

國立臺灣大學生物資源暨農學院生物機電工程學系

碩士論文

Department of Biomechatronics Engineering

College of Bioresources and Agriculture

National Taiwan University

Master Thesis

應用深度學習方法於茶菁影像識別

及分級系統之開發

Application of Deep Learning Methods to Develop

Plucked Tea Shoot Image Identification

and Grading System

王鼎慈

Ding-Ci Wang

指導教授：陳世芳 博士

Advisor: Shih-Fang Chen, Ph.D.

中華民國 112 年 8 月

August 2023





## 致謝

碩士生活即將結束，回顧這兩年，彷彿無比漫長，卻又轉瞬即逝。那個拜訪實驗室的我，難以想像如今能懷著滿胸腔的感激寫下致謝。鼓起勇氣來台大攻讀碩士是我無悔且正確的選擇。兩年前懵懂的大學生，現在即將成為不畏挑戰，有足夠能力的畢業生，因此，謝謝當初那個勇敢的自己。

感謝兩年來栽培我的陳世芳老師，在剛入學時引導我以適當的速度自學、成長。也謝謝老師，在繁忙的茶葉計畫中信任我，讓我們一同順利完成計畫與研究。老師在生活上的關心也讓本該枯燥的碩士生活充斥溫情，謝謝最關心學生的世芳老師。另外，也特別感謝林秀榮老師，教導我茶葉的知識、提供研究資料、在計畫上處處幫忙，謝謝最罩、最愛吃辣的秀榮老師。也感謝口試委員陳右人老師、蔡志賢老師與郭彥甫老師，謝謝老師們的指導，讓我能以更全面的角度檢視這份研究，我能受到老師們的肯定倍感榮光。

謝謝實驗室的夥伴們，知識與程式能力甲冠天下的育堂大神、細心負責的依芳、實作強大的熊哥、擅長模型的薛哥、做事周到的世鈺，無所不能的彥碩，這些實驗室前輩為我樹立了優良楷模。還有一同排憂解難的 R10 們，彥成與昱宏，和你們共度的時光，在計劃上、論文上、修課上、實驗室中，得以新增了不少歡笑。還有篆澤、Jimmy、知芸、廷睿、煒翔、凱鈞、騏瑞，謝謝你們的陪伴，我即將卸下碩士身份，交棒給你們了，祝福研究順利。

最後，感謝我的父母，謝謝你們即使不了解我的研究也無條件支持我。謝謝我在成功高中交的六個兄弟，一同吃飯出遊喝酒解憂。我終於要畢業了，碩士兩年的所學，將使我一生受用。


## 摘要



臺灣茶聞名世界，優良的品質與豐富的香氣受到廣大茶飲愛好者的喜愛。優良的成茶品質，除受製造過程影響外，原料端鮮採茶菁的狀態更是影響品質的關鍵因素，因此製茶廠會對原料端之茶菁予以品質分級，建立以質制價標準，並依茶菁之適製性調整後續製茶之目標品項。然而，現今茶廠的茶菁分級方式高度仰賴人力主觀地依其經驗，以肉眼判別分級，難以自動化執行分級制價。且因各有其主觀立場，容易產生產購雙方，即茶農及製茶廠間的採購制價糾紛。因此，若能建立一套智慧茶菁分級系統，此系統可依明確且客觀的方式，自動且快速的執行茶菁分級，將有助於緩解產購糾紛。

