Strawberry Ripeness Detection Using Deep Learning

Zhiyuan Mi

A project report submitted to the Auckland University of Technology
in partial fulfillment of the requirements for the degree of
Master of Computer and Information Sciences (MCIS)

2024

School of Engineering, Computer & Mathematical Sciences

Abstract

In agriculture, timely and accurate assessment of fruit ripeness is crucial to optimize harvest planning and reduce waste. In this report, we explore the integration of two cutting-edge deep learning models, YOLOv9 and Swin Transformer, to develop a complex model for detecting strawberry ripeness. Trained and tested on a specially curated dataset, our model achieves a mean precision (mAP) 87.3% by using the metric intersection over union (IoU) at threshold 0.5. This performance outperforms over the model by using YOLOv9 alone, which achieved an mAP 86.1%. Our model also demonstrated the improved Precision and Recall, with Precision rising to 85.3% and Recall rising to 84.0%, reflecting its ability to accurately and consistently detect different stages of strawberry ripeness.

Keywords: Transformer, YOLOv9, Swin Transformer, Deep Learning, Computer Vision

Table of Contents

List of Figures

List of Tables

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: Date: 14 April 2024

Acknowledgment

Fisrt of all, I would like to thank my parents for their finanical support. Owing to the unselfish and generous sponsor from them, I have this invaluable opportunity to complete my Master's study with the Auckland University of Technology (AUT), New Zealand.

I would also like to express my deepest gratitude to my primary supervisor Wei Qi Yan. In this study, he not only provided me with professional knowledge support and careful guidance, but also helped me enrich my learning experience. I believe I could not complete my study without Dr Yan's supervision and instructions.

Finally, my cordial thanks also go to my girlfriend and my best friends who love and care me and whom I love and care.

Zhiyuan Mi

Auckland, New Zealand

April 2024

Chapter 1
Introduction

This chapter is composed of five parts: The first part introduces the background and motivations, the second part includes the research question, followed by the contributions, objectives, and structure of this report.

1.1 Background and Motivation

The agricultural sector is at a critical juncture, facing global challenges such as population growth, climate change, and the need for sustainable development practices, which urgently require innovative solutions. Strawberries, as a high-value crop, symbolize these challenges. They are highly perishable, sensitive to environmental conditions, and require precise harvest timing to ensure optimal quality and yield. Traditionally, assessing strawberry ripeness has been a manual process relying on labor-intensive testing, which is time-consuming and error-prone. This approach faces significant challenges in terms of scalability, efficiency, and objectivity, especially in large-scale commercial farming.

The emergence of precision agriculture empowered by artificial intelligence and machine learning offers a promising solution to these challenges. AI technologies, including deep learning, have shown great potential in transforming various industries by automating complex tasks efficiently and accurately. In agriculture, these technologies have started paving the way for applications such as automated plant disease detection, yield prediction and crop monitoring.

YOLOv9 and Swin Transformer are two popular models in the field of artificial intelligence (AI) and computer vision.YOLOv9, the latest iteration of the models “You Only Look Once" series, is known for its real-time object detection capabilities, which have been significantly improved over its predecessors in terms of speed and accuracy. On the other hand, Swin Transformer introduces a layered transformer with an architecture suitable for efficient processing of image data, especially for tasks requiring detailed visual understanding.

Therefore, integrating these technologies for strawberry ripeness detection is a leap forward for precision agriculture. By automating ripeness detection, this approach aims to reduce reliance on manual labor, minimize human error, and make more accurate and timely harvest decisions. This is especially important for strawberries, which have a narrow window of optimal ripeness and whose quality directly impacts market value and

consumer satisfaction.

1.2 Research Questions

The main objective of this report is how to effectively integrate and apply YOLOv9 and Swin Transformer technologies for strawberry ripeness detection, so the main research questions of this report are:

How can we effectively combine YOLOv9 and Swin Transformer for strawberry ripeness detection?

In order to solve this research question, we split it into the following questions:

(1) How can the integrated YOLOv9 and Swin Transformer model be trained, validated ?

(2) What role does the Swin Transformer play in enhancing the accuracy of ripeness detection, and how does it complement the object detection capabilities of YOLOv9?

(3) How do we evaluate the model and prove that our improvements to the model are effective?

The core concept of this project is to utilize the complementary advantages of YOLOv9 and Swin Transformer to develop a robust, efficient, and highly accurate strawberry ripeness detection system. We need train the dataset to get the best results and evaluate the model thoroughly.

1.3 Contributions

The focus of this report is to train a strawberry ripeness detection model. The main contributions of this report will be:

(1) **Dataset creation and enhancement**: We will create, annotate, and preprocess a dataset of strawberry images and videos. The dataset will contain images under various environmental conditions to enhance the robustness and applicability of

detection models in different agricultural environments.

(2) **Integration of YOLOv9 and Swin Transformer:** This report will integrate YOLOv9 with Swin Transformer to create a powerful strawberry ripeness detection model. This model will take advantage of the efficient object detection capabilities of YOLOv9 and the powerful feature extraction capabilities of Swin Transformer to significantly improve the accuracy of ripeness detection.

(3) Model evaluation: We will evaluate the model using comprehensive evaluation methods, including precision recall analysis, intersection of unions (IoU), and mean average precision (mAP). We compare it with the YOLOv9 model to prove the effectiveness of our model improvements.

1.4 Objectives of This Report

This report aims to demonstrate the application of advanced deep learning technology in the agricultural field, focusing on strawberry ripeness detection. Firstly, a comprehensive review of related technologies and their applications in agriculture is provided, with a focus on deep learning. Secondly, a model for detecting strawberry ripeness based on YOLOv9 and Swin Transformer is trained. Thirdly, rigorous methods are employed to evaluate the performance of the proposed model against the existing benchmarks. At the end of the report, we will discuss limitations and future work.

1.5 Structure of This Report

This report is organized into six chapters, each focuses on a specific aspect of the research and development process:

In Chapter 2, we introduce the relevant technologies in this report based on the literature. The research outcomes related to the application of deep learning in agriculture and current fruit ripeness detection methods are also discussed.

In Chapter 3, we describe the technical approach and experimental design, including

the integration of YOLOv9 and Swin Transformer as well as the evaluation metrics to evaluate model performance.

In Chapter 4, we present the results of the experiments, including the performance comparison of the proposed model with the standard model.

In Chapter 5, we will analyze and discuss the experimental results. We will dosome extra research to analyze the nutritional value of strawberries.

Finally, in Chapter 6 we conclude this report and discuss our limitations and future work.

:

Chapter 2
Literature Review

In this chapter, we will introduce the technology appeared in this report. The application of deep learning in agriculture and existing fruit ripeness detection technology will also be discussed.

2.1 Introduction

The field of computer vision has made significant progress in deep learning techniques, especially in object detection. The application of these technologies in agriculture is also becoming widespread.

2.2 You Only Look Once (YOLO)

YOLO (You Only Look Once) is a popular real-time object detection system first proposed in 2016 by Joseph Redmon et al. The core idea of YOLO models lies in transforming the object detection task as a single regression problem directly from image pixels to bounding-box coordinates and category probabilities, an approach that, compared to previous systems, dramatically increases the speed and efficiency, allowing it to run in real time(Redmon et al., 2016).

2.2.1 Version Iterations of YOLO

YOLOv1 is the first version of the YOLO family, which takes use of GoogLeNet as its backbone. It was characterized by its speed and ability to perform real-time detection. However, its accuracy was slightly lower compared to other detection systems available at the time(Redmon et al., 2016).

YOLOv2 is the second version of the YOLO algorithm, which was proposed in 2016. The YOLOv2 model improves on YOLOv1 by increasing the detection precision and recall while maintaining high speed detection. YOLOv2 is use of Darknet-19 as the backbone network, which is faster and more accurate than GoogLeNet. It also introduces anchor boxes to improve detection accuracy(Redmon & Farhadi, 2016).

YOLOv3 makes use of Darknet-53 as its backbone, which is more accurate than Darknet-19. It also takes advantage of multi-label classification to detect multiple target categories. These improvements have led to further improvements in detection accuracy,

especially on small objects(Redmon & Farhadi, 2018).

The YOLOv4 algorithm is improved based on YOLOv3. It adopts CSPDarknet as the backbone network, which is faster and more accurate than Darknet-53. YOLOv4 also takes in the Bag of Freebies (BoF) strategy, which further improves the detection accuracy and speed(Bochkovskiy et al., 2020).

YOLOv5 has the Focus module to dramatically improve network efficiency. At the same time YOLOv5 provides models of multiple sizes, which greatly improves flexibility and ease of use, and is widely used in industrial and research projects(Cao et al., 2023).

The YOLOv6 algorithm is improved on YOLOv5, mainly includes: Using EfficientRep as the backbone network to improve the model inference speed; Using Rep-PAN as the neck network to enhance feature fusion(Li et al., 2022).

YOLOv7 is the sequel of YOLOv4 team, which mainly focuses on the optimization of model structure re-referencing and dynamic label assignment issues(Wang, Bochkovskiy, et al., 2022).

YOLOv8, from the same team as YOLOv5, is a cutting-edge, state-of-the-art (SOTA) model that builds on the success of the previous YOLOv5 version and introduces new features and enhancements to further improve performance and flexibility.YOLOv8 was designed to be fast, accurate, and easy to use, making it an excellent choice for a wide variety of object detection and tracking, instance segmentation, image categorization, and pose estimation tasks. YOLOv8(Hussain, 2023).

2.2.2 YOLOv9 Model

YOLOv9 (Wang et al. , 2024) has taken significant advances in the field of deep learning target detection. The proposed concept of programmable gradient information (PGI) was employed to cope with the various variations required for deep networks to achieve

multiple goals. Meanwhile, a new lightweight network structure was designed based on gradient path planning, the generalized efficient layer aggregation network (GELAN)(Wang et al., 2024) .

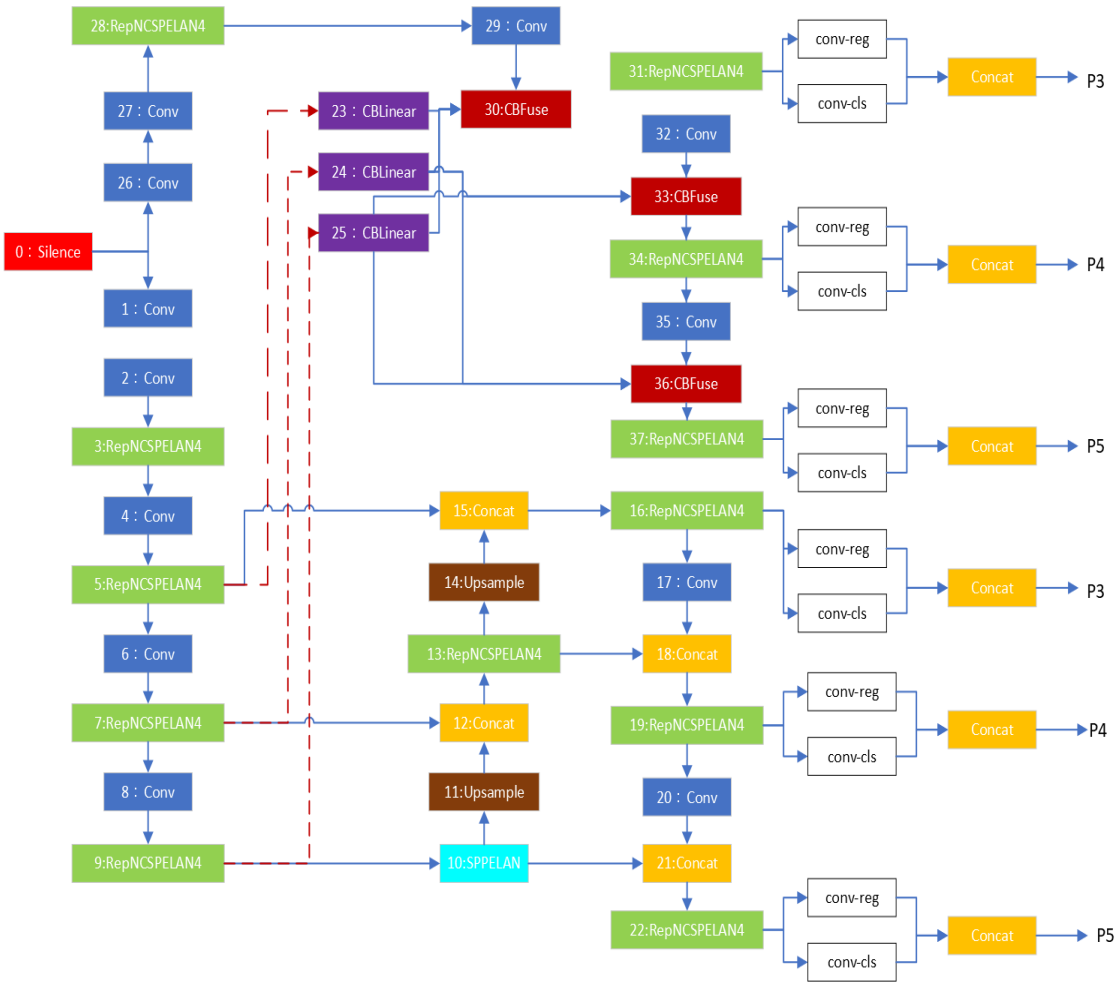

Figure 2.1. YOLOv9 structure

Figure 2.1 is a structural diagram based on YOLOv9.yaml. On the left side of the figure are consecutively numbered convolutional layers (Conv) that are typically used for feature extraction. The backbone is composed of Silence, Conv and RepNCSPELAN4.

