, can reduce interference during model recognition. It is also beneficial to improving the accuracy of the model that can achieve faster training and verification of the model. The next is WasteNet dataset, which has a total of 1,326 images. This dataset is different from WasteData in that each image is occupied by different waste objects. And these waste objects are stacked and twisted, which is close to the waste state in the real world. This kind of dataset is closer to the real scene, because in actual applications, waste is often mixed together, and there may be mutual occlusion. The model trained using this kind of dataset has stronger generalization ability and can better cope with complex real-world environments. Our proposed dataset alleviates the problem of lack of diversity in waste datasets and helps improve the generalization ability and practicality of waste classification models.

Besides, if a small amount of waste data can be employed to train the model and achieve the desired result, the cost of manual collection and labeling of waste data can be greatly saved. Therefore, the starting point for the improvement of data augmentation methods and training models in this thesis is semi-supervised learning. Nonuniform color data augmentation and non-uniform offset data augmentation are proposed. The function of the nonuniform color data augmentation is to simulate natural light to change the color in the image. In the process of collecting waste data, we can observe that some waste images will show instant light changes. For example, a waste image may appear dark in the upper left corner and bright in the lower right corner, and the training of the model will also be affected by these factors.

Then, nonuniform offset data augmentation changes the shape of waste in the image, such as distortion, to simulate the real waste state. Thus, we can utilize nonuniform data augmentation to expand the number of waste datasets and simulate real-world scenarios, which helps to improve the generalization ability and robustness of the model. Moreover, faced with the increasing cost of misclassification, we design an adaptive weighted loss function that allocates weights according to the data volume characteristics of different categories. Our experimental results show that our proposed data augmentation technology can help the model improve accuracy and correct for waste images recognition ability.

Finally, we propose CISO, an object detection model structure based on semi-supervised learning. This method can learn data features through pseudo-label and a large amount of unlabeled data, and in some cases can show higher robustness to noise. Pseudo-labeling, mean iteration, and data augmentation are the main strategies covered by this model. Our experimental results show that CISO performs well and the model accuracy is improved. The ablation experiments also prove the robustness of our model. CISO can effectively utilize limited labeled data and combine it with a large amount of unlabeled data to significantly reduce the workload of manual collection and labeling of data without sacrificing model performance.

Our research results not only provide effective solutions for waste classification and improve the automation level of waste treatment, but also have great significance in promoting environmental protection, resource utilization, and cost saving, thereby achieving sustainable development of resources.

Although our research work has achieved pretty rich results, there are still a few challenges and limitations as follows:

(1) Due to the wide variety of waste types, currently, the dataset we propose does not include all waste objects, such as durian peel (e.g., durian peel is not wet waste, it belongs to the dry waste category). In addition, for WasteData, though its advantage is that it can reduce interference during model recognition, but since there is only one object in each image, it is more suitable for simple waste classification tasks and can quickly complete model training. Furthermore, though WasteNet is closer to real scenes, it is more suitable for complex scene recognition. Therefore, both datasets also have room for improvement.

(2) As for the nonuniform data augmentation technique, we believe that it still has scope for further enhancement. Currently, non-uniform color data augmentation only simulates natural light. If there are other situations, such as changes in the color of the waste object itself, these also need to be considered. In addition, the nonuniform data augmentation in this thesis has an initial value. This initial value is the best value obtained from the experiments based on the dataset we created. While changing the dataset or using this method on other classification tasks, multiple experiments are required to obtain the most appropriate initial value. Finally, our adaptive weighted loss function, while effective, simply assigns the weights of different waste categories. Perhaps there are more effective and appropriate loss function algorithms for waste classification that can be studied.

(3) For the CISO model, our study did not involve the precise selection of training samples, but only randomly selected training samples from the dataset. However, in actual application scenarios, there is often a certain distribution

difference between unlabeled data and labeled data, because unlabeled data may come from a completely different environment than labeled data. This distribution inconsistency may affect the learning result and generalization ability of the model.

Throughout the course of the research, some findings have been disseminated through publication of academic papers, such as at conferences and journals. This not only represents the results of our research on the topic of using deep learning for waste classification, but also highlights the importance and impact of our work. Each publication makes a unique contribution to the field of waste classification, and provides novel insights. These publications form the basis of our overall research objectives and the key findings presented from these publications provide directions for further research into waste classification. The publication list included in this thesis not only serves as evidence of our contribution but also serves as a guide for future inquiry, with both practical and academic value.

8.2 Future Work

In the future, we need to further improve our methods and datasets as follows:

(1) The number of waste images in the dataset needs to continue to increase. Moreover, in the future, we may apply WasteData to train a basic model to quickly achieve high-accuracy classification, then utilize WasteNet to further train and optimize the model. We speculate whether this strategy, which combines the advantages of the two datasets, can improve the model's performance and generalization ability in complex scenarios while ensuring model accuracy.

(2) For non-uniform data augmentation techniques, we will explore strategies to simulate other colors besides natural light. In addition, whether there is a better definition of the initial value of nonuniform data augmentation in different classification tasks also requires further research. Finally, a more effective and

appropriate waste classification adaptive weighted loss function algorithm is also one of our future works.

(3) In CISO model, the impact of the number of iterations on model performance was verified. Our future work will still concentrate on how to adjust the iteration strategy according to specific task requirements and data characteristics to ensure the accuracy and practicality of the model. Besides, the precise selection of training samples also needs to be considered.

To sum up, our future work will focus on further optimizing the model structure and training strategy, exploring new feature extraction and classification methods, and verifying the effectiveness and applicability of the model in a wider range of practical waste application scenarios. Additionally, considering the diversity and variability of waste classification standards, future research will also explore the adaptability and flexibility of the model so that it can quickly respond to updates and changes in classification standards. Through continuous exploration and improvement, we look forward to providing more accurate, efficient, and reliable solutions for waste classification tasks, and making greater contributions to environmental protection.

References

Abdel-Hamid, O., Mohamed, A. R., Jiang, H., Deng, L., Penn, G., & Yu, D. (2014). Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 22*(10), 1533-1545.

Adedeji, O., & Wang, Z. (2019). Intelligent waste classification system using deep learning convolutional neural network. *Procedia Manufacturing, 35*, 607-612.

Al-Sarayreha, M., Reis, M., Yan, W., Klette, R. (2020) Potential of deep learning and snapshot hyperspectral imaging for classification of species in meat. Food Control.

Al-Sarayreha, M. (2020) Hyperspectral Imaging and Deep Learning for Food Safety. PhD Thesis. Auckland University of Technology, New Zealand.

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

Almahairi, A., Ballas, N., Cooijmans, T., Zheng, Y., Larochelle, H., & Courville, A. (2016). Dynamic capacity networks. In *International Conference on Machine Learning* (pp. 2549-2558). PMLR.