為達成智慧分級的目的，本研究應用深度學習方法建立此一茶菁影像分級系統，本系統主要由兩部分組成，包含硬體端的茶菁取像模組，及軟體端的茶菁分級模型。整體系統運作方式如下：首先，取像模組拍攝待分級之茶菁影像，將其上傳至雲端系統，接續由茶菁分級模型辨識影像中每株茶菁的等級。茶菁分級模型由茶菁提取模型、部位識別模型與茶菁分級標準三部分組成，茶菁提取模型偵測影像中的單株茶菁，而部位識別模型辨識單株茶菁之茶芽、嫩葉、熟葉、老葉、莖、魚葉及紅梗等七類，根據部位辨識結果，茶菁分級標準可將每株茶菁分為 A (優良)、B (尚可)、C (欠佳)三個等級。本研究茶菁影像蒐集自桃園、新竹、苗栗及南投等地茶廠，包含'臺茶 1 號'、'臺茶 12 號'、'臺茶 17 號'與'青心大有'等四茶樹品種，蒐集之 825 張批次茶菁影像用於建立茶菁提取模型，亦蒐集 1337 張單株茶菁影像用於建立部位識別模型。茶菁提取模型與部位識別模型採用 Mask2Former (Masked-attention Mask Transformer) 深度學習方法。

茶菁分級系統在單株茶菁提取的平均準確度均值 (mean Average Precision, mAP) 達 0.88；而在物件辨識上，對於嫩葉、熟葉及老葉等大物件的辨識率，分



別為 0.60、0.71 及 0.53，對於莖、茶芽、紅梗及魚葉等小物件的辨識率，分別在 AP50 時得相應表現為 0.86、0.51、0.48 及 0.65，整體茶菁分級準確率達 0.87。為使茶菁分級模型可迎合使用者便利性與紀錄查詢之功能，故建置資料庫與服務網頁於茶菁分級系統。服務網頁為應用 NodeJS 進行開發，使用者可登入網頁觀察茶菁影像、分級資訊與部位辨識結果，並根據產季、茶樹品種或採購策略彈性調整茶菁分級標準。應用本茶菁分級系統，鮮採茶菁可依明確且客觀的方式進行品質分級，期望緩解茶農及製茶廠間的茶菁採購制價糾紛。

關鍵詞：茶菁分級、深度學習、物件辨識、Transformer 模型。

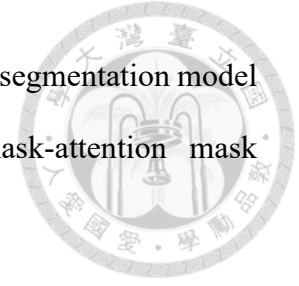
## ABSTRACT



Taiwanese tea is renowned worldwide for its excellent quality and fragrant aroma. The plucked tea shoot highly influences the quality of tea production. Therefore, tea factories grade the plucked tea shoot according to its quality, and the purchasing price can be determined by its grade. Tea factories also adjust the tea processing based on the grade to achieve better production. However, the grading method is performed with naked-eye observation and subjective judgment by human experience, leading to procurement conflict between the factories and tea farmers. An objective and automatic grading method is necessary and expected to mitigate this issue.

This study aimed to develop a tea shoot grading system (TSGS) using deep learning methods. The system consisted of a tea shoot image capture module and a tea shoot grading model. The overall operation of the system is as follows: Firstly, the module captured the image of plucked tea shoot and then uploaded it to the system. Next, the grading model graded the quality of each tea shoot in the uploaded image. The grading model included a single tea shoot segmentation model, an organ identification model, and grading criteria. The single tea shoot segmentation model extracted single tea shoot from a batch image, and the organ identification model identify seven types of organs of each single tea shoot, including buds, fresh leaves, mature leaves, old leaves, stems, fish leaf, and red stalks. Based on the identified organs, the grading criteria could classified tea shoots into three grades: A (excellent), B (ordinary), and C (defective). The collected tea samples in this study included “*Taiwan tea experiment station (TTES) No.12*”, “*TTES No.12*”, “*TTES No.17*”, and “*Chin-Shin-Dapan*”, acquired from Taoyuan, Hsinchu, Miaoli, and Nantou regions. There were 825 tea shoot batch images being collected to develop the single tea shoot segmentation model, and 1337 single tea shoot images were

used to establish the organ identification model. The single tea shoot segmentation model and organ identification model were developed using the mask-attention mask transformer (Mask2Former).