YOLOv9 proposed a new network architecture, GELAN.GELAN designs a generalized and efficient layer aggregation network by combining two neural network architectures, CSPNet(Wang et al., 2019) and ELAN(Wang, Liao, et al., 2022). GELAN combines lightweight, inference speed, and accuracy, and its design allows for the flexible

integration of a variety of computational modules. This allows YOLOv9 to be adapted to a wide range of applications without sacrificing speed or accuracy

Table 2.1 The performance of various versions of YOLO on the MS COCO dataset

Model	mAP (50-95)	mAP (50)	FLOPs
YOLOv9t	38.3	53.1	7.7
YOLOv9s	46.8	63.4	26.7
YOLOv9m	51.4	68.1	76.8
YOLOv9c	53.0	70.2	102.8
YOLOv9e	55.6	72.8	192.5
YOLOv7	53.9	72.2	181.7
YOLOv7-X	55.0	73.2	307.9
YOLOv6-N	37.5	53.1	11.4
YOLOv6-S	45.0	61.8	45.3

Table 2.1 shows the performance of each version of YOLO on the MS COCO dataset. The lightweight model YOLOv9s improves on mAP, the significant progress has been made in balancing the trade-off between model complexity and detection performance in the medium- to large-scale models YOLOv9m and YOLOv9e, which improves the accuracy while significantly reducing parameters and computation. The results demonstrate strategic advances in model design for YOLOv9, emphasizing that it improves efficiency without degrading the accuracy necessary for real-time object detection tasks. The model not only pushes the boundaries of performance metrics, but also emphasizes the importance of computational efficiency, making it a key development in computer vision.

2.3 Transformer

The Transformer model has become a revolutionary technology in the field of Natural Language Processing (NLP) since it was firstly introduced (Vaswani et al., 2017). The core idea is to utilize the Self-Attention mechanism to capture long-distance dependencies in the input sequence, which is crucial for understanding and generating natural language(Woo et al., 2018).The success of the Transformer model has spawned a series of models based on this architecture. Figure 2.3 shows the transformer structure.

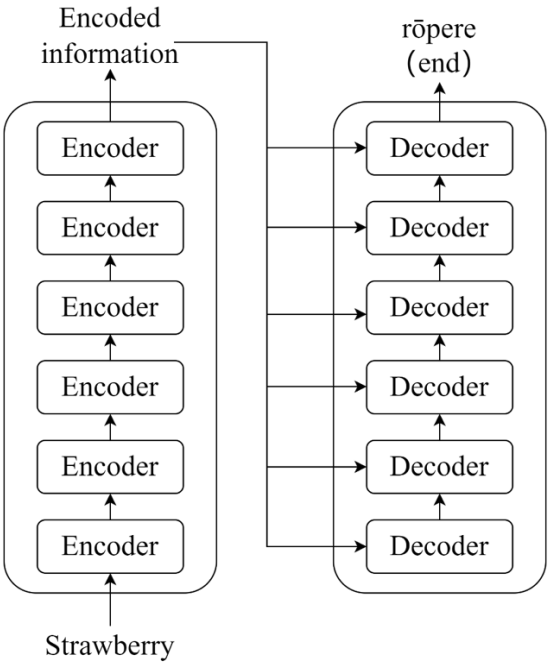

Figure 2.3. Transformer structure

BERT (Bidirectional Encoder Representations from Transformers) is an approach for pre-training language representations, proposed by the Google AI team in 2018.The innovation of the BERT model is its ability to capture context based on both sides of the entire input data (i.e., bi-directionally). It has significant implications for improving the performance of natural language processing (NLP) tasks. This model marks an important advancement in the field of deep learning and NLP by setting a new performance standard in multilingual comprehension tasks(Devlin et al., 2019).

XLNet is an advanced natural language processing model proposed at Google Brain and Carnegie Mellon University in 2019. It is an improvement on the BERT model and aims to address some of the limitations inherent in the Masked Language Model (MLM) by BERT in the pre-training process. XLNet proposes a new pre training method: Permutation Language Model (PMLM)(Z. Yang et al., 2020).

RoBERTa (Robustly optimized BERT approach) is a language representation model based on the BERT architecture, proposed by Facebook AI in 2019. RoBERTa has achieved significant performance gains in several natural language processing (NLP)

tasks, keyed by several Key Improvements(Y. Liu et al., 2019).

GPT (Generative Pre-trained Transformer) is the first one in a series of Natural Language Processing (NLP) models developed by OpenAI, designed to understand, and generate natural language text through deep learning techniques. The first generation of GPT models which introduced the Transformer architecture to the NLP domain, demonstrating the potential for strong performance on a wide range of NLP tasks through pre-training and fine-tuning (Radford et al., 2018). Based on GPT, GPT-2 significantly improves the quality and diversity of text generation by increasing the model size and training dataset. The release of GPT-2 has led to a wide-ranging discussion on the potential impact of AI text generation(Radford et al., 2019). GPT-3 further scales the model with 175 billion parameters.GPT-3 demonstrates amazing text comprehension and generation capabilities that can be adapted to multiple language tasks with little or no fine-tuning(Brown et al., 2020). GPT-4 is the latest generation of natural language processing models introduced by OpenAI, inheriting, and significantly improving the capabilities of its predecessor, GPT-3. As a large-scale multimodal model, GPT-4 not only makes significant progress in text generation and comprehension, but also adds the ability to process image input, further extending the model's range of applications(OpenAI et al., 2024).

2.4 Transformer in Computer Vision

Transformer architecture was originally designed to handle Natural Language Processing (NLP) tasks but has been shown to be well suited for processing image data as well. The introduction of Transformer has brought innovative approaches and models to the field of computer vision (Bi et al., 2021).

Vision Transformer (ViT) segments the input image into multiple fixed-size patches, which are tiled and converted into a series of vectors (analogous to word embeddings in NLP) via a linear projection layer, and these vectors are then fed into the standard Transformer model for processing. By pre-training on large-scale image datasets, ViT

demonstrates impressive performance, especially on image classification tasks(Dosovitskiy et al., 2021).

DETR utilizes Transformer's self-attention mechanism to handle the task of target detection, and by directly predicting object bounding boxes and categories throughout the image, it simplifies the process of target detection by removing many of the complex steps that need to be manually designed (Carion et al., 2020).

2.5 Deep Learning in Agriculture

Deep learning has emerged as a powerful tool in the field of agriculture, offering innovative solutions to various challenges faced in the industry. A slew of studies have highlighted the potential of deep learning in agriculture(Kamilaris & Prenafeta-Boldu, 2018).

The effectiveness of using pretrained deep neural networks (DNN) on agricultural datasets was explored(Espejo-Garcia et al. 2020) to improve weed identification accuracy in precision agriculture.

Convolutional Neural Networks (CNN) are particularly effective in analyzing plant disease images (Sharma et al., 2023)). The Faster R-CNN model they designed achieved a detection rate of 99.39% in pepper plants, highlighting the potential of deep learning models to revolutionize agricultural disease management.

The application of machine learning and deep learning has been explored in crop biomass feedstock production (Peng & Karimi Sadaghiani, 2023), including optimization of photosynthesis, crop improvement, and overall sustainability of crop production. The important role these technologies play in improving the sustainability and efficiency of agricultural practices is highlighted.

The applications of deep learning in agriculture have been exposed (Ren et al., 2020) (Bharman et al., 2022). They all highlighted the potential of deep learning to enhance

precision agriculture by enabling real-time decision-making and optimizing agricultural operations. The challenges have been discussed such as the need for large data sets and the high computational costs associated with training complex models. Despite these challenges, integrating deep learning into agriculture is expected to drive innovation that will significantly increase the productivity and sustainability of the industry.

2.6 Fruit Ripeness Detection

Fruit ripeness detection is an important aspect in agriculture as it affects the quality and shelf life of the fruit. Various technologies and sensors have been developed to detect fruit ripeness at different stages of ripeness.

A photon ripeness detection system (Hasanuddin et al., 2015)) was introduced that relies on light reflectance to differentiate between ripe and unripe fruits. The use of non-invasive a thermal microwave spectroscopy (Korostynska et al., 2018) has been employed for real-time fruit ripeness detection, particularly focusing on automated strawberry picking. Astuti et al. (2019) addressed the lack of ripeness knowledge among farmers by developing a tool for oil palm fruit ripeness detection using the K-Nearest Neighbor algorithm. Jiao et al. (2021) introduced a miniature spectral sensor for fruit ripeness detection, emphasizing the importance of accurately assessing fruit maturity during picking and transportation processes. Several studies have focused on the detection of fruit ripeness using deep learning algorithms.

A method (Mu et al., 2020) was proposed to automatically detect unripe tomatoes by using Faster Region-based Convolutional Neural Network (Faster R-CNN) and ResNet-101 model to learn from the COCO dataset through transfer learning. The method performed well on immature and occluded tomatoes that are difficult to detect through traditional image analysis methods.

A CNN was introduced for automated, lossless classification of mulberry maturity (Ashtiani et al., 2021). The method improved the accuracy and efficiency of the sorting

process by automatically classifying fruit into different ripeness categories based on visual cues.

The application of transfer learning using the VGG-16 model in fruit ripeness detection (Pardede et al., 2021). Their study highlights the effectiveness of deep learning relative to traditional machine learning for this task, with particular emphasis on the important role regularization techniques play in enhancing model performance.

A powerful CNN model was proposed to detect citrus black spot disease and evaluate fruit ripeness through deep learning (Momeny et al., 2022). One of their key innovations is the use of a learning augmentation strategy that generates new data from noisy and recovered images to enhance model training. Momeny et al. utilized Bayesian algorithm-optimized noise parameters to create noisy images and then took use of convolutional autoencoders to restore these images, effectively augmenting the training data.

A systematic review of methods was conducted for oil palm fresh fruit bunch ripeness detection (Lai et al., 2023). Significantly, they identify computer vision combined with deep learning as the most promising approach for field applications due to its real-time operation, cost-effectiveness, and non-contact nature. This method outshines others in terms of adaptability and accuracy, offering substantial improvements over traditional visual assessments by workers, which are subjective and labor-intensive.

In summary, it is very feasible to use deep learning for fruit ripeness detection. The potential of deep learning in fruit ripeness detection represents an important step forward in agricultural technology, with the potential not only to reduce labor costs but also to increase efficiency and reduce waste.

Chapter 3
Methodology

The main content of this chapter is to clearly articulate research methods which satisfy the objectives of this report. This chapter mainly introduces YOLOv9 and Swin Transformer. We will also introduce our evaluation methods.

3.1 Introduction

In this chapter, we outline the research methodology for the development and evaluation of a deep learning-based system for dynamic detection of strawberry ripeness through video analysis. The integration of YOLOv9 and Swin Transformer technologies forms the core of our approach, leveraging their capabilities to achieve real-time, accurate ripeness detection.