Altikat, A. A. A. G. S., Gulbe, A., & Altikat, S. (2022). Intelligent solid waste classification using deep convolutional neural networks. *International Journal of Environmental Science and Technology,* 1-8.

Amasuomo, E., & Baird, J. (2016). The concept of waste and waste management. *J. Mgmt. & Sustainability, 6,* 88.

An, N., & Yan, W. (2021). Multitarget tracking using Siamese neural networks. *ACM Transactions on Multimedia Computing Communications and Applications, 17*(2s), 1-16.

Arazo, E., Ortego, D., Albert, P., O'Connor, N. E., & McGuinness, K. (2020). Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *International Joint Conference on Neural Networks (IJCNN)* (pp. 1-8). IEEE.

Atallah, S. B., Banda, N. R., Banda, A., & Roeck, N. A. (2023). How large language models including generative pre-trained transformer (GPT) 3 and 4 will impact medicine and surgery. *Techniques in Coloproctology, 27*(8), 609-614.

Atito, S., Awais, M., & Kittler, J. (2021). SIT: Self-supervised vision transformer. *arXiv:2104.03602.*

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. *arXiv:1607.06450.*

Bachman, P., Alsharif, O., & Precup, D. (2014). Learning with pseudo-ensembles. In *Advances in Neural Information Processing Systems, 27.*

Bar, A., Wang, X., Kantorov, V., Reed, C. J., Herzig, R., Chechik, G., ... & Globerson, A. (2022). DetReg: Unsupervised pretraining with region priors for object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 14605-14615).

Barrault, L., Chung, Y. A., Meglioli, M. C., Dale, D., Dong, N., Duquenne, P. A., ... & Wang, S. (2023). SeamlessM4T-massively multilingual & multimodal machine translation. *arXiv:2308.11596.*

Bashkirova, D., Abdelfattah, M., Zhu, Z., Akl, J., Alladkani, F., Hu, P., ... & Saenko, K. (2022). ZeroWaste dataset: Towards deformable object segmentation in cluttered scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 21147-21157).

Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., & Raffel, C. A. (2019). MixMatch: A holistic approach to semi-supervised learning. *In Advances in Neural Information Processing Systems, 32.*

Bhunia, A. K., Chowdhury, P. N., Yang, Y., Hospedales, T. M., Xiang, T., & Song, Y. Z. (2021). Vectorization and rasterization: Self-supervised learning for sketch and handwriting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 5672-5681).

Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020). YOLOv4: Optimal speed and accuracy of object detection. *arXiv:2004.10934*.

Cai, Z., Fan, Q., Feris, R. S., & Vasconcelos, N. (2016). A unified multi-scale deep convolutional neural network for fast object detection. In *European Conference on Computer Vision* (pp. 354-370). Springer International Publishing.

Cao, C., Wang, B., Zhang, W., Zeng, X., Yan, X., Feng, Z., ... & Wu, Z. (2019). An improved Faster R-CNN for small object detection. *IEEE Access, 7*, 106838-106846.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. In *European Conference on Computer Vision* (pp. 213-229). Cham: Springer.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In *IEEE/CVF International Conference on Computer Vision* (pp. 9650-9660).

Cascante-Bonilla, P., Tan, F., Qi, Y., & Ordonez, V. (2021). Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning. In *AAAI Conference on Artificial Intelligence* (Vol. 35, No. 8, pp. 6912-6920).

Chen, B., Chen, W., Yang, S., Xuan, Y., Song, J., Xie, D., ... & Zhuang, Y. (2022). Label matching semi-supervised object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 14381-14390).

Chen, G., Wang, H., Chen, K., Li, Z., Song, Z., Liu, Y., ... & Knoll, A. (2020). A survey of the four pillars for small object detection: Multiscale representation,

contextual information, super-resolution, and region proposal. *IEEE Transactions on Systems, Man, and Cybernetics: Systems, 52*(2), 936-953.

Chen, H., Jin, Y., Jin, G., Zhu, C., & Chen, E. (2021). Semisupervised semantic segmentation by improving prediction confidence. *IEEE Transactions on Neural Networks and Learning Systems, 33*(9), 4991-5003.

Chen, J., Guo, H., Yi, K., Li, B., & Elhoseiny, M. (2022). Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 18030-18040).

Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision (ECCV)* (pp. 801-818).

Chen, M., Du, Y., Zhang, Y., Qian, S., & Wang, C. (2022). Semi-supervised learning with multi-head co-training. In *AAAI Conference on Artificial Intelligence* (Vol. 36, No. 6, pp. 6278-6286).

Chen, S., Huang, J., Xiao, T., Gao, J., Bai, J., Luo, W., & Dong, B. (2020). Carbon emissions under different domestic waste treatment modes induced by garbage classification: Case study in pilot communities in Shanghai, China. *Science of the Total Environment,* 717, 137193.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning* (pp. 1597-1607).

Chen, X., Xie, S., & He, K. (2021). An empirical study of training self-supervised vision transformers. In *CVF International Conference on Computer Vision (ICCV)* (pp. 9620-9629).

Chen, X., Yuan, Y., Zeng, G., & Wang, J. (2021). Semi-supervised semantic segmentation with cross pseudo supervision. In *IEEE/CVF Conference on Computer*

Vision and Pattern Recognition (pp. 2613-2622).

Cheng, K., Guo, Q., He, Y., Lu, Y., Gu, S., & Wu, H. (2023). Exploring the potential of GPT-4 in biomedical engineering: the dawn of a new era. *Annals of Biomedical Engineering,* 1-9.

Cheng, Y., & Wang, H. (2019). A modified contrastive loss method for face recognition. *Pattern Recognition Letters, 125,* 785-790.

Clark, K., Luong, M. T., Le, Q. V., & Manning, C. D. (2020). ELECTRA: Pre-training text encoders as discriminators rather than generators. *arXiv:2003.10555.*

Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., & Le, Q. V. (2018). AutoAugment: Learning augmentation policies from data. *arXiv:1805.09501.*

Cui, L., Lv, P., Jiang, X., Gao, Z., Zhou, B., Zhang, L., ... & Xu, M. (2020). Context-aware block net for small object detection. *IEEE Transactions on Cybernetics, 52*(4), 2300-2313.

Cui, W., Zhang, W., Green, J., Zhang, X., & Yao, X. (2019). YOLOv3-DarkNet with adaptive clustering anchor box for garbage detection in intelligent sanitation. In *International Conference on Electronic Information Technology and Computer Engineering (EITCE)* (pp. 220-225).

Dai, W., Li, J., Li, D., Tiong, A. M. H., Zhao, J., Wang, W., … & Hoi, S. (2023). InstructBLIP: Towards general-purpose vision-language models with instruction tuning. *arXiv:2305.06500.*

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'05)* (Vol. 1, pp. 886-893).

De Carolis, B., Ladogana, F., & Macchiarulo, N. (2020). YOLO TrashNet: Garbage

detection in video streams. In *IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS)* (pp. 1-7).