The identification performances of TSGS were as follows. The mean average precision (AP) of the single tea shoot segmentation model was 0.88. The organ identification model reached mAP values of 0.60, 0.71, and 0.53 for fresh, mature, and old leaves, respectively. The AP50 values for stems, buds, red stalks and fish leaves were 0.86, 0.51, 0.48, and 0.65, respectively. The overall accuracy of the grading model was 0.87. The TSGS also implemented a database and web development to enhance user experience, facilitating for convenient utilization of the grading model and data retrieval. The web development was built on NodeJS. Users could log in the web page to observe uploaded images, grading information, and the organ identification results. The web development also allowed users to adjust the grading criteria based on the harvest season, tea cultivar, and purchasing strategy. By applying TSGS, the plucked tea shoot could be graded clearly and objectively. The promising result was expected to alleviate pricing disputes between tea farmers and tea processing factories.

**Keywords:** plucked tea shoot grading, deep learning, object detection, transformer.

# TABLE OF CONTENTS



致謝 .....	i
摘要 .....	ii
ABSTRACT .....	iv
TABLE OF CONTENTS .....	vi
LIST OF FIGURES .....	viii
LIST OF TABLES .....	x
ABBREVIATIONS .....	xi
CHAPTER 1. INTRODUCTION .....	1
1.1 Research Background .....	1
1.2 Objectives .....	2
CHAPTER 2. LITERATURE REVIEW .....	3
2.1 Tea shoots Quality in Different Plucking Method .....	3
2.2 Quality Evaluation of Plucked Tea Shoot .....	4
2.3 Computer Vision on Tea Related Study .....	5
2.3.1 Plucked Tea Shoot Identification .....	5
2.3.2 Tea Quality Estimation .....	6
2.4 Deep Learning .....	7
2.4.1 Object Detection .....	8
2.4.2 Transformer .....	9
CHAPTER 3. MATERIALS AND METHODS .....	11
3.1 Data Acquisition .....	11
3.2 Development of Grading Criteria .....	12
3.3 Datasets Preparation .....	14
3.3.1 Single Tea Shoot Segmentation Dataset .....	14
3.3.2 Organ Identification Dataset .....	15
3.4 Models Architecture .....	16
3.4.1 Masked-attention Mask Transformer .....	16

3.4.2 Shifted Windows Transformer .....	18
3.5 Development of TSGS.....	19
3.5.1 Tea Shoot Grading Model .....	19
3.5.2 Design of Tea Shoot Image Capture Module .....	21
3.5.3 Design of Web Developments .....	22
3.6 Evaluation Metrics.....	23
3.7 Tea Leaf Overlap Ratio Test .....	24
<b>CHAPTER 4. RESULTS AND DISCUSSION .....</b>	<b>26</b>
4.1 Single Tea Shoot Segmentation Model .....	26
4.2 Organ Identification Model Performance.....	29
4.2.1 Metrics in Large Object and Small Object .....	29
4.2.2 Backbone Substitution.....	30
4.2.3 Data Augmentation.....	32
4.3 TSGS Performance .....	33
4.3.1 Grading Accuracy.....	35
4.3.2 Processing Speed .....	36
4.4 Discussion: Overlapping leaves .....	37
4.4.1 Evaluation of Overlap Ratio Test .....	37
4.4.2 Tea Shoot Segmentation in Overlapping.....	40
4.4.3 Grading Accuracy in Overlapping.....	41
4.5 Tea Shoot Image Capture Module.....	43
4.6 Web Developments .....	44
<b>CHAPTER 5. CONCLUSION AND FUTURE WORK .....</b>	<b>49</b>
5.1 Conclusion.....	49
5.2 Future work .....	50
<b>REFERENCES .....</b>	<b>51</b>
<b>APPENDIX .....</b>	<b>57</b>
<b>Appendix A. Tea Shoot Labeling for Overlap Ratio Calculation .....</b>	<b>57</b>