3.2 Research Design

3.2.1 Overall of the Proposed Model

In this report, we propose a strawberry ripeness detection method based on YOLOv9 network and Swin Transformer. The method can automatically detect the position of the strawberries from a video with multiple frames and track their movement trajectory to mark the strawberries at the site and predict their ripeness. This method will be a great convenience for growing and picking strawberries.

We trained a hybrid model by combining YOLOv9 and Swin Transformer, which enhances the model's ability to generalize and rely on modeling capabilities at a distance, resulting in better overall performance(S. Yang et al., 2023).

The overall structure of the strawberry ripeness detection model is shown in Figure 3.1. Firstly, YOLOv9 model is trained by using the pre-prepared dataset. The model is improved by combining Swin Transformer, which can better extract the target feature information(Liu et al., 2022). Then, the video was processed using a fusion network of YOLOv9 and Swin Transformer to detect strawberry ripeness with high accuracy. The model will classify the strawberries as the classes “Unripe”, “Half-ripe”, and “Ripe”, and outputs detection frames and feature vectors for each frame of the given video.

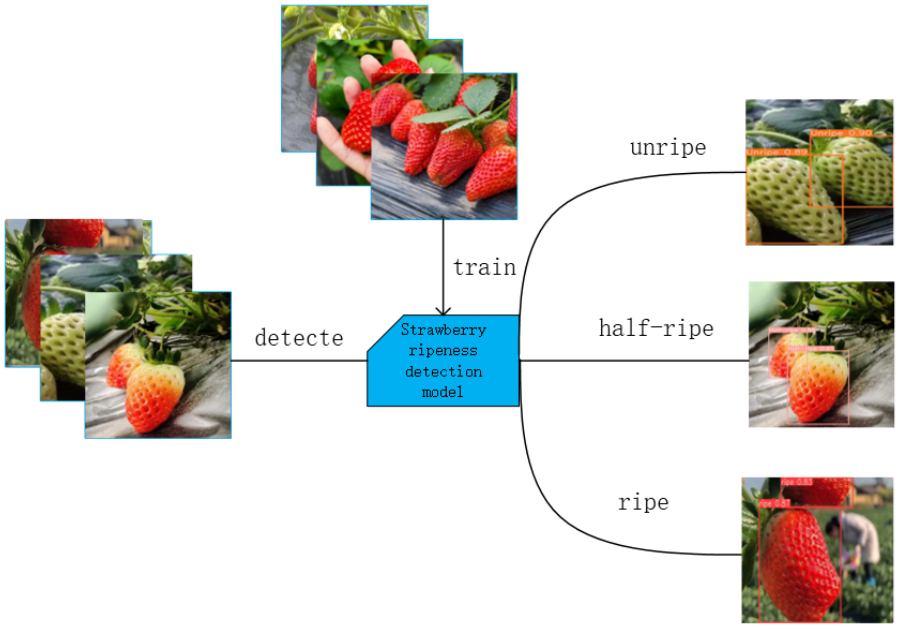

Figure 3.1 Overall structure of the strawberry ripeness detection model

3.2.2 Research Design of YOLOv9 Model

YOLOv9 is the latest version of the YOLO algorithm family. YOLOv9 is improved on its predecessor with the aims to address the problem of information loss in deep learning and to improve the performance of the model on a variety of tasks. YOLOv9 introduces Programmable Gradient Information (PGI), which preserves important data throughout the depth of the proposed network, ensuring more reliable gradient generation and thus improving model convergence and performance. Meanwhile, YOLOv9 designs a new lightweight network structure based on gradient path planning: generalized efficient layer aggregation network (GELAN). By using only conventional convolution, GELAN achieves higher parameter utilization than deeply differentiable convolutional designs based on state-of-the-art techniques, while demonstrating the great advantages of being lightweight, fast, and accurate.

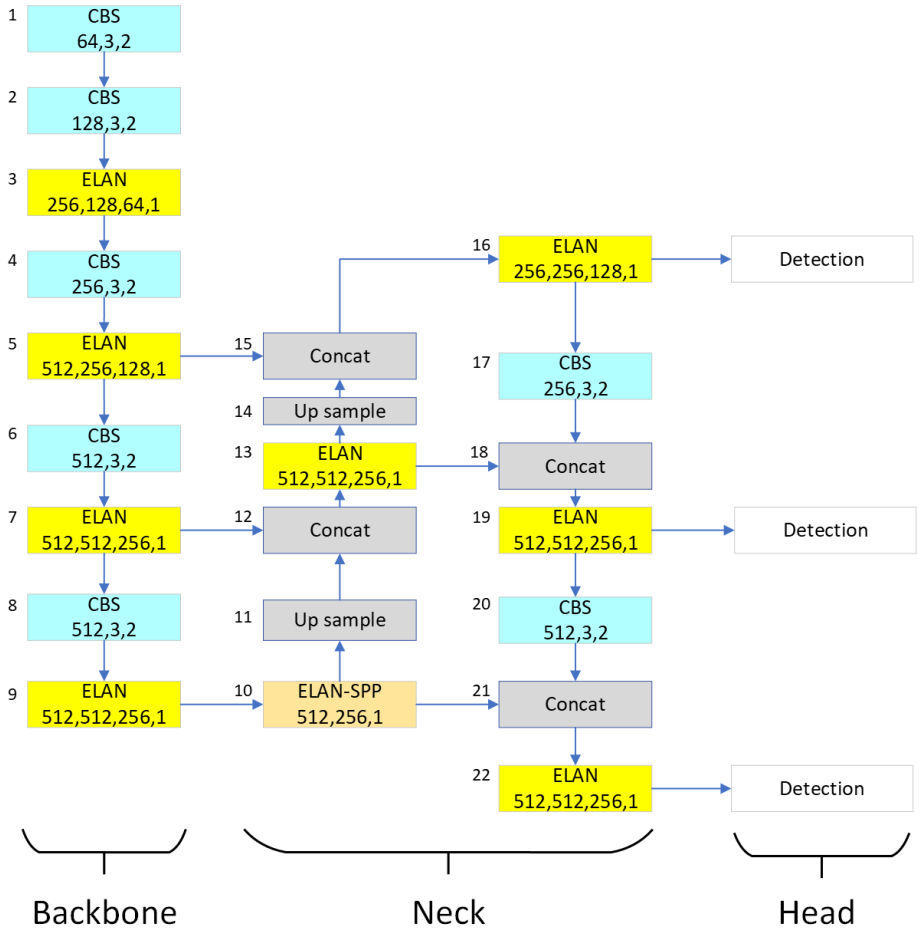

Figure 3.2 YOLOv9 structure

Figure 3.2 illustrates the convolutional neural network architecture of YOLOv9 model. This model is divided into three main parts: Backbone, Neck, and Head. The Backbone is the main feature extraction part of the model. It consists of multiple convolutional layers that are responsible for extracting useful features from the input image. Backbone consists of multiple layers that progressively reduce the spatial dimensions and increase the number of channels through different depth and step configurations, which helps in capturing features at different levels of abstraction of the image.

Neck is the part that connects Backbone and Head, which serves to perform feature fusion and realignment for object recognition by the detector. This part consists of Up sample and Concat operations, which combine high level, smaller feature maps with low level, larger feature maps, thus preserving spatial information at different scales. This

helps to detect objects at different scales of the image. Head is the last part of the model and is responsible for object detection based on the features coming from Backbone and Neck.

In deep learning networks, information loss may occur as data passes through each layer. This loss of information may lead to bias in the gradient flow, which in turn affects the learning of the model. The PGI (Programmable Gradient Information) proposed by YOLOv9 aims to solve this problem. The core idea of PGI is to incorporate an auxiliary reversible branch in the model that generates reliable gradients for deeper features to use, even though these deeper features may have lost critical information due to information bottleneck effects. Auxiliary reversible branching ensures that critical features are maintained even in deep networks to perform the target task. One of the main advantages of PGI over traditional deep supervision methods is its wider applicability, which makes it suitable not only for very deep neural networks, but also for lightweight models. GI allows the model to generate reliable gradient updates to the network parameters and is free to choose a loss function that is suitable for the target task, overcoming the problems encountered in masked modeling. With the PGI mechanism, even lightweight models can benefit from an auxiliary supervision mechanism. In YOLOv9, PGI enables models to achieve or even surpass the results of existing models pre-trained with large datasets even when trained from scratch. YOLOv9 is able to obtain complete information on models of different sizes, enabling highly accurate target detection.

3.2.3 Research Design of Swin Transformer

Swin Transformer (Shifted Window Transformer) is a computer vision model based on Transformer. Swin Transformer overcomes the problems of computational inefficiency and difficulty in handling high-resolution images of traditional Transformer models(Z. Liu et al., 2021).

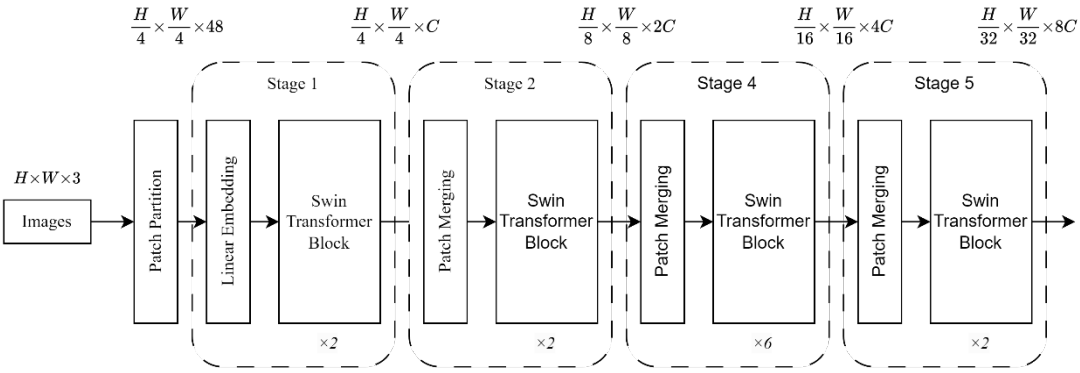

Figure 3.3 Swin Transformer structure

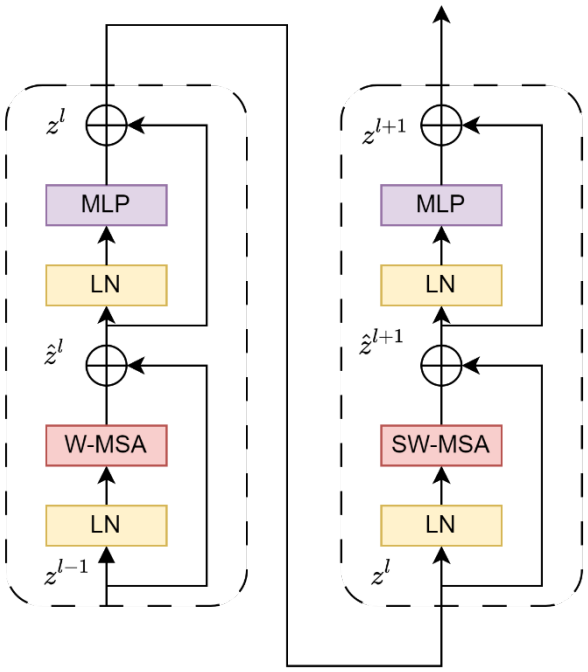

Figure 3.4. Swin Transformer Blocks

Figure 3.3 shows the structure of the Swin Transformer. At the beginning, the image is divided into multiple small blocks, each small block is usually a small square. These patches are flattened into vectors and passed into the model for processing. The model adopts a layered design and consists of four stages, each stage will reduce the resolution of the input feature map. The four stages construct feature maps of different sizes. Except for the first stage, which first passes through a Linear Embedding layer, the remaining three stages first pass through a Patch Merging layer for down sampling. The four stages construct feature maps of different sizes. Except for the first stage, which first passes

through a Linear Embedding layer, the remaining three stages first pass through a Patch Merging layer for down sampling. Each stage has repeatedly stacked Swin Transformer Blocks.