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805.*

DeVries, T., & Taylor, G. W. (2017). Improved regularization of convolutional neural networks with Cutout. *arXiv:1708.04552.*

Ding, N., Tang, Y., Fu, Z., Xu, C., Han, K., & Wang, Y. (2023). Can large pre-trained models help vision models on perception tasks? *arXiv:2306.00693.*

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16×16 words: Transformers for image recognition at scale. *arXiv:2010.11929.*

Driess, D., Xia, F., Sajjadi, M. S., Lynch, C., Chowdhery, A., Ichter, B., ... & Florence, P. (2023). PaLM-E: An embodied multimodal language model. *arXiv:2303.03378.*

Egli, A. (2023). ChatGPT, GPT-4, and other large language models: The next revolution for clinical microbiology? *Clinical Infectious Diseases, 77*(9), 1322-1328.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (VoC) challenge. *International Journal of Computer Vision, 88,* 303-338.

Fan, J., Cui, L., & Fei, S. (2023). Waste detection system based on data augmentation and YOLO_EC. *Sensors, 23*(7), 3646.

Ferronato, N., & Torretta, V. (2019). Waste mismanagement in developing countries: A review of global issues. *International Journal of Environmental Research and Public Health, 16*(6), 1060.

French, G., Laine, S., Aila, T., Mackiewicz, M., & Finlayson, G. (2019). Semi-supervised semantic segmentation needs strong, varied perturbations. *arXiv:1906.01916.*

Fu, J., Sun, X., Wang, Z., & Fu, K. (2020). An anchor-free method based on feature balancing and refinement network for multiscale ship detection in SAR images. *IEEE Transactions on Geoscience and Remote Sensing, 59*(2), 1331-1344.

Fu, K., Li, J., Ma, L., Mu, K., & Tian, Y. (2020). Intrinsic relationship reasoning for small object detection. *arXiv:2009.00833.*

Fu, Y., Nguyen, M., Yan, W. (2022) Grading methods for fruit freshness based on deep learning. Springer Nature Computer Science.

Fu, Y. (2020) Fruit Freshness Grading Using Deep Learning. Master's Thesis. Auckland University of Technology, New Zealand.

Funch, O. I., Marhaug, R., Kohtala, S., & Steinert, M. (2021). Detecting glass and metal in consumer trash bags during waste collection using convolutional neural networks. *Waste Management, 119*, 30-38.

Fran, C. (2017). Deep learning with depth wise separable convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Gao, X., Nguyen, M., & Yan, W. Q. (2023). Enhancement of human face mask detection performance by using ensemble learning models. In *Pacific-Rim Symposium on Image and Video Technology* (pp. 124-137). Singapore: Springer Nature Singapore.

Gedara, K. M., Nguyen, M., & Yan, W. Q. (2023). Enhancing privacy protection in intelligent surveillance: Video blockchain solutions. In *International Congress on Blockchain and Applications* (pp. 42-51). Cham: Springer Nature Switzerland.

Geirhos, R., Jacobsen, J. H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., &

Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence, 2*(11), 665-673.

Girshick, R. (2015). Fast R-CNN. In *IEEE International Conference on Computer Vision* (pp. 1440-1448).

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 580-587).

Gong, C., Wang, D., & Liu, Q. (2021). AlphaMatch: Improving consistency for semi-supervised learning with alpha-divergence. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 13683-13692).

Gonzalez, S., & Miikkulainen, R. (2020). Improved training speed, accuracy, and data utilization through loss function optimization. In *IEEE Congress on Evolutionary Computation (CEC)* (pp. 1-8). IEEE.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM, 63*(11), 139-144.

Grill, J. B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., ... & Valko, M. (2020). Bootstrap your own latent-a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems, 33*, 21271-21284.

Gundupalli, S. P., Hait, S., & Thakur, A. (2017). Multi-material classification of dry recyclables from municipal solid waste based on thermal imaging. *Waste Management, 70*, 13-21.

Guo, C., Fan, B., Zhang, Q., Xiang, S., & Pan, C. (2020). AugFPN: Improving multi-scale feature learning for object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 12595-12604).

Guo, H., Mao, Y., & Zhang, R. (2019). Mixup as locally linear out-of-manifold regularization. In *AAAI Conference on Artificial Intelligence* (Vol. 33, No. 01, pp. 3714-3722).

Guo, M. H., Xu, T. X., Liu, J. J., Liu, Z. N., Jiang, P. T., Mu, T. J., ... & Hu, S. M. (2022). Attention mechanisms in computer vision: A survey. *Computational Visual Media, 8*(3), 331-368.

He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 9729-9738).

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask R-CNN. In *IEEE International Conference on Computer Vision* (pp. 2961-2969).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770-778).

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 37*(9), 1904-1916.

He, R., Yang, J., & Qi, X. (2021). Re-distributing biased pseudo labels for semi-supervised semantic segmentation: A baseline investigation. In *IEEE/CVF International Conference on Computer Vision* (pp. 6930-6940).

Hendrycks, D., & Gimpel, K. (2016). Gaussian error linear units (GELUs). *arXiv:1606.08415*.

Heo, B., Kim, J., Yun, S., Park, H., Kwak, N., & Choi, J. Y. (2019). A comprehensive overhaul of feature distillation. In *IEEE/CVF International Conference on Computer Vision* (pp. 1921-1930).

Heo, B., Yun, S., Han, D., Chun, S., Choe, J., & Oh, S. J. (2021). Rethinking spatial dimensions of vision transformers. In *IEEE/CVF International Conference on Computer Vision* (pp. 11936-11945).

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv:1503.02531.*

Hou, Q., Zhou, D., & Feng, J. (2021). Coordinate attention for efficient mobile network design. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 13713-13722).

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 7132-7141).

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4700-4708).

Huang, G. L., He, J., Xu, Z., & Huang, G. (2020). A combination model based on transfer learning for waste classification. *Concurrency and Computation: Practice and Experience, 32*(19), e5751.

Huang, K., Lei, H., Jiao, Z., & Zhong, Z. (2021). Recycling waste classification using vision transformer on portable device. *Sustainability, 13*(21), 11572.

Hung, W. C., Tsai, Y. H., Liou, Y. T., Lin, Y. Y., & Yang, M. H. (2018). Adversarial learning for semi-supervised semantic segmentation. *arXiv:1802.07934.*

Huo, X., Xie, L., He, J., Yang, Z., Zhou, W., Li, H., & Tian, Q. (2021). ATSO: Asynchronous teacher-student optimization for semi-supervised image segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 1235-1244).

Inoue, H. (2018). Data augmentation by pairing samples for images classification.

arXiv:1801.02929.

Iscen, A., Tolias, G., Avrithis, Y., & Chum, O. (2019). Label propagation for deep semi-supervised learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 5070-5079).