## LIST OF FIGURES



<b>Figure 3.1</b> Five harvest seasons in a year. ....	11
<b>Figure 3.2</b> Examples of two different backgrounds.....	12
<b>Figure 3.3</b> Grading criteria for plucked tea shoot.....	13
<b>Figure 3.4</b> Tea shoot annotation in labelme.....	14
<b>Figure 3.5</b> Organs of tea shoots. ....	15
<b>Figure 3.6</b> The architecture of Mask2Former. ....	16
<b>Figure 3.7</b> The architecture of Swin transformer. ....	18
<b>Figure 3.8</b> The architecture of TSGS.....	19
<b>Figure 3.9</b> The architecture of tea shoot grading model. ....	20
<b>Figure 3.10</b> Components for capture module. ....	21
<b>Figure 3.11</b> Web developments architecture diagram of TSGS.....	22
<b>Figure 3.12</b> Definition of confusion matrix. ....	23
<b>Figure 3.13</b> An example of the overlapped tea shoots.....	25
<b>Figure 4.1</b> Training curves of single tea shoot segmentation model. ....	26
<b>Figure 4.2</b> Successful tea shoot segmentation. ....	27
<b>Figure 4.3</b> Leaves calculation of tea shoot segmentation model. ....	27
<b>Figure 4.4</b> Error cases of tea shoot segmentation. ....	28
<b>Figure 4.5</b> Incorrectly identification of overgrowth tea shoot. ....	29
<b>Figure 4.6</b> Small organ annotation and prediction.....	30
<b>Figure 4.7</b> Successful organ identification. ....	30
<b>Figure 4.8</b> Error case of red stalk identification. ....	31
<b>Figure 4.9</b> Model predictions in rotated case.....	32
<b>Figure 4.10</b> Identification of augmented model and base model.....	33

<b>Figure 4.11</b> The confusion matrix of manual grading. ....	34
<b>Figure 4.12</b> Confusion matrix between model and human. ....	35
<b>Figure 4.13</b> Inaccurate leaf number calculations. ....	35
<b>Figure 4.14</b> The distribution of overlapping tea shoots. ....	38
<b>Figure 4.15</b> Misclassification of identifying the overlapped tea shoots. ....	40
<b>Figure 4.16</b> Demonstration of the tea shoot image capture module. ....	43
<b>Figure 4.17</b> General grading function mode for tea shoot grading. ....	45
<b>Figure 4.18</b> Standalone function mode for organ investigation. ....	45
<b>Figure 4.19</b> Database query webpage. ....	46
<b>Figure 4.20</b> Decision tree for grading criteria. ....	47
<b>Figure 4.21</b> Flexible database design for user-definable grading criteria. ....	48
<b>Figure A.1</b> Overlapping tea shoot labeling using fixed-contour method. ....	57



## LIST OF TABLES



<b>Table 3.1</b> Number of tea shoots images and cultivars in five harvest seasons. ....	11
<b>Table 3.2</b> Number of images of single tea shoot segmentation dataset. ....	14
<b>Table 3.3</b> Number of instances of organ identification. ....	16
<b>Table 4.1</b> Performance of organ identification model.....	31
<b>Table 4.2</b> Performance of augmented model and base model.....	33
<b>Table 4.3</b> Scenarios demonstration: overlapped tea leaves under various IoU and overlap ratio. ....	39
<b>Table 4.4</b> Number of images of overlap ratio. ....	39
<b>Table 4.5</b> Separation performance in overlapping. ....	40
<b>Table 4.6</b> Grading accuracy for overlapped tea shoots in two cases.....	41