Transformer Block has two structures, as shown in Figure 3.4. The difference between the two structures is that one uses the W-MSA structure and the other uses the SW-MSA structure. Moreover, these two structures are used in pairs, first using a W-MSA structure and then using a SW-MSA structure.

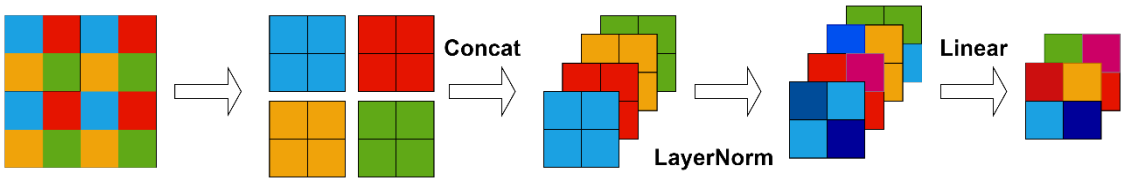

Figure 3.5. Patch merging

Patch Merging: As shown in Figure 3.5, assume that the input to Patch Merging is a 4×4 size single-channel feature map. Patch Merging will divide each 2×2 adjacent pixel into a patch, and then divide the same position in each patch (The same color) pixels are put together to get 4 feature maps. Then these four feature maps are concat spliced in the depth direction, and then passed through a LayerNorm layer. Finally, a fully connected layer is used to make a linear change in the depth direction of the feature map, changing the depth of the feature map from C to C/2. In other words, after passing the Patch Merging layer, the height and width of the feature map will be halved, and the depth will be doubled.

W-MSA: Another important improvement of Swin Transformer is the window-based self-attention layer, which is W-MSA (Windows Multi-head Self-Attention). As shown in Figure 3.6, the left side is the ordinary Multi-head Self-Attention (MSA) module. For each pixel in the feature map, it needs to be calculated with all pixels during the Self-Attention calculation process. However, when using the Windows Multi-head Self-Attention (W-MSA) module, first divide the feature map into windows according to the size of $M \times M$ (e.g., $M=2$ in Figure 3.6), and then perform self-attention inside each

windows individually.

Eq. (3.1) and Eq. (3.2) are the calculation formulas of MSA and W-MSA, h is the height of the feature map, w is the width of the feature map, C is the depth of the feature map, and M is the size of each window. Since the number of patches inside the window is much smaller than the number of image patches, and the number of windows remains unchanged, the computational complexity of W-MSA is linearly related to the image size, thus greatly reducing the computational complexity of the model.

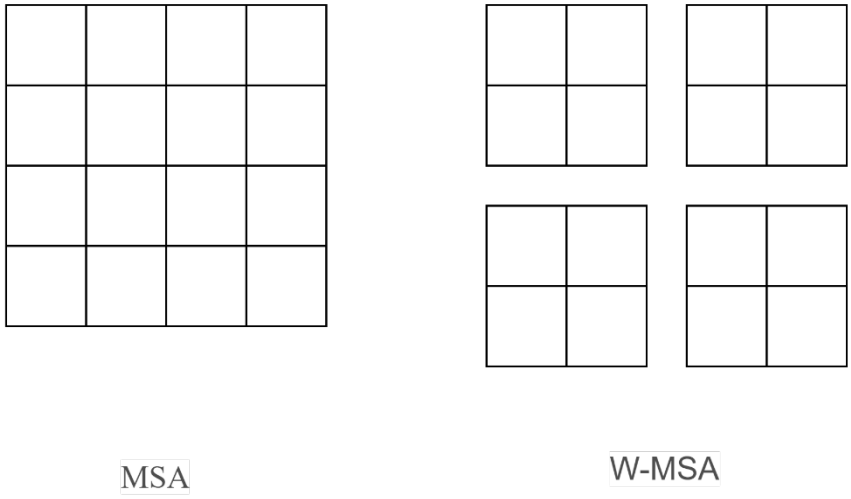

Figure 3.6 MSA and W-MSA

$$\Omega(MSA) = 4hwC^2 + 2(hw)^2C \tag{3.1}$$

$$\Omega(W-MSA) = 4hwC^2 + 2M^2hwC \tag{3.2}$$

SW-MSA: While using the W-MSA module, self-attention calculation will only be performed within each window, so information cannot be transferred between windows. To solve this problem, Swin Transformer introduces Shifted Windows Multi-Head Self-Attention (SW-MSA).

In this report, we built the strawberry ripeness detection model based on YOLOv9. We propose a method to replace the backbone network in YOLOv9 with Swin Transformer. This hybrid model combines the fast and efficient detection capabilities of

YOLOv9 with the powerful and flexible feature representation of Swin Transformer, designed to enhance the system's ability to accurately identify and classify strawberry ripeness from video input.

In the hybrid model, Swin Transformer acts as a powerful feature extractor by capturing the details and variations of strawberry appearance. These details and changes mark different stages of maturity. Swin Transformer ensures that global and local features are effectively captured and used for prediction. This is particularly useful for detecting strawberries under varying lighting, occlusion, and background complexity conditions.

3.3 Evaluation Methods

Our evaluation is a critical step for computer vision models, which helps measure model performance and guide future improvements. In deep learning, all evaluation methods are based on confusion matrix. Table 3.1 shows the confusion matrix. In Table3.1, True Positive(TP) means that the true category of the sample is a positive example, and the result predicted by the model is also a positive example, so the prediction is correct. True Negative (TN) means that the true category of the sample is a negative example, and the model predicts it as a negative example, so the prediction is correct. False Positive (FP) means that the true category of the sample is a negative example, but the model predicts it as a positive example, so the prediction is wrong. False Negative (FN) means that the true category of the sample is a positive example, but the model predicts it as a negative example, so the prediction is wrong (Caelen, 2017).

Table 3.1 Confusion matrix

	Positive	Negative
True	TP	TN
False	FP	FN

IoU (Intersection over Union) is a general evaluation index in the field of computer

vision, especially in tasks such as target detection and image segmentation. IoU mainly reflects the degree of overlap between the predicted bounding box and the ground truth bounding box. As shown in Figure 3.7, the green box is the truth bounding box, which is the box marked when labeling the data set. The red box is the predicted bounding box, which is the prediction box predicted by the trained model. As shown in Figure 3.8, IoU is the result of dividing the overlapping part of two areas by the set part of the two areas(Everingham et al., 2010).

$$IoU = \frac{Area\ of\ Overlap}{Area\ of\ Union} \tag{3.3}$$

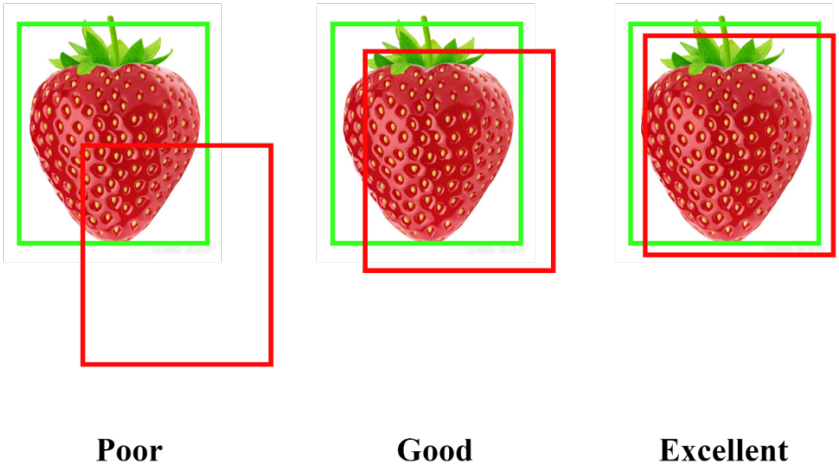

Figure 3.7 An example of bounding box

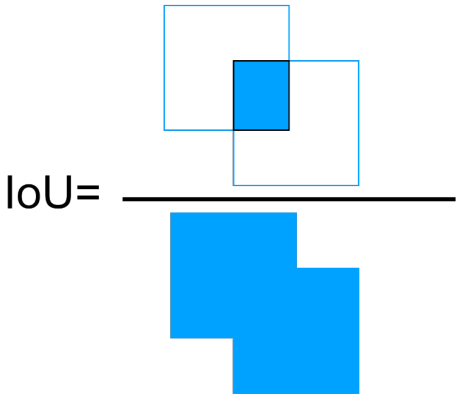

Figure 3.8 The method of IoU

Precision is an indicator for evaluating the performance of a classification model. It

measures the proportion of items that the model correctly identifies as positive out of all items that the model identifies as positive(Buckland & Gey, 1994).

$$\text{Precision} = \frac{TP}{TP+FP} \tag{3.4}$$

Although precision is an important metric, it does not provide a complete view of model performance on its own. Therefore, precision is often combined together with recall.

$$\text{Recall} = \frac{TP}{TP+FN} \tag{3.5}$$

To take into consideration of both Precision and Recall, F1 score is usually employed as an indicator to measure the overall performance of the model. The F1 score is the harmonic mean of the Precision and Recall.

$$\text{F1 score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.6}$$

where mAP (mean Average Precision) is an indicator widely used to evaluate model performance in computer vision tasks, especially in the fields of target detection and image retrieval. The mAP provides a single performance metric to evaluate the overall effectiveness of the model by comprehensively considering the precision and recall of the model under different categories and different detection difficulties. By plotting the curve of Precision versus Recall and calculating the area under the curve (AUC), the AP value of a single category is obtained. The mAP value is the average of the AP values of all categories. The higher the value, the better the performance of the model. The mathematical expressions are shown in Eq. (3.7) and Eq. (3.8). In this report, because the prediction results are divided into three classes, k=3.

$$AP = \int_0^1 p(r)dr \tag{3.7}$$

$$mAP = \frac{\sum_{i=1}^{k} AP_i}{k} \tag{3.8}$$

where mAP@IoU represents the mAP value calculated under a specific IoU threshold.

For example, mAP@0.5 means that the result is considered correct only when IoU$\geqslant$0.5. mAP@0.5 is a very popular evaluation metric because it considers the recognition accuracy and positioning accuracy of the model.

mAP@[.5:.95] is also a commonly used evaluation index, which calculates the average of all mAP values with IoU from 0.5 to 0.95 (in steps of 0.05). This approach is more rigorous as it considers a range of different IoU thresholds, providing a more comprehensive perspective on model performance.

Chapter 4
Results

The main content of this chapter is to collect video data and demonstrate the experimental results. In the end, this chapter will also discuss the limitations of the project.

4.1 Data Preparation

In deep learning, data collection is a critical step in the effective models. The performance of deep learning models largely depends on the quality and diversity of data used for training.

In this report, we collected various strawberry images and videos datasets to ensure the quality and accuracy of our models. For the data set, we used pre-processing techniques such as image cropping, resizing, and labeling to ensure that the data set is processed into the form required by the model.

We downloaded test videos of strawberry plantations from the internet and performed images extraction on the videos. We are use of a Python script to assist us in extracting images. This script allows us to split the video into images at set intervals. This is helpful for reducing data redundancy. Additionally, we downloaded strawberry images from the Internet to increase robustness.

As shown in Figure 4.1, we collected a total of 722 strawberry images. In addition, we also downloaded the strawberry image dataset, open sourced by the StrawDI team and selected images that met our requirements(Pérez-Borrero et al., 2020). Finally, we collected a total of 2,000 strawberry images from different regions and under different lighting and weather conditions, which helped to enhance model diversity.