Jaderberg, M., Simonyan, K., & Zisserman, A. (2015). Spatial transformer networks. In *Advances in Neural Information Processing Systems, 28*.

Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. *Electronic Markets, 31*(3), 685-695.

Jeong, J., Lee, S., Kim, J., & Kwak, N. (2019). Consistency-based semi-supervised learning for object detection. In *Advances in Neural Information Processing Systems, 32*.

Ji, H., Liu, Z., Yan, W. Q., & Klette, R. (2019). Early diagnosis of Alzheimer's disease based on selective kernel network with spatial attention. In *Asian Conference on Pattern Recognition* (pp. 503-515). Cham: Springer International Publishing.

Ji, H., Liu, Z., Yan, W. Q., & Klette, R. (2019). Early diagnosis of Alzheimer's disease using deep learning. In *International Conference on Control and Computer Vision* (pp. 87-91).

Jing, L., & Tian, Y. (2020). Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 43*(11), 4037-4058.

Joseph, K. J., Khan, S., Khan, F. S., & Balasubramanian, V. N. (2021). Towards open world object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 5830-5840).

Kang, Z., Yang, J., Li, G., & Zhang, Z. (2020). An automatic garbage classification

system based on deep learning. *IEEE Access, 8,* 140019-140029.

Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4401-4410).

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., ... & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, 102274.

Kaza, S., Yao, L., Bhada-Tata, P., & Van Woerden, F. (2018). *What a Waste 2.0: A Global Snapshot of Solid Waste Management to 2050.* World Bank Publications.

Kenton, J. D. M. W. C., & Toutanova, L. K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *In NAACL-HLT* (Vol. 1, p. 2).

Kim, J., Hur, Y., Park, S., Yang, E., Hwang, S. J., & Shin, J. (2020). Distribution aligning refinery of pseudo-label for imbalanced semi-supervised learning. In *Advances in Neural Information Processing Systems, 33*, 14567-14579.

Kisantal, M., Wojna, Z., Murawski, J., Naruniec, J., & Cho, K. (2019). Augmentation for small object detection. *arXiv:1902.07296.*

Kohli, A. P. S., Sitzmann, V., & Wetzstein, G. (2020). Semantic implicit neural scene representations with semi-supervised training. In *International Conference on 3D Vision (3DV)* (pp. 423-433). IEEE.

Kolesnikov, A., Zhai, X., & Beyer, L. (2019). Revisiting self-supervised visual representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 1920-1929).

Kong, T., Sun, F., Yao, A., Liu, H., Lu, M., & Chen, Y. (2017). RON: Reverse connection with objectness prior networks for object detection. In *IEEE Conference*

on Computer Vision and Pattern Recognition (pp. 5936-5944).

Kraft, M., Piechocki, M., Ptak, B., & Walas, K. (2021). Autonomous, onboard vision-based trash and litter detection in low altitude aerial images collected by an unmanned aerial vehicle. *Remote Sensing, 13*(5), 965.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems, 25*.

Le, H., Nguyen, M., Yan, W. Q., & Nguyen, H. (2021). Augmented reality and machine learning incorporation using YOLOv3 and ARKit. *Applied Sciences, 11*(13), 6006.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436-444.

Lee, D. H. (2013). Pseudo-Label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on Challenges in Representation Learning* (Vol. 3, No. 2, p. 896).

Lee, S., Lee, M., Lee, J., & Shim, H. (2021). Railroad is not a train: Saliency as pseudo-pixel supervision for weakly supervised semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 5495-5505).

Li, B., Zhang, Y., Chen, L., Wang, J., Yang, J., & Liu, Z. (2023). Otter: A multi-modal model with in-context instruction tuning. *arXiv:2305.03726.*

Li, C., Leng, Z., Yan, C., Shen, J., Wang, H., MI, W., ... & Sun, H. (2023). ChatHaruhi: Reviving anime character in reality via large language model. *arXiv:2308.09597.*

Li, J., Li, D., Xiong, C., & Hoi, S. (2022). BLIP: Bootstrapping language-image

pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning* (pp. 12888-12900). PMLR.

Li, J., Wei, Y., Liang, X., Dong, J., Xu, T., Feng, J., & Yan, S. (2016). Attentive contexts for object detection. *IEEE Transactions on Multimedia, 19*(5), 944-954.

Li, K., Liu, C., Zhao, H., Zhang, Y., & Fu, Y. (2021). ECACL: A holistic framework for semi-supervised domain adaptation. In *IEEE/CVF International Conference on Computer Vision* (pp. 8578-8587).

Li, S., Yan, M., & Xu, J. (2020). Garbage object recognition and classification based on Mask Scoring RCNN. In *International Conference on Culture-oriented Science & Technology (ICCST)* (pp. 54-58). IEEE.

Li, W., Liu, K., Zhang, L., & Cheng, F. (2020). Object detection based on an adaptive attention mechanism. *Scientific Reports, 10*(1), 11307.

Li, X., Wang, W., Hu, X., & Yang, J. (2019). Selective kernel networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 510-519).

Li, Y., Xu, C., & Yan, W. (2023). Forest phenomics: how does developing sensor technology improve the growth of forest plantations?. *Frontiers in Forests and Global Change, 6*, 1327850.

Li, Z., & Zhou, F. (2017). FSSD: Feature fusion single shot multibox detector. *arXiv:1712.00960.*

Liang, J. C., Cui, Y., Wang, Q., Geng, T., Wang, W., & Liu, D. (2023). ClusterFormer: Clustering as a universal visual learner. *arXiv:2309.13196.*

Liang, S., & Yan, W. Q. (2022). A hybrid CTC+ Attention model based on end-to-end framework for multilingual speech recognition. *Multimedia Tools and Applications, 81*(28), 41295-41308.

Lieskovská, E., Jakubec, M., Jarina, R., & Chmulík, M. (2021). A review on speech

emotion recognition using deep learning and attention mechanism. *Electronics, 10*(10), 1163.

Lin, C., Guo, M., Li, C., Yuan, X., Wu, W., Yan, J., ... & Ouyang, W. (2019). Online hyper-parameter learning for auto-augmentation strategy. In *IEEE/CVF International Conference on Computer Vision* (pp. 6579-6588).

Liu, J., Pan, C., Yan, W. (2022) Litter detection from digital images using deep learning. Springer Nature Computer Science.

Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2024). Visual instruction tuning. In *Advances in Neural Information Processing Systems, 36.*

Liu, S., Zhi, S., Johns, E., & Davison, A. J. (2021). Bootstrapping semantic segmentation with regional contrast. *arXiv:2104.04465.*

Liu, X., Yan, W. Q., & Kasabov, N. (2024). Moving vehicle tracking and scene understanding: A hybrid approach. *Multimedia Tools and Applications, 83*(17), 51541-51558.