## ABBREVIATIONS

Aggregated residual transformations	ResNeXt
Average precision	AP
Average Precision for Intersection of Union at 0.50	AP50
Average Precision for Intersection of Union at 0.75	AP75
Convolution neuron network	CNN
Data-efficient image transformers	DeiT
Deep learning	DL
Detection transformer	DETR
Efficient neural network	ENet
Faster region-based convolutional neural network	Faster R-CNN
Internet of things	IoTs
Intersection over union	IoU
Mask region-based convolutional neural network	Mask R-CNN
Masked-attention mask transformer	Mask2Former
Natural language processing	NLP
Radial basis function kernel	RBF
Region proposal networks	RPNs
Regional convolutional neural network	R-CNN
Residual neural network	ResNet
Secure file transfer protocol	SFTP
Shifted windows transformer	Swin
Taiwan tea experiment station	TTES
Tea shoot grading system	TSGS
Tea research and extension station	TRES
Transformer-based set prediction with RCNN	TSP-RCNN
Transformer in transformer	TNT
Vision transformer	ViT
Visual geometry group	VGG
You-Only-Look-Once	YOLO



## CHAPTER 1. INTRODUCTION



### 1.1 Research Background

Tea is one of the most popular beverages all over the world. Countries like China, India, Sri Lanka, Japan, and Taiwan are known for their tea production. The global tea market increased to 55.1 billion US dollars in 2019 and is estimated to surpass 68.9 billion US dollars by 2027 (Kumar and Deshmukh 2022). According to the report from the FAOSTAT, global tea production reached 6.5 million tonnes in 2021 (Food and Agriculture Organization Statistical Database, 2022).

Quality assurance for raw materials is vital to maintaining the overall quality of tea production. Plucked tea shoot is the raw material used to produce tea production. The quality of each plucked tea shoot is significantly influenced by its characteristics, such as the tenderness of leaves, the stem's length, and the lignified portion. Tea shoots were classified into different grades according to their quality. The grade of each tea shoot should be determined by examining their characteristics. According to the grade, tea factories and farmers can agree on fair prices. Tea shoot grading relies on subjective judgment according to experience and observation. However, this subjective method often leads to procurement conflicts between tea factories and farmers due to cognitive disparity.

An objective and standardized grading method could be beneficial in addressing this issue by providing a clear and reliable grading process based on the characteristics of tea shoots, aiming to minimize cognitive disparity. This grading method could be established by transforming subjective experiences into an objective and explicit grading process.

Such a method would benefit tea factories and farmers by reducing procurement conflict. Additionally, sharing grading information with farmers would be helpful for further tea plantation management and harvesting strategies accordingly.



## 1.2 Objectives

This study aimed to develop a tea shoot grading system (TSGS) to provide automated and objective grading procedures (Figure 1.1). TSGS was designed to acquire images of tea shoots and grade each tea shoot into one of three grades: A (excellent), B (ordinary), or C (defective). Moreover, a web application was developed to display the images and grading results. The objectives of this study were to:

1. Build a tea shoot image capture module to automatically capture and upload on-site images to the grading model for tea shoot samples.
2. Develop empirical grading criteria based on traits of tea shoot and a tea shoot grading model for the tea shoot using deep learning architectures.
3. Establish web developments with user-friendly interfaces and flexible applications.

## CHAPTER 2. LITERATURE REVIEW



### 2.1 Tea shoots Quality in Different Plucking Method

The plucking methods profoundly affect the quality of harvested tea shoots. The tea-plucking machine provides several advantages, such as increased speed, labor savings, and automated procedure, and it is suitable for large-scale plantations with flat terrain. However, the machine has certain drawbacks, including the inconsistent appearance of the harvested tea shoots, broken leaves, and twigs. Conversely, hand-plucking allows for selective picking points, involving meticulous cutting points of tea shoots, resulting in a more uniform quality while minimizing the presence of undesirable parts. However, this approach is slower and requires higher labor costs due to the manual effort demanded. The tea-plucking machine could reduce production costs by 50 to 70% compared to hand-plucking and increase efficiency by 8 to 15 times (Wu, 2015). Nevertheless, the amount of undesirable parts content in mechanically harvested tea shoots is three times higher than in manually harvested ones (Wijeratne, 2012).