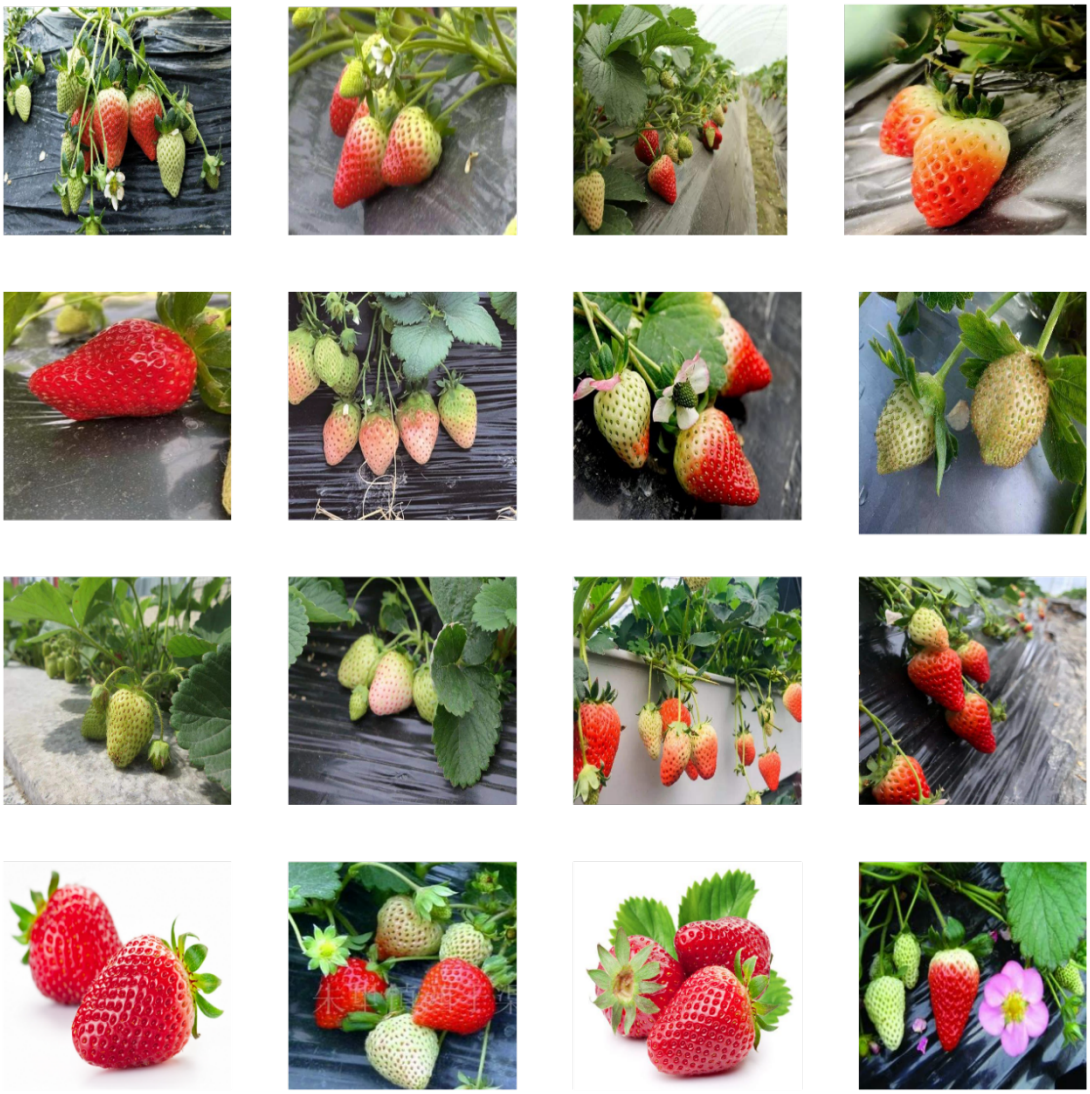

Figure 4.1 Samples of our dataset

Data labeling is an important part of preparing training data for machine learning models. This process involves assigning an accurate label or category to each sample in the dataset, enabling the model to learn from these labels and make predictions. In this report, we are use of EISeg to label the collected strawberry images. Figure 4.2 illustrates the results after labeling. EISeg (Efficient and Interactive Segmentation) is an efficient interactive image segmentation tool, mainly used in geospatial analysis, remote sensing image processing, medical image processing and other fields. EISeg provides a method to achieve precise segmentation with minimal user interaction, greatly improving the efficiency and accuracy of image segmentation (Xian et al., 2016).

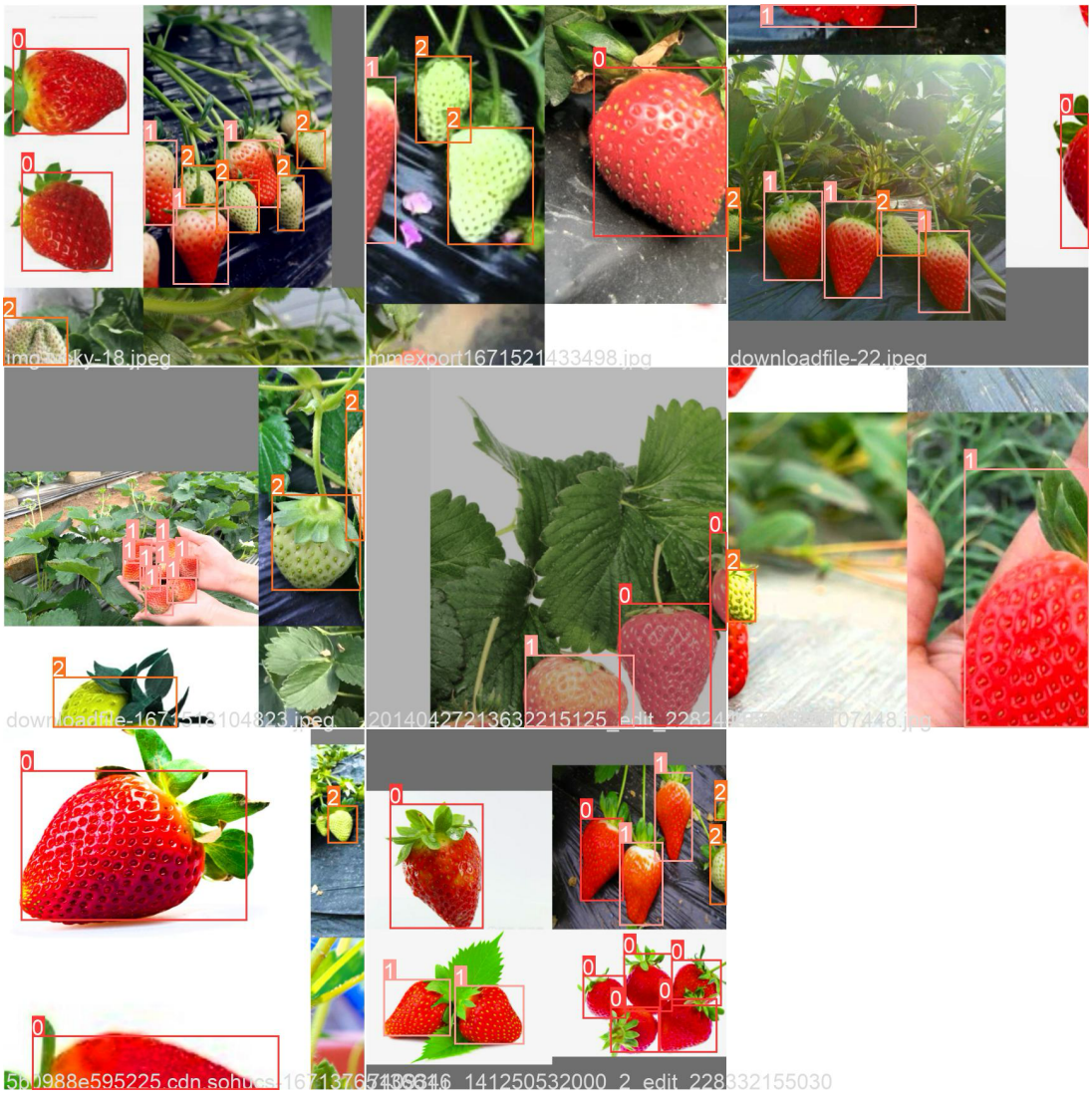

Figure 4.2 The example of the results after labeling

Data splitting is a fundamental technique in machine learning for training models and evaluating model performance. It involves dividing the dataset into separate subsets to provide an honest assessment of the performance of our proposed models on unseen data. The three main subsets commonly used are: Training dataset, validation dataset and test dataset. The training set is the largest part of the dataset used to train the model. The validation set is employed to provide an unbiased assessment of the model which fits on the training data set when adjusting the model's hyperparameters. After the model has been trained and validated, the test set is used for an unbiased evaluation of the final model. The correct data splitting can avoid model overfitting problems and significantly improve the validity and reliability of model evaluation.

In this report, we also split the data. We take use of 80% of the dataset for training,

10% for validation, and 10% for testing. Figure 4.3 clearly illustrates the dataset splitting.

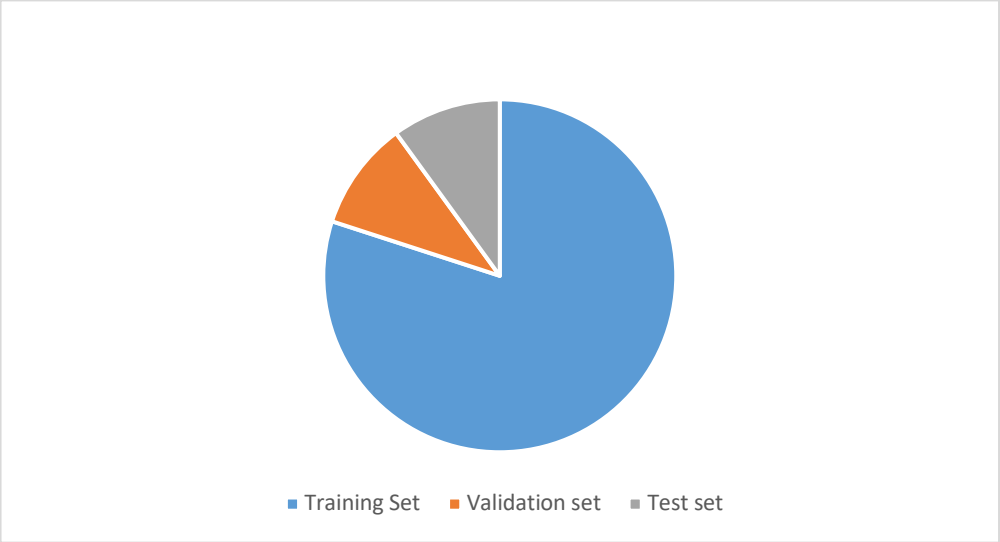

Figure 4.3 Data splitting pie chart

4.2 Performance of Strawberry Ripeness Detection Model

In this report, we trained a YOLOv9 network and Swin Transformer hybrid model for strawberry ripeness detection. We will analyze and discuss the results of the YOLOv9 model combined with Swin Transformer. In addition, we also used the dataset to train a model using only YOLOv9, and we compared the results of the two models to verify the advantages of adding Swin Transformer.

The experimental environment for the strawberry ripeness detection model is shown in Table 4.1. CUDA supports performing multiple operations in parallel by leveraging the power of the GPU, which can effectively improve the efficiency of training models.

Table 4.1 Experimental environment

Platform	ASUS TUF3
Python	vision 3.9
Pytorch	vision 2.2.1
CUDA	vision 12.1
GPU	RTX3060

As shown in Table 4.2, the parameters of model training are listed. The training

process includes 100 epochs, and the image batch processed in each epoch is 8. The initial value of the learning rate is 0.01. The table also details the parameters of various data enhancement techniques, such as HSV color space adjustment, scaling, and translation of images. Together, these parameters define how the model is trained, with the goal of optimizing the model's performance on a specific data set so that it can detect strawberry ripeness more accurately.

Table 4.2 Parameters of model training

Parameters	Values	Parameters	Values
models	yolov9-c.pt	lr0	0.01
cfg	yolov9_swin_ transfomrer.yaml	lrf	0.01
data	strawberry.yaml	momentum	0.937
epoch	100	weight_decay	0.0005
batch_size	8	warmup_epochs	3
warmup_momentum	0.8	warmup_bias_lr	0.1
box	7.5	cls	0.5
obj	0.7	dfl	1.5
iou_t	0.2	fl_gamma	0
hsv_h	0.015	hsv_s	0.7
hsv_v	0.4	degrees	0
translate	0.1	scale	0.9
shear	0	perspective	0
flipud	0	fliplr	0.5
mosaic	1.0	mixup	0.15

Figure 4.4 shows the true labels of the validation set in the strawberry ripeness data set. These validation sets can be employed to evaluate the performance of our detection model. As we know, in the dataset, the strawberries of different ripeness levels have completely different sizes, shapes, orientations, and environments, which is helpful for evaluating model predictions.