Liu, X. (2023) Vehicle-Related Scene Understanding Using Deep Learning. PhD Thesis, Auckland University of Technology, New Zealand.

Liu, Z., Yan, W., Yang, B. (2018) Image denoising based on a CNN model. International Conference on Control, Automation and Robotics.

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In *European Conference on Computer Vision* (pp. 740-755). Springer International Publishing.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016). SSD: Single shot multibox detector. In *European Conference on Computer Vision* (pp. 21-37). Springer.

Liu, X., Neuyen, M., & Yan, W. Q. (2020). Vehicle-related scene understanding

using deep learning. In *ACPR 2019 Workshops,* (pp. 61-73). Springer Singapore.

Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., ... & Ge, B. (2023). Summary of ChatGPT-related research and perspective towards the future of large language models. *Meta-Radiology,* 100017.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv:1907.11692.*

Liu, Y. C., Ma, C. Y., He, Z., Kuo, C. W., Chen, K., Zhang, P., ... & Vajda, P. (2021). Unbiased teacher for semi-supervised object detection. *arXiv:2102.09480.*

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *IEEE/CVF International Conference on Computer Vision* (pp. 10012-10022).

Liu, Z., Mao, H., Wu, C. Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A ConvNet for the 2020s. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 11976-11986).

Lun, Z., Pan, Y., Wang, S., Abbas, Z., Islam, M. S., & Yin, S. (2023). Skip-YOLO: Domestic garbage detection using deep learning method in complex multi-scenes. *International Journal of Computational Intelligence Systems, 16*(1), 139.

Luo, Z., Nguyen, M., & Yan, W. Q. (2021). Sailboat detection based on automated search attention mechanism and deep learning models. In *International Conference on Image and Vision Computing New Zealand (IVCNZ)* (pp. 1-6).

Luo, Z., Yan, W. Q., & Nguyen, M. (2022). Kayak and sailboat detection based on the improved YOLO with Transformer. In *International Conference on Control and Computer Vision* (pp. 36-41).

Luo, Z. (2022) Sailboat and Kayak Detection Using Deep Learning Methods. Masters Thesis, Auckland University of Technology, New Zealand.

Ma, W., Wang, X., & Yu, J. (2020). A lightweight feature fusion single shot multibox detector for garbage detection. *IEEE Access, 8,* 188577-188586.

Mahajan, D., Girshick, R., Ramanathan, V., He, K., Paluri, M., Li, Y., ... & Van Der Maaten, L. (2018). Exploring the limits of weakly supervised pretraining. In *European Conference on Computer Vision (ECCV)* (pp. 181-196).

Mao, W. L., Chen, W. C., Wang, C. T., & Lin, Y. H. (2021). Recycling waste classification using optimized convolutional neural network. *Resources, Conservation and Recycling, 164*, 105132.

Mehtab, S. Yan, W., Narayanan, A. (2022) 3D vehicle detection using cheap LiDAR and camera sensors. International Conference on Image and Vision Computing New Zealand.

Mehtab, S. (2022) Deep Neural Networks for Road Scene Perception in Autonomous Vehicles Using LiDARs and Vision Sensors. PhD Thesis, Auckland University of Technology, New Zealand.

Meng, S., & Chu, W. T. (2020). A study of garbage classification with convolutional neural networks. In *International Conference on Computing, Analytics and Networks (Indo-Taiwan ICAN)* (pp. 152-157).

Mikolov, T., Kombrink, S., Burget, L., Černocký, J., & Khudanpur, S. (2011). Extensions of recurrent neural network language model. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5528-5531).

Mittal, S., Tatarchenko, M., & Brox, T. (2019). Semi-supervised semantic segmentation with high-and low-level consistency. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 43*(4), 1369-1379.

Miyato, T., Maeda, S. I., Koyama, M., & Ishii, S. (2018). Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 41*(8), 1979-1993.

Mohanraj, C., Senthilkumar, T., Chandrasekar, M., & Arulmozhi, M. (2023). Conversion of waste plastics into sustainable fuel. In *Waste to Profit* (pp. 41-52). CRC Press.

Najibi, M., Rastegari, M., & Davis, L. S. (2016). G-CNN: An iterative grid based object detector. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2369-2377).

Nayan, A. A., Saha, J., Mozumder, A. N., & Mahmud, K. R. (2020). Real time detection of small objects. *Int. J. Innov. Technol. Explor. Eng.*

Neverova, N., Wolf, C., Taylor, G. W., & Nebout, F. (2015). Multi-scale deep learning for gesture detection and localization. In *ECCV* (pp. 474-490). Springer International Publishing.

Nguyen, M., & Yan, W. Q. (2023). From faces to traffic lights: A multi-scale approach for emotional state representation. In *IEEE International Conference on High Performance Computing & Communications, Data Science & Systems, Smart City & Dependability in Sensor, Cloud & Big Data Systems & Application* (pp. 942-949).

Nguyen, M., & Yan, W. Q. (2021). Temporal colour-coded facial-expression recognition using convolutional neural network. In *International Summit Smart City 360°* (pp. 41-54). Cham: Springer International Publishing.

Nie, Z., Duan, W., & Li, X. (2021). Domestic garbage recognition and detection based on Faster R-CNN. In *Journal of Physics: Conference Series* (Vol. 1738, No. 1, p. 012089). IOP Publishing.

Niu, Z., Zhong, G., & Yu, H. (2021). A review on the attention mechanism of deep learning. *Neurocomputing, 452,* 48-62.

Nixon, M., & Aguado, A. (2019). *Feature Extraction and Image Processing for Computer Vision.* Academic Press.

Oliva, A., & Torralba, A. (2007). The role of context in object recognition. *Trends in Cognitive Sciences, 11*(12), 520-527.

Olivier, C. (2006). Semi-Supervised Learning (Adaptive Computation and Machine Learning). *MIT Press, 2006.*

Oord, A. V. D., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv:1807.03748.*

OpenAI. (2023). GPT-4 technical report. *arXiv.2303.08774.*

Ouali, Y., Hudelot, C., & Tami, M. (2020). Semi-supervised semantic segmentation with cross-consistency training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 12674-12684).

Pan, C., Li, X., & Yan, W. Q. (2018). A learning-based positive feedback approach in salient object detection. In *International Conference on Image and Vision Computing New Zealand (IVCNZ)* (pp. 1-6). IEEE.

Pan, C., Liu, J., Yan, W. Q., Cao, F., He, W., & Zhou, Y. (2021). Salient object detection based on visual perceptual saturation and two-stream hybrid networks. *IEEE Transactions on Image Processing, 30*, 4773-4787.

Pan, C., & Yan, W. Q. (2020). Object detection based on saturation of visual perception. *Multimedia Tools and Applications, 79,* 19925-19944.

Papageorgiou, C., & Poggio, T. (2000). A trainable system for object detection. *International Journal of Computer Vision, 38*, 15-33.