Several tea-producing countries have utilized Mechanized tea harvesting, including Taiwan, Japan, Argentina, China, Sri Lanka, and India. Technological advancements have further improved the efficiency of tea-plucking machines. While tea-plucking machines reduce production costs and time, careful control is required to avoid impacts on tea quality. The choice between the tea-plucking machine and hand-plucking depends on various factors, such as the scale of the plantation, terrain, labor availability, and the desired quality management. It is crucial to balance these considerations in selecting the plucking method to achieve optimal alignment with specific requirements.

## 2.2 Quality Evaluation of Plucked Tea Shoot

The quality of plucked tea shoot is related to its chemical composition, the plantations where it grown, and its traits. Koch et al. (2018) demonstrated the correlation between the composition of plucked tea shoot, such as catechins, antioxidants, and metal content, and its quality. The composition is significantly influenced by environmental conditions. Therefore, even for tea plants belonging to the same species, their metabolism and mineral status can vary depending on the soil and climate conditions. This variation can ultimately affect the final quality of the produced tea (Hazra et al., 2021).

Also, plucked tea with different traits may present varying flavor profiles. For example, the younger part of the tea plant, such as the bud and younger leaf, contains higher levels of amino acids and catechins, resulting in a savory taste (Xu et al., 2021). In addition, tea made from plucked tea shoots with longer twigs tends to present more undesirable flavors. As a result, the made tea produced by younger parts is generally sold at a higher price in the premium tea market. In contrast, tea made from mature and old leaves with higher fiber content is more suitable for being the material for general beverage.



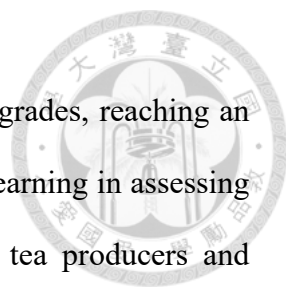


## 2.3 Computer Vision on Tea Related Study

### 2.3.1 Plucked Tea Shoot Identification

The application of deep learning can benefit in assessing the quality of plucked tea shoot. Gayathri et al. (2020) used a convolution neuron network (CNN) to identify the tea disease on the image of a single leaf, reaching an accuracy of 0.90. However, the actual field images were possible to contain a complex background and multiple leaves. Lee et al. (2020) applied a faster region-based convolutional neural network (Faster R-CNN) to detect seven diseases on actual field images, achieving a mean average precision (AP) of 0.66 and an accuracy of 0.89. Based on the diagnosis, practitioners could choose appropriate pesticides in plantation management. In addition, deep learning has been applied during the harvest stage to address the labor shortages. Lin and Chen (2019) developed a navigation system for tea harvesting machines using an efficient neural network (ENet), obtaining a mean accuracy of 0.94 and an inference time of 0.17s. Their study proposed an automatic method for harvesting machines.

Xu et al. (2022) constructed a You-Only-Look-Once v3 (YOLOv3) model to locate the position of tea buds. The model reached an accuracy of 0.95 and has the potential to pluck good quality tea shoot. Kamrul et al. (2020) applied the Faster R-CNN model to classify tea into five classes depending on its condition, with an accuracy of 0.96. These five classes described the shape, organs, maturity, damage, and completeness during the production process. Kamrul's study showed a practical method for tea producers to estimate the tea. Zhang et al. (2023) developed a quality identification model based on CNN and transfer learning, providing consumers with a practical method to examine the grade of made tea. The grades described the processing techniques and quality of the raw



materials. In their study, the model categorized made tea into three grades, reaching an accuracy of 0.93. These studies demonstrated the potential of deep learning in assessing the quality of plucked tea shoot and its practical application for tea producers and consumers.