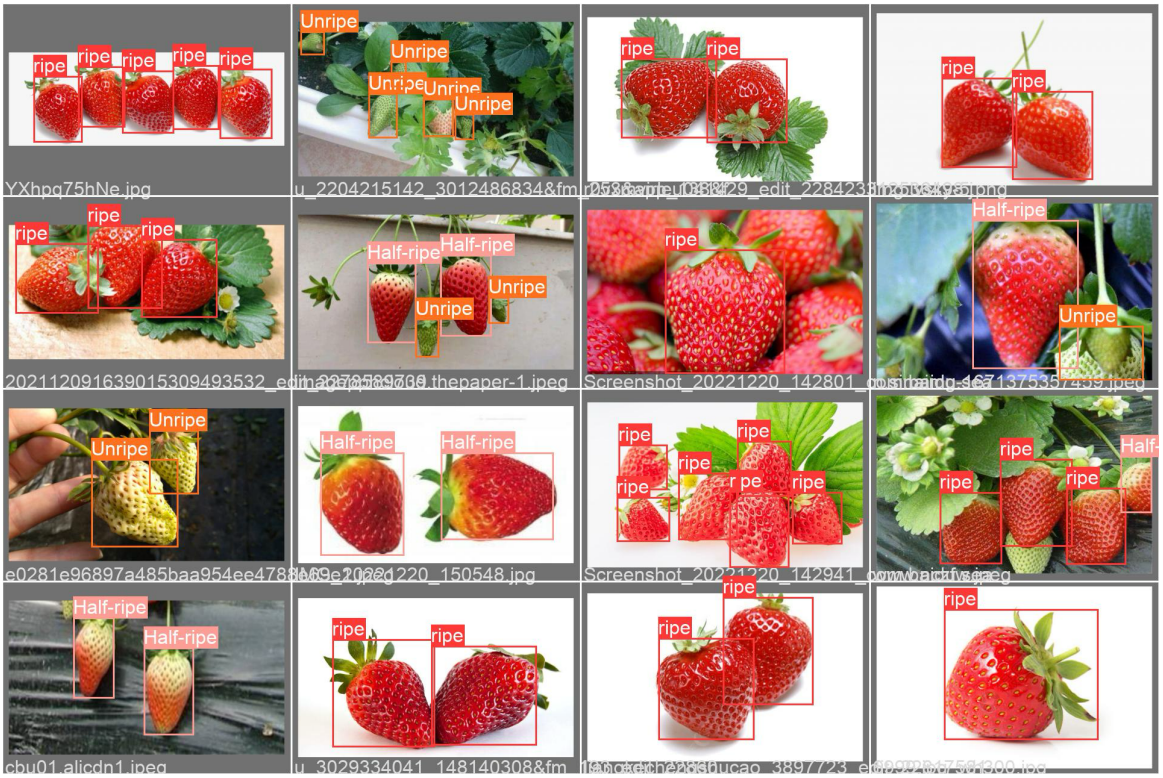

Figure 4.4 The true label situation of the validation set

Figure 4.5 shows the results of our model on the validation set. In the same validation dataset, our model is able to detect most strawberries of different sizes, orientations, environments, and ripeness. By comparing Figure 4.4 and Figure 4.5, it shows that our model is very effective.

We plotted the Precision-Recall (PR) curves of the two models as shown in Figures 4.6 and 4.7. The PR curve represents the relationship between precision and recall, where the thin line represents the PR curve of each category, and the thick line represents the average PR curve of all categories. The area under the PR curve (AUC) can be used to reflect the performance of the model. In the two figures, comparing the AUC values can compare the performance gap between the two models(Davis & Goadrich, 2006). For the two figures , comparing the AUC values can show how much the performance of the model has been improved by adding Swin Transformer. For the "ripe", the AUC increased from 0.925 to 0.933. For the "half- ripe", the AUC increased from 0.789 to 0.804. For the "unripe", the AUC increased from 0.868 to 0.882. For the sum of all, the AUC of mAP

@0.5 improves from 0.861 to 0.873. It is obvious that the performance of the model with Swin Transformer is improved at all ripeness, with higher precision and recall values. In short, the PR curve of our model performs well, can accurately detect the ripeness of strawberries, and is significantly improved compared to the YOLOv9 model.

Figure 4.5 Results of our model on the validation set

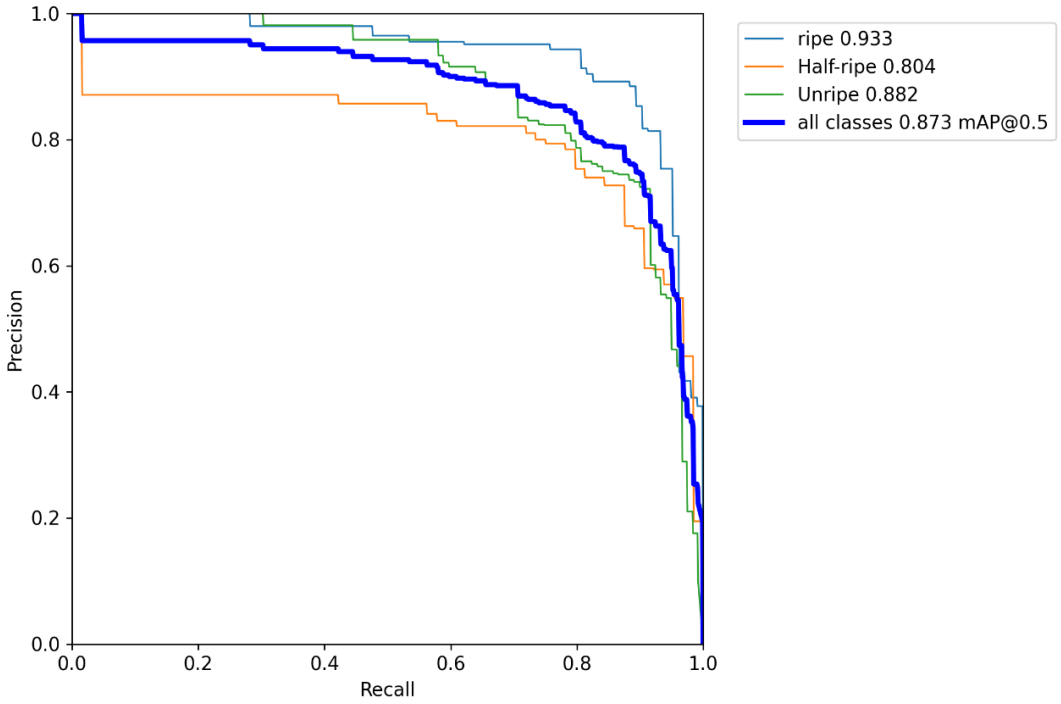

Figure 4.6 PR curve of YOLOv9+Swin Transformer model

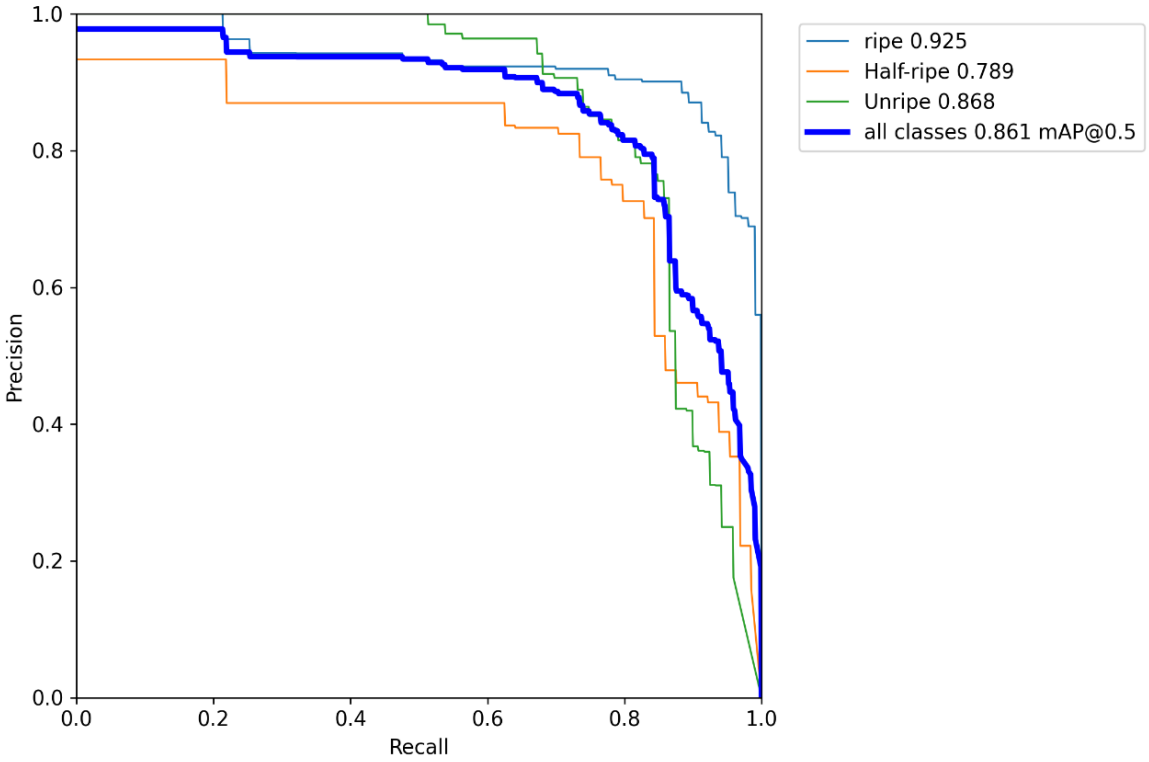

Figure 4.7 PR curve of YOLOv9 model

Figure 4.8 shows the evaluation metrics of the model. The “box loss” illustrates the bounding box regression loss for the training and validation datasets. The loss is significantly reduced in Figure 4.7, indicating that the model is getting better at predicting the correct bounding box of the object. Similarly, “obj loss” and “cls loss” represent objectivity loss and classification loss in model training. Their same losses decrease over time, which is positive.

The Precision and Recall of our model are both high and stable, indicating that the model performs well. mAP@0.5 is the average average accuracy with IoU threshold 0.5. mAP@[.5:.95] This shows the average accuracy calculated over multiple IoU thresholds from 0.5 to 0.95 in steps of 0.05. Figure 4.8 shows that mAP@ and mAP@[.5:.95] continue to increase, indicating that the model performs well under different levels of detection stringency. Overall, the model shows improvement over time in all aspects: bounding box prediction, object presence confidence, and classification. Precision and recall are both high. mAP is also excellent, reflecting excellent model performance.