Park, W., Kim, D., Lu, Y., & Cho, M. (2019). Relational knowledge distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3967-3976).

Passalis, N., & Tefas, A. (2018). Probabilistic knowledge transfer for deep representation learning. *CoRR, abs/1803.10837, 1*(2), 5.

Pham, H., Dai, Z., Xie, Q., & Le, Q. V. (2021). Meta pseudo labels. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 11557-11568).

Poth, C., Pfeiffer, J., Rücklé, A., & Gurevych, I. (2021). What to pre-train on? efficient intermediate task selection. *arXiv:2104.08247*.

Prince, S. J. (2012). *Computer Vision: Models, Learning, and Inference*. Cambridge University Press.

Proença, P. F., & Simoes, P. (2020). TACO: Trash annotations in context for litter detection. *arXiv:2003.06975*.

Qi, C., & Su, F. (2017). Contrastive-center loss for deep neural networks. In *International Conference on Image Processing (ICIP)* (pp. 2851-2855). IEEE.

Qi, J., Nguyen, M., & Yan, W. Q. (2022). Waste classification from digital images using ConvNeXt. In *Pacific-Rim Symposium on Image and Video Technology* (pp. 1-13). Cham: Springer International Publishing.

Qi, J., Nguyen, M., & Yan, W. Q. (2022). Small visual object detection in smart waste classification using transformers with deep learning. In *International Conference on Image and Vision Computing New Zealand* (pp. 301-314). Cham: Springer Nature Switzerland.

Qi, J., Nguyen, M., & Yan, W. Q. (2024). CISO: Co-iteration semi-supervised learning for visual object detection. *Multimedia Tools and Applications*, *83*(11), 33941-33957.

Qi, J., Nguyen, M., & Yan, W. Q. (2024). NUNI-Waste: novel semi-supervised semantic segmentation waste classification with non-uniform data augmentation. *Multimedia Tools and Applications*, 1-19.

Qiao, S., Shen, W., Zhang, Z., Wang, B., & Yuille, A. (2018). Deep co-training for semi-supervised image recognition. In *European Conference on Computer Vision*

(ECCV) (pp. 135-152).

Qin, Z., Yan, W. (2021) Traffic-sign recognition using deep learning. International Symposium on Geometry and Vision.

Qiu, L., Xiong, Z., Wang, X., Liu, K., Li, Y., Chen, G., ... & Cui, S. (2022). ETHSeg: An amodel instance segmentation network and a real-world dataset for X-Ray waste Inspection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2283-2292).

Rabano, S. L., Cabatuan, M. K., Sybingco, E., Dadios, E. P., & Calilung, E. J. (2018). Common garbage classification using MobileNet. In *IEEE International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM)* (pp. 1-4). IEEE.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning* (pp. 8748-8763).

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. OpenAI Blog.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog,* 1(8), 9.

Rasmus, A., Berglund, M., Honkala, M., Valpola, H., & Raiko, T. (2015). Semi-supervised learning with ladder networks. In *Advances in Neural Information Processing Systems, 28.*

Reddy, Y. C. A. P., Viswanath, P., & Reddy, B. E. (2018). Semi-supervised learning: A brief review. *Int. J. Eng. Technol, 7*(1.8), 81.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once:

Unified, real-time object detection. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 779-788).

Redmon, J., & Farhadi, A. (2017). YOLO9000: Better, faster, stronger. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 7263-7271).

Ren, Y. (2017) Banknote Recognition in Real Time Using ANN. Master's Thesis, Auckland University of Technology, New Zealand.

Ren, Y., Nguyen, M., Yan, W. (2018) Real-time recognition of series seven New Zealand banknotes. International Journal of Digital Crime and Forensics (IJDCF) 10 (3), 50-66.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems, 28.*

Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., & Savarese, S. (2019). Generalized intersection over union: A metric and a loss for bounding box regression. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 658-666).

Romero, A., Ballas, N., Kahou, S. E., Chassang, A., Gatta, C., & Bengio, Y. (2014). FitNets: Hints for thin deep nets. *arXiv:1412.6550.*

Romera-Paredes, B., & Torr, P. H. S. (2016). Recurrent instance segmentation. In *European Conference on Computer Vision* (pp. 312-329). Springer.

Sakalle, A., Tomar, P., Bhardwaj, H., Acharya, D., & Bhardwaj, A. (2021). A LSTM based deep learning network for recognizing emotions using wireless brainwave driven system. *Expert Systems with Applications, 173*, 114516.

Sajjadi, M., Javanmardi, M., & Tasdizen, T. (2016). Mutual exclusivity loss for semi-supervised deep learning. In *IEEE International Conference on Image*

Processing (ICIP) (pp. 1908-1912). IEEE.

Sajjadi, M., Javanmardi, M., & Tasdizen, T. (2016). Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *Advances in Neural Information Processing Systems, 29.*

Sermanet, P., Lynch, C., Chebotar, Y., Hsu, J., Jang, E., Schaal, S., ... & Brain, G. (2018). Time-contrastive networks: Self-supervised learning from video. In *IEEE International Conference on Robotics and Automation (ICRA)* (pp. 1134-1141).

Shen, D., Chen, X., Nguyen, M., & Yan, W. Q. (2018). Flame detection using deep learning. In *International Conference on Control, Automation and Robotics (ICCAR)* (pp. 416-420). IEEE.

Shi, C., Tan, C., Wang, T., & Wang, L. (2021). A waste classification method based on a multilayer hybrid convolution neural network. *Applied Sciences, 11*(18), 8572.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556.*

Singla, A. (2023). Evaluating ChatGPT and GPT-4 for visual programming. In *ACM Conference on International Computing Education Research-Volume 2* (pp. 14-15).

Shrivastava, A., Gupta, A., & Girshick, R. (2016). Training region-based object detectors with online hard example mining. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 761-769).

Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C. A., ... & Li, C. L. (2020). FixMatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems, 33,* 596-608.

Sohn, K., Zhang, Z., Li, C. L., Zhang, H., Lee, C. Y., & Pfister, T. (2020). A simple semi-supervised learning framework for object detection. *arXiv:2005.04757.*

Sousa, J., Rebelo, A., & Cardoso, J. S. (2019). Automation of waste sorting with deep learning. In *IEEE Workshop de Visão Computacional (WVC)* (pp. 43-48).

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research, 15*(1), 1929-1958.

Srinivas, A., Lin, T. Y., Parmar, N., Shlens, J., Abbeel, P., & Vaswani, A. (2021). Bottleneck transformers for visual recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 16519-16529).

Suzuki, T. (2022). TeachAugment: Data augmentation optimization using teacher knowledge. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10904-10914).

Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning* (pp. 6105-6114).

Tang, Y., Chen, W., Luo, Y., & Zhang, Y. (2021). Humble teachers teach better students for semi-supervised object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3132-3141).