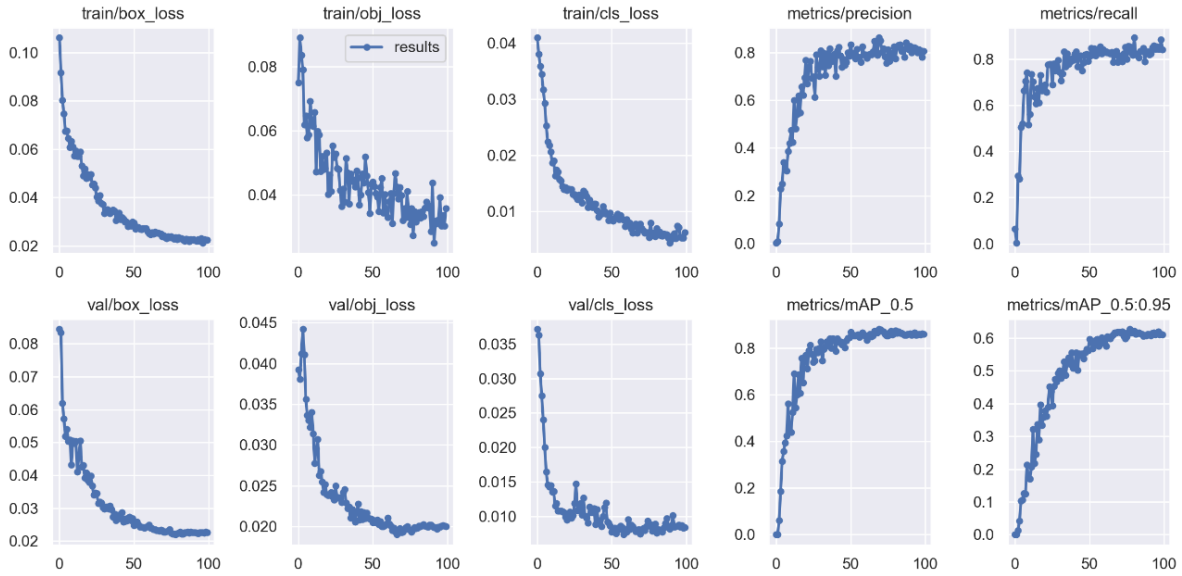

Figure 4.8 Plots of results of YOLOv9+Swin Transformer model

Based on the performance metrics provided in Table 4.2 for the two models, YOLOv9 with Swin Transformer and YOLOv9, at epoch 100, we make the following observations: YOLOv9+Swin Transformer has a higher Precision (0.853) compared to YOLOv9 (0.777) This means that the YOLOv9+Swin Transformer has a higher prediction rate. YOLOv9+Swin Transformer also scores higher on recall (0.840) compared to YOLOv9 (0.800). This indicates that YOLOv9+Swin Transformer is better at detecting all relevant objects in the dataset. The mAP@0.5 of YOLOv9+Swin Transformer (0.873) is slightly higher than that of YOLOv9 (0.861). And mAP@[.5:.95]: YOLOv9+Swin Transformer (0.627) is higher than YOLOv9 (0.610). Taken consideration of all these metrics, the YOLOv9+Swin Transformer model outperforms the standard YOLOv9 model on all listed performance metrics. This shows that our addition of Swin Transformer to YOLOv9 has significantly improved model performance.

Table 4.3 Performance comparison of the two models

Model	Epoch	Precision	Recall	mAP@0.5	mAP@[.5:.95]
YOLOv9+Swin Transformer	100	0.853	0.840	0.873	0.627
YOLOv9	100	0.777	0.800	0.861	0.610

Table 4.4 Comparison of strawberry ripeness detection models

Model	Precision	Recall	mAP@0.5	mAP@[.5:.95]
YOLOv9	0.777	0.800	0.861	0.610
YOLOv8	0.774	0.749	0.823	0.552
YOLOv8+Swin Transformer	0.815	0.831	0.861	0.613
YOLOv9+Swin Transformer	**0.853**	**0.840**	**0.873**	**0.627**

From Table 4.4, we can analyze the performance of different versions of the YOLO model in terms of Precision, Recall and mAP. The YOLOv9+Swin Transformer model has the highest Precision, reaching 0.853. In comparison, the Precision of the original YOLOv8 and YOLOv9 is slightly lower, with YOLOv9 being 0.777 and YOLOv8 being 0.774. YOLOv9+Swin Transformer reaches 0.840 in Recall, higher than the other three models. YOLOv9+Swin Transformer also has the highest mAP@0.5, reaching 0.873. On the more stringent mAP@[.5:.95], YOLOv9+Swin Transformer also showed the best performance, reaching 0.627. In summary, the YOLOv9+Swin Transformer model we proposed performs optimally on all major performance indicators. This further demonstrates that our method combining YOLOv9 and Swin Transformer can improve the performance of the strawberry detection model.

All in all, our strawberry ripeness detection model can accurately detect the ripeness of strawberries. All indicators of the model are very good, and our improvements to the model have proven to be very effective.

4.3 Demos and Discussions

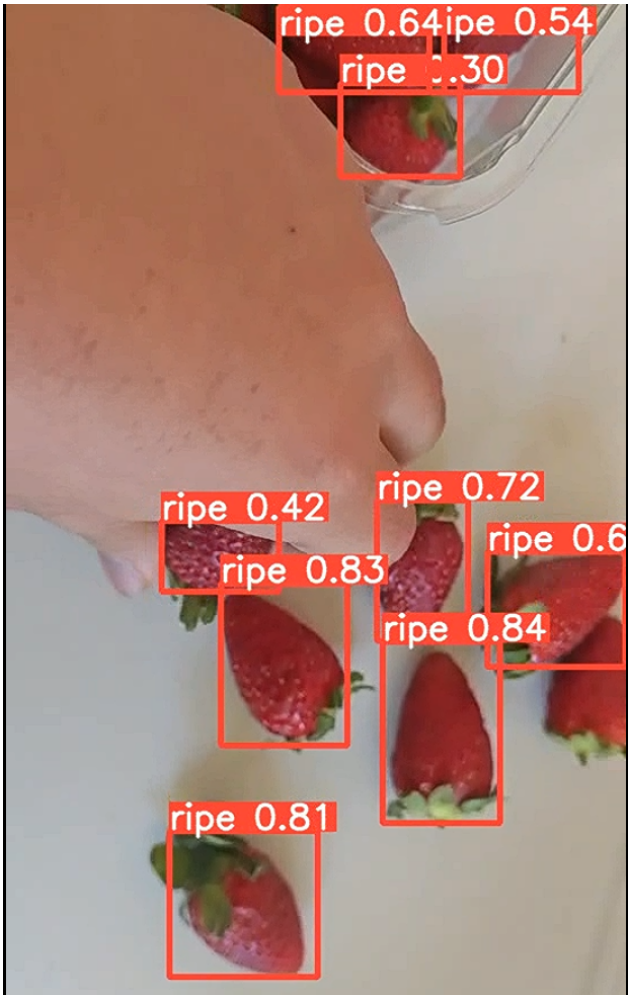

Figure 4.9 The demo for strawberry ripeness detection

Figure 4.9 is to show a strawberry ripeness detection demo based on our model. It clearly shows that our model can accurately identify the strawberries in the video and detect their ripeness. Even the strawberries in the image are not complete, which are blocked by a hand. However, our model can still clearly identify their ripeness, which shows that the performance of the model is robust and excellent.

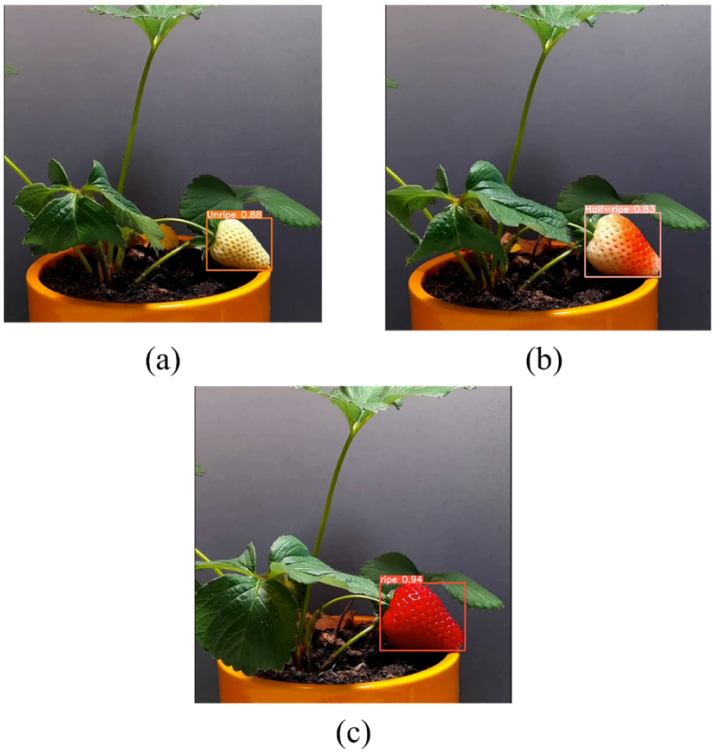

Figure 4.10 Different ripeness stages of strawberries

Figure 4.10 shows the model's detection results of different ripeness stages of strawberries. Our model successfully detected the ripeness as shown in Figure 4.10 (a), Figure 4.10 (b), and Figure 4.10 (c).

Based on the analysis of the two demos, our model has excellent performance and can detect strawberries of different ripeness in complex environments

Chapter 5
Analysis and Discussions

This chapter delves into a comprehensive analysis and discussion of the findings from the experimental results of the strawberry ripeness detection model.

5.1 Analysis

In summary, the hybrid model of YOLOv9 and Swin Transformer effectively improves the accuracy and reliability of strawberry ripeness detection. Indicators such as precision, recall, and mAP all show that the hybrid model of YOLOv9 and Swin Transformer has better detection results for strawberries of various ripeness levels.

5.2 Discussions

As shown in the previous chapters, our experimental results show that the hybrid model of YOLOv9 and Swin Transformer performs better than the YOLOv9 model. The key factors enabling these advances include:

Firstly, Swin Transformer can capture detailed and subtle features of strawberries, which greatly improves detection rates. This works particularly well in complex scenes where strawberries appear under various lighting and occlusion conditions.

Secondly, the architecture of YOLOv9, especially the integration of Programmable Gradient Information (PGI) and its lightweight and powerful network structure (GELAN), is able to locate strawberries quickly and accurately within video frames.

5.3 Further Research Work

We also conducted further research work. We juiced 500g of strawberries and analyzed the nutritional content. We list the nutritional value of 500g of strawberry juice. We calculate the main nutritional values of 500g of strawberry juice in Table 5.1.

Table 5.1 Nutritional value of 500g strawberry juice

Water	450g
Energy	680kJ
Sugars	35g
Fat	1.5g
Vitamin B9(Folate)	120µg
Vitamin C	295mg

Protein	0.25mg

In Table 5.1, we found 95% strawberry juice is water. The fat content is very low at 1.5 grams. The energy content is 680 kJ. This is equivalent to approximately 163 kcal. The calorie of 500 g kiwi juice is as high as 300 kcal (Dawes & Keene, 1999), while 500g apple juice is 240 kcal (Mattick & Moyer, 1983). Compared with other fruits, strawberries and strawberry juice are a low-calorie option and are more suitable for fitness and sports enthusiasts.

In Table 5.1, we find 500 g of strawberry juice contains 35 g of sugar, this value is low compared to other fruits. The high-water content and low sugar content make strawberries very diabetic friendly. In addition, strawberries are rich in dietary fiber, which slows down the rise of blood sugar after meals and helps control blood sugar. It should be noted that juicing may destroy the dietary fiber of strawberries. Therefore, diabetics should refrain from juicing strawberries (Xi et al., 2014).

The most important nutrients in strawberries are vitamins, of which, the most important are vitamin B9 and vitamin C.

Vitamin B9, also known as folate, is an important water-soluble vitamin. Folate acid plays an important role in DNA synthesis and repair, red blood cell formation and healthy brain function. Therefore, folate is also one of the most important nutrients for pregnant women. For adults, the recommended daily intake is 400 μg. Pregnant women are advised to increase their intake to 600 μg to support fetal development. 500 g of strawberry juice provides up to 120 μg of vitamin B9 (Ebara, 2017). There is no doubt that strawberries are very suitable for pregnant women because of their high vitamin B9 content and low sugar content.

Strawberries are also an excellent source of vitamin C, an antioxidant vital for immune and skin health.

In conclusion, strawberries have great nutritional value, the fruits are very friendly to fitness enthusiasts, diabetics, pregnant women and babies.

Chapter 6
Conclusion and Future Work

In this chapter, we will summarize the main findings, contributions, limitations, and future work of this report.

6.1 Conclusion

In this research project, we successfully demonstrated the integration of YOLOv9 and Swin Transformer models to detect strawberry ripeness with high accuracy. The hybrid model achieved a mean Average Precision (mAP) at an IoU of 0.5 of 87.3%, surpassing the performance of traditional models by using YOLOv9 alone, which registered a mAP of 86.1%. The Precision and Recall are better. This improvement underscores the effectiveness of combining these advanced deep learning technologies to enhance precision in agricultural applications. The ability of this proposed model to accurately categorize strawberries into unripe, half-ripe, and ripe stages can significantly aid in optimizing harvest times, thus reducing waste and increasing yield quality.