Tarvainen, A., & Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems, 30.*

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. *Nature Medicine, 29*(8), 1930-1940.

Tian, Z., Shen, C., Chen, H., & He, T. (2019). FCOS: Fully convolutional one-stage object detection. In *IEEE/CVF International Conference on Computer Vision* (pp. 9627-9636).

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

Vallayil, M., Nand, P., Yan, W. Q., & Allende-Cid, H. (2023). Explainability of automated fact verification systems: A comprehensive review. *Applied Sciences, 13*(23), 12608.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information processing Systems* (pp. 5998- 6008).

Waisberg, E., Ong, J., Masalkhi, M., Kamran, S. A., Zaman, N., Sarker, P., ... & Tavakkoli, A. (2023). GPT-4: A new era of artificial intelligence in medicine. *Irish Journal of Medical Science (1971-),* 1-4.

Wan, L., Zeiler, M., Zhang, S., Le Cun, Y., & Fergus, R. (2013). Regularization of neural networks using dropconnect. In *International Conference on Machine Learning* (pp. 1058-1066). PMLR.

Wang, C. Y., Bochkovskiy, A., & Liao, H. Y. M. (2023). YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 7464-7475).

Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., ... & Tang, X. (2017). Residual attention network for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3156-3164).

Wang, G., Wu, X., Yan, W. (2017) The state-of-the-art technology of currency identification: A comparative study. International Journal of Digital Crime and Forensics 9 (3), 58-72.

Wang, J., Yan, W. (2016) BP-neural network for plate number recognition. International Journal of Digital Crime and Forensics (IJDCF) 8 (3), 34-45.

Wang, J., Bacic, B., Yan, W. (2018) An effective method for plate number recognition. Multimedia Tools and Applications, 77 (2), 1679-1692.

Wang, L., Yan, W. (2021) Tree leaves detection based on deep learning. International Symposium on Geometry and Vision.

Wang, S., Khabsa, M., & Ma, H. (2020). To pretrain or not to pretrain: Examining the benefits of pretraining on resource rich tasks. *arXiv:2006.08671*.

Wang, T., Cai, Y., Liang, L., & Ye, D. (2020). A multi-level approach to waste object segmentation. *Sensors, 20*(14), 3816.

Wang, X., & Gupta, A. (2015). Unsupervised learning of visual representations using videos. In *IEEE International Conference on Computer Vision* (pp. 2794-2802).

Wang, X., & Yan, W. Q. (2023). Human identification based on gait manifold. *Applied Intelligence, 53*(5), 6062-6073.

Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018). CBAM: Convolutional block attention module. In *European Conference on Computer Vision (ECCV)* (pp. 3-19).

Wu, M., Yue, H., Wang, J., Huang, Y., Liu, M., Jiang, Y., ... & Zeng, C. (2020). Object detection based on RGC Mask R - CNN. *IET Image Processing, 14*(8), 1502-1508.

Wu, X., Sahoo, D., & Hoi, S. C. (2020). Recent advances in deep learning for object detection. *Neurocomputing, 396*, 39-64.

Xia, Y., Nguyen, M., Yan, W. (2022) A real-time Kiwifruit detection based on improved YOLOv7. International Conference on Image and Vision Computing New Zealand (IVCNZ)

Xia, Y., Nguyen, M., Yan, W. (2023) Kiwifruit counting using KiwiDetector and KiwiTracker. IntelliSys conference.

Xia, Y., Nguyen, M., Yan, W. (2023) Multiscale Kiwifruit detection from digital images. PSIVT.

Xiang, Y., Yan, W. (2021) Fast-moving coin recognition using deep learning. Springer Multimedia Tools and Applications.

Xiao, B., Nguyen, M., & Yan, W. Q. (2021). Apple ripeness identification using deep learning. In *International Symposium on Geometry and Vision* (pp. 53-67). Springer International Publishing.

Xiao, S., Dong, H., Geng, Y., & Brander, M. (2018). An overview of China's recyclable waste recycling and recommendations for integrated solutions. *Resources, Conservation and Recycling, 134,* 112-120.

Xie, Q., Dai, Z., Hovy, E., Luong, T., & Le, Q. (2020). Unsupervised data augmentation for consistency training. In *Advances in Neural Information Processing Systems, 33*, 6256-6268.

Xie, Q., Luong, M. T., Hovy, E., & Le, Q. V. (2020). Self-training with noisy student improves ImageNet classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10687-10698).

Xin, C., Nguyen, M., & Yan, W. Q. (2020). Multiple flames recognition using deep learning. In *Handbook of Research on Multimedia Cyber Security* (pp. 296-307). IGI Global.

Xing, J., Yan, W. (2021) Traffic sign recognition using guided image filtering. International Symposium on Geometry and Vision.

Xing, J., Nguyen, M., Yan, W. (2022) The improved framework of traffic sign recognition by using guided image filtering. Springer Nature Computer Science.

Xing, J. (2022) Traffic Sign Recognition from Digital Images Using Deep Learning. Master's Thesis, Auckland University of Technology, New Zealand.

Xing, J., Nguyen, M., Yan, W. (2022) Traffic sign recognition from digital images by using deep learning. Pacific-Rim Symposium on Image and Video Technology.

Xu, M., Zhang, Z., Hu, H., Wang, J., Wang, L., Wei, F., ... & Liu, Z. (2021). End-to-end semi-supervised object detection with soft teacher. In *IEEE/CVF International Conference on Computer Vision* (pp. 3060-3069).

Yan, C., Tu, Y., Wang, X., Zhang, Y., Hao, X., Zhang, Y., & Dai, Q. (2019). STAT: Spatial-temporal attention mechanism for video captioning. *IEEE Transactions on Multimedia, 22*(1), 229-241.

Yan, W. Q. (2023). *Computational methods for deep learning: theory, algorithms, and implementations*. Springer Nature.

Yan, W. Q. (2019). *Introduction to Intelligent Surveillance: Surveillance Data Capture, Transmission, and Analytics.* Springer.

Yan, Z., Zhang, H., Wang, B., Paris, S., & Yu, Y. (2016). Automatic photo adjustment using deep neural networks. ACM Transactions on Graphics, 35(2), 1-15.

Yang, B., Xu, K., Wang, H., & Zhang, H. (2022). Random Transformation of image brightness for adversarial attack. *Journal of Intelligent & Fuzzy Systems, 42*(3), 1693-1704.

Yang, F., Wu, K., Zhang, S., Jiang, G., Liu, Y., Zheng, F., ... & Zeng, L. (2022). Class-aware contrastive semi-supervised learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 14421-14430).

Yang, L., Qi, L., Feng, L., Zhang, W., & Shi, Y. (2023). Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 7236-7246).

Yang, L., Zhuo, W., Qi, L., Shi, Y., & Gao, Y. (2022). ST++: Make self-training

work better for semi-supervised semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4268-4277).

Yang, M., & Thung, G. (2016). Classification of trash for recyclability status. *CS229 project report, Stanford University*.

Yang, Q., Wei, X., Wang, B., Hua, X. S., & Zhang, L. (2021). Interactive self-training with mean teachers for semi-supervised object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 5941-5950).

Yang, X., Song, Z., King, I., & Xu, Z. (2022). A survey on deep semi-supervised learning. *IEEE Transactions on Knowledge and Data Engineering.*

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). XLNet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems, 32.*

Yim, J., Joo, D., Bae, J., & Kim, J. (2017). A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4133-4141).

Yin, X., Goudriaan, J. A. N., Lantinga, E. A., Vos, J. A. N., & Spiertz, H. J. (2003). A flexible sigmoid function of determinate growth. *Annals of Botany, 91*(3), 361-371.

Yong, L., Ma, L., Sun, D., & Du, L. (2023). Application of MobileNetV2 to waste classification. *Plos One, 18*(3), e0282336.

Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., & Sang, N. (2018). BiSeNet: Bilateral segmentation network for real-time semantic segmentation. In *European Conference on Computer Vision (ECCV)* (pp. 325-341).

Yu, J., Jiang, Y., Wang, Z., Cao, Z., & Huang, T. (2016). UnitBox: An advanced object detection network. In *ACM international Conference on Multimedia* (pp.

516-520).

Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., & Yoo, Y. (2019). CutMix: Regularization strategy to train strong classifiers with localizable features. In *IEEE/CVF International Conference on Computer Vision* (pp. 6023-6032).

Zagoruyko, S., & Komodakis, N. (2016). Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv:1612.03928.*

Zaman, A. U. (2015). A comprehensive review of the development of zero waste management: Lessons learned and guidelines. *Journal of Cleaner Production, 91,* 12-25.

Zan, D., Chen, B., Zhang, F., Lu, D., Wu, B., Guan, B., ... & Lou, J. G. (2023). Large language models meet NL2Code: A survey. In *Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 7443-7464).

Zhang, Q. (2018) Currency Recognition Using Deep Learning. Master's Thesis, Auckland University of Technology, New Zealand.

Zhang, Q., Yan, W. (2018) Currency detection and recognition based on deep learning. IEEE International Conference on Advanced Video and Signal Based Surveillance.

Zhang, Q., Yan, W., Kankanhalli, M. (2019) Overview of currency recognition using deep learning. Journal of Banking and Financial Technology, 3 (1), 59–69.

Zhang, D. Q., Tan, S. K., & Gersberg, R. M. (2010). Municipal solid waste management in China: status, problems and challenges. *Journal of Environmental Management, 91*(8), 1623-1633.

Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2017). mixup: Beyond

empirical risk minimization. *arXiv:1710.09412.*

Zhang, H., Zu, K., Lu, J., Zou, Y., & Meng, D. (2022). EPSANet: An efficient pyramid squeeze attention block on convolutional neural network. In *Asian Conference on Computer Vision* (pp. 1161-1177).

Zhang, J., Zhao, Y., Saleh, M., & Liu, P. (2020). PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning* (pp. 11328-11339). PMLR.

Zhang, L., Hu, Q., Zhang, S., & Zhang, W. (2020). Understanding Chinese residents' waste classification from a perspective of intention–behavior gap. *Sustainability, 12*(10), 4135.

Zhang, Q., Zhang, X., Mu, X., Wang, Z., Tian, R., Wang, X., & Liu, X. (2021). Recyclable waste image recognition based on deep learning. *Resources, Conservation and Recycling, 171,* 105636.

Zhang, S., Chen, Y., Yang, Z., & Gong, H. (2021). Computer vision based two-stage waste recognition-retrieval algorithm for waste classification. *Resources, Conservation and Recycling, 169*, 105543.

Zhai, X., Oliver, A., Kolesnikov, A., & Beyer, L. (2019). S4L: Self-supervised semi-supervised learning. In *IEEE/CVF International Conference on Computer Vision* (pp. 1476-1485).

Zhao, K. (2021) Fruit Detection Using CenterNet. Master's Thesis, Auckland University of Technology, New Zealand.

Zhao, K., Yan, W. (2021) Fruit detection from digital images using CenterNet. International Symposium on Geometry and Vision.

Zhao, Z. Q., Zheng, P., Xu, S. T., & Wu, X. (2019). Object detection with deep learning: A review. *IEEE Transactions on Neural Networks and Learning Systems,*

30(11), 3212-3232.

Zhao, Z., Yang, L., Long, S., Pi, J., Zhou, L., & Wang, J. (2023). Augmentation matters: A simple-yet-effective approach to semi-supervised semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 11350-11359).

Zheng, H., & Gu, Y. (2021). EnCNN-UPMWS: Waste classification by a CNN ensemble using the UPM weighting strategy. *Electronics, 10*(4), 427.

Zhou, H., Yu, X., Alhaskawi, A., Dong, Y., Wang, Z., Jin, Q., ... & Lu, H. (2022). A deep learning approach for medical waste classification. *Scientific Reports, 12*(1), 2159.

Zhou, Q., Yu, C., Wang, Z., Qian, Q., & Li, H. (2021). Instant-teaching: An end-to-end semi-supervised object detection framework. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4081-4090).

Zhu, D., Chen, J., Shen, X., Li, X., & Elhoseiny, M. (2023). MiniGPT-4: Enhancing vision-language understanding with advanced large language models. *arXiv:2304.10592.*

Zhu, W., Peng, B., & Yan, W. Q. (2024). Dual knowledge distillation on multiview pseudo labels for unsupervised person re-identification. *IEEE Transactions on Multimedia, 26*, 7359-7371.

Zhu, X., Cheng, D., Zhang, Z., Lin, S., & Dai, J. (2019). An empirical study of spatial attention mechanisms in deep networks. In *IEEE/CVF international conference on computer vision* (pp. 6688-6697).

Zhu, X., Lyu, S., Wang, X., & Zhao, Q. (2021). TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios. In *IEEE/CVF International Conference on Computer Vision* (pp. 2778-2788).

Ziouzios, D., & Dasygenis, M. (2019). A smart recycling bin for waste classification. In *Panhellenic Conference on Electronics & Telecommunications (PACET)* (pp. 1-4). IEEE.

Zoph, B., Cubuk, E. D., Ghiasi, G., Lin, T. Y., Shlens, J., & Le, Q. V. (2020). Learning data augmentation strategies for object detection. In *European Conference on Computer Vision* (pp. 566-583). Springer.

Zvyagin, M., Brace, A., Hippe, K., Deng, Y., Zhang, B., Bohorquez, C. O., ... & Ramanathan, A. (2023). GenSLMs: Genome-scale language models reveal SARS-CoV-2 evolutionary dynamics. *International Journal of High Performance Computing Applications, 37*(6), 683-705.