6.2 Limitations

While the results are promising, our research has several limitations:

Firstly, though the dataset includes images of strawberries from a variety of conditions, they are primarily from one variety. This limitation may affect the applicability of the model to different varieties of strawberries, such as strawberries that are white when ripe.

Secondly, our model has good performance. However, the performance of this proposed model in actual strawberry planting may be affected by external factors such as lighting and camera clarity.

Finally, the strawberry dataset we have proposed is limited in size and variety, and using more datasets may further improve model performance.

6.3 Future Work

Our future work remains to solidify the findings of this report and address its limitations.

Firstly, we will collect and integrate data from a wider range of climate and

geographic regions to improve the model's robustness and applicability in different agricultural settings.

Secondly, we will improve the model according to different varieties of strawberries to improve the general applicability of our model to various varieties of strawberries.

Finally, we will combine visual data with input from environmental sensors (e.g., humidity, temperature), which can improve the accuracy of maturity detection under different environmental conditions.

References

Al-Sarayreh, M., Reis, M., Yan, W., Klette, R. (2019) A sequential CNN approach for foreign object detection in hyperspectral images. International Conference on Information, Communications and Signal.

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

An, N., Yan, W. (2021) Multitarget tracking using Siamese neural networks. ACM Transactions on Multimedia Computing, Communications and Applications.

An, N. (2020) Anomalies Detection and Tracking Using Siamese Neural Networks. Master's Thesis. Auckland University of Technology, New Zealand.

Ashtiani, S. H. M., Javanmardi, S., Jahanbanifard, M., Martynenko, A., & Verbeek, F. J. (2021). Detection of mulberry ripeness stages using deep learning models. *IEEE Access*, 9, 100380-100394.

Astuti, I. F., Nuryanto, F. D., Widagdo, P. P., & Cahyadi, D. (2019, July). Oil palm fruit ripeness detection using K-Nearest neighbour. In Journal of Physics: Conference Series (Vol. 1277, No. 1, p. 012028). IOP Publishing.

Bharman, P., Saad, S. A., Khan, S., Jahan, I., Ray, M., & Biswas, M. (2022). Deep learning in agriculture: a review. *Asian J. Res. Comput. Sci*, 13, 28-47.

Bi, J., Zhu, Z., & Meng, Q. (2021, September). Transformer in computer vision. In *IEEE International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI)* (pp. 178-188). IEEE.

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

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei,

D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33, 1877-1901.

Buckland, M., & Gey, F. (1994). The relationship between recall and precision. *Journal of the American society for information science*, 45(1), 12-19.

Caelen, O. (2017). A Bayesian interpretation of the confusion matrix. *Annals of Mathematics and Artificial Intelligence, 81(3)*, 429-450.

Cao, L., Kang, S. B., & Chen, J. P. (2023, July). Improved lightweight YOLOv5s Algorithm for Traffic Sign Recognition. In *2023 3rd International Symposium on Computer Technology and Information Science (ISCTIS)* (pp. 289-294). IEEE.

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 International Publishing.

Davis, J., & Goadrich, M. (2006, June). The relationship between Precision-Recall and ROC curves. In *International Conference on Machine Learning* (pp. 233-240).

Dawes, H. M., & Keene, J. B. (1999). Phenolic composition of kiwifruit juice. *Journal of Agricultural and Food Chemistry*, *47*(6), 2398-2403.

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

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929.*

Ebara, S. (2017). Nutritional role of folate. *Congenital Anomalies*, *57*(5), 138-141.

Espejo-Garcia, B., Mylonas, N., Athanasakos, L., & Fountas, S. (2020). Improving weeds identification with a repository of agricultural pre-trained deep neural networks. *Computers and Electronics in Agriculture*, 175, 105593.

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.

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.

Gowdra, N., Sinha, R., MacDonell, S., Yan, W. (2021) Maximum Categorical Cross Entropy (MCCE): A noise-robust alternative loss function to mitigate racial bias in Convolutional Neural Networks (CNNs) by reducing overfitting. Pattern Recognition.

Gowdra, N. (2021) Entropy-Based Optimization Strategies for Convolutional Neural Networks. PhD Thesis, Auckland University of Technology, New Zealand.

Hasanuddin, N. H., Wahid, M. H. A., Azidin, M. A. M., Ahmad Hambali, N. A. M., Yusof, N. R., & Shahimin, M. M. (2015). Nondestructive fruit ripeness detection system and its spectral analyses. *Applied Mechanics and Materials*, 815, 394-397.

Hussain, M. (2023). YOLO-v1 to YOLO-v8, the rise of YOLO and its complementary nature toward digital manufacturing and industrial defect detection. *Machines*, 11(7), 677.

Jiao, Y., Zhang, Q., Luo, X., & Liang, Z. (2021, June). Fruit ripeness detection based on miniature spectral sensor. In *International Symposium on Computer Technology and Information Science (ISCTIS)* (pp. 93-98). IEEE.

Kamilaris, A., & Prenafeta-Boldu, F. X. (2018). Deep learning in agriculture: A survey. Computers and Electronics in Agriculture, 147, 70–90. https://doi.org/10.1016/j.compag.2018.02.016

Korostynska, O., Mason, A., & From, P. J. (2018, December). Electromagnetic sensing for non-destructive real-time fruit ripeness detection: Case-study for automated

strawberry picking. In *Proceedings* (Vol. 2, No. 13, p. 980). MDPI.

Lai, J. W., Ramli, H. R., Ismail, L. I., & Wan Hasan, W. Z. (2023). Oil palm fresh fruit bunch ripeness detection methods: a systematic review. *Agriculture*, 13(1), 156.

Li, C., Li, L., Jiang, H., Weng, K., Geng, Y., Li, L., ... & Wei, X. (2022). YOLOv6: A single-stage object detection framework for industrial applications. *arXiv preprint arXiv:2209.02976.*

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

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

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

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., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., & Hu, H. (2022). Video swin transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3202-3211).

Mattick, L. R., & Moyer, J. C. (1983). Composition of apple juice. *Journal of the Association of Official Analytical Chemists*, *66*(5), 1251-1255.

Momeny, M., Jahanbakhshi, A., Neshat, A. A., Hadipour-Rokni, R., Zhang, Y. D., & Ampatzidis, Y. (2022). Detection of citrus black spot disease and ripeness level in orange fruit using learning-to-augment incorporated deep networks. *Ecological Informatics*, 71, 101829.

Mu, Y., Chen, T. S., Ninomiya, S., & Guo, W. (2020). Intact detection of highly occluded immature tomatoes on plants using deep learning techniques. *Sensors*, 20(10), 2984.

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., … Zoph, B. (2024). GPT-4 technical report. *arXiv preprint arXiv:2303.08774.*

Pan, C., Yan, W. (2018) A learning-based positive feedback in salient object detection. International Conference on Image and Vision Computing New Zealand.

Pan, C., Yan, W. (2020) Object detection based on saturation of visual perception. Multimedia Tools and Applications, 79 (27-28), 19925-19944.

Pan, C., Liu, J., Yan, W., Zhou, Y. (2021) Salient object detection based on visual perceptual saturation and two-stream hybrid networks. IEEE Transactions on Image Processing.

Pardede, J., Sitohang, B., Akbar, S., & Khodra, M. L. (2021). Implementation of transfer learning using VGG16 on fruit ripeness detection. *Int. J. Intell. Syst. Appl, 13*(2), 52-61.

Peng, W., & Karimi Sadaghiani, O. (2023). A review on the applications of machine learning and deep learning in agriculture section for the production of crop biomass raw materials. *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, 45(3), 9178-9201.

Pérez-Borrero, I., Marín-Santos, D., Gegundez-Arias, M. E., & Cortés-Ancos, E. (2020). A fast and accurate deep learning method for strawberry instance segmentation. *Computers and Electronics in Agriculture*, 178, 105736.

Qi, J., Nguyen, M., Yan, W. (2022) Small visual object detection in smart waste classification using Transformers with deep learning. International Conference on Image and Vision Computing New Zealand (IVCNZ).

Qi, J., Nguyen, M., Yan, W. (2022) Waste classification from digital images using ConvNeXt. Pacific-Rim Symposium on Image and Video Technology (PSIVT).

Qi, J., Nguyen, M., Yan, W. (2023) CISO: Co-iteration semi-supervised learning for visual object detection. Multimedia Tools and Applications.

Qi, J., Nguyen, M., Yan, W. (2024) NUNI-Waste: Novel semi-supervised semantic segmentation for waste classification with non-uniform data augmentation. Multimedia Tools and Applications.

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).

Redmon, J., & Farhadi, A. (2018). YOLOv3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.

Ren, C., Kim, D.-K., & Jeong, D. (2020). A survey of deep learning in agriculture: techniques and their applications. *Journal of Information Processing Systems*, 16(5), 1015–1033.

Sharma, R., Kukreja, V., & Bordoloi, D. (2023, May). Deep learning meets agriculture: A Faster R-CNN based approach to pepper leaf blight disease detection and multi-classification. In *International Conference for Emerging Technology (INCET)* (pp. 1-5). IEEE.

Tang, S., Yan, W. (2024) Utilizing RT-DETR model for fruit calorie estimation from digital images. Information.

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

Processing Systems, 30.

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, C. Y., Liao, H. Y. M., & Yeh, I. H. (2022). Designing network design strategies through gradient path analysis. *arXiv preprint arXiv:2211.04800.*

Wang, C. Y., Liao, H. Y. M., Wu, Y. H., Chen, P. Y., Hsieh, J. W., & Yeh, I. H. (2020). CSPNet: A new backbone that can enhance learning capability of CNN. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 390-391).

Wang, C. Y., Yeh, I. H., & Liao, H. Y. M. (2024). YOLOv9: Learning what you want to learn using programmable gradient information. *arXiv preprint arXiv:2402.13616.*

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

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).

Xi, B., Li, S., Liu, Z., Tian, H., Yin, X., Huai, P., ... & Steffen, L. M. (2014). Intake of fruit juice and incidence of type 2 diabetes: A systematic review and meta-analysis. *PloS one*, *9*(3), e93471.

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.

Xian, M., Xu, F., Cheng, H. D., Zhang, Y., & Ding, J. (2016, December). EISeg: Effective interactive segmentation. In *International Conference on Pattern Recognition (ICPR)* (pp. 1982-1987).

Xiao, B., Nguyen, M., Yan, W. (2021) Apple ripeness identification using deep learning. International Symposium on Geometry and Vision.

Xiao, B., Nguyen, M., Yan, W. (2023) Apple ripeness identification from digital images using transformers. Multimedia Tools and Applications, Springer Science and Business Media LLC.

Xiao, B., Nguyen, M., Yan, W. (2023) Fruit ripeness identification using transformers. Applied Intelligence, Springer Science and Business Media LLC.

Xiao, B., Nguyen, M., Yan, W. (2023) A mixture model for fruit ripeness identification in deep learning. Handbook of Research on AI and ML for Intelligent Machines and Systems, pp.1-16, Chapter 16, IGI Global.

Xiao, B., Nguyen, M., Yan, W. (2023) Fruit ripeness identification using YOLOv8 model. Multimedia Tools and Applications.

Xiao, B. (2024) Fruit Ripeness Identification from Digital Images Using Deep Learning. PhD Thesis, Auckland University of Technology, New Zealand.

Xue, Y. Yan, W. (2023) YOLO models for fresh fruit classification from digital videos. Handbook of Research on AI and ML for Intelligent Machines and Systems, pp. 421-435, Chapter 17, IGI Global.

Xue, Y. (2024) YOLO Models for Fresh Fruit Classification from Digital Videos. Master's Thesis. Auckland University of Technology, New Zealand.

Yan, W. (2023) Computational Methods for Deep Learning: Theory, Algorithms, and Implementations (2nd Edition). Springer.

Yan, W. (2019) Introduction to Intelligent Surveillance: Surveillance Data Capture,

Transmission, and Analytics (3rd Edition). Springer.

Yang, S., Wang, W., Gao, S., & Deng, Z. (2023). Strawberry ripeness detection based on YOLOv8 algorithm fused with LW-Swin Transformer. *Computers and Electronics in Agriculture*, 215, 108360.

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

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.

Zhu, Y., Yan, W. (2022) Image-based storytelling using deep learning. ACM ICCCV.