### 2.3.2 Tea Quality Estimation

Color and texture are essential characteristics of tea that could be analyzed using image processing methods. Borah and Bhuyan (2003) developed a non-destructive method for measuring tea fermentation during production by applying color space conversion and histogram intersection techniques. Chen et al. (2022) designed an automatic controller to monitor the color change of tea and adjust the related parameters, such as hot air, steam, and pan-fire to optimize the manufacturing process.


For texture measurements, it refers to the size or the shape of the targets. Wu et al. (2007) utilized entropy to characterize the texture of plucked tea shoot and built a classifier to discriminate four categories of green tea with an accuracy of 0.97. Borah et al. (2007) developed a texture-based estimation method to classify eight types of tea in Crush, Tear, and Curl status, obtaining an accuracy of 0.80. Moreover, Zhu et al. (2017) combined color and texture extraction methods for tea sensory analysis. The extracted parameters described comprehensive information about texture and color. Based on that, a radial basis function kernel (RBF) model was used to predict sensory quality with a  $p$ -value of 0.60. The result showed significant potential of computer vision in describing visual sensory and confirmed its effectiveness in tea quality estimation.

## 2.4 Deep Learning

With the increment of computer speed and its advanced performance, deep learning (DL) provides an effective solution for various fields, including numerical analysis, natural language processing (NLP), and computer vision. DL models comprise multiple layers of neural networks that estimate the correlation between input and output by the weight of neurons (Krizhevsky et al., 2012). An optimal model could be established by training model with adequate data and parameters setting. During model training, the loss function and backpropagation are used to obtain the optimal model (Werbos, 1990). The loss function maps the error to a real number to evaluate the error, while backpropagation is a widely used algorithm to efficiently minimize a model's error by updating the model's weight.


The deep learning applications in computer vision relied on feature extraction to characterize numerical features based on an input image. To achieve higher performance in feature extraction, researchers have proposed various backbones. The visual geometry group (VGG) backbone is a well-known architecture for its convolution layers and the adoption of hidden layers (Simonyan and Zisserman., 2015). The implementation of hidden layers allows for the extraction of features with greater diversity. However, increasing hidden layers can lead to vanishing gradients, making the model difficult to train. To avoid this issue, He et al. (2016) proposed a residual neural network (ResNet) that skips redundant connections in hidden layers to preserve gradients. Inspired by ResNet, aggregated residual transformations (ResNeXt) were designed as a multi-branch architecture to reduce the computational complexity of the model (Xie et al., 2017).

### 2.4.1 Object Detection




Object detection, a sophisticated computer vision technique involving both recognition and localization, has become increasingly prominent with the widespread adoption of DL. Detection models could be broadly categorized into two-stage and one-stage models, based on the prediction procedure. In a two-stage model, a set of proposal boxes is generated in the images, and recognition is executed across all proposal boxes to determine if objects exist. Region proposal networks (RPNs) and their successor, the regional convolutional neural network (R-CNN), are representative examples of two-stage models (Girshick et al., 2014). In a one-stage model, such as the YOLO family, recognition and localization are performed in a single process (Redmon et al., 2016). In object detection, an object could be detected through detection boxes, pixels, or boundaries. For more detailed information about the object's area or boundary, semantic segmentation and instance segmentation are used. Semantic segmentation involves clustering every pixel in an image, while instance segmentation detects and delineates each distinct object. The Efficient neural network is a famous model for semantic segmentation, achieving an intersection over union (IoU) of 0.80 on Cityscapes (Paszke et al., 2017). For instance segmentation, Mask region-based convolutional neural network (Mask R-CNN) is a widely used model, with an mAP of 0.37 on COCO (He et al., 2017).

## 2.4.2 Transformer



Recently, the transformer model received great attention for its significant performance in NLP, such as speech recognition, question answering, and text categorization. Sequence data is one of the most common data types in NLP, where the characteristic of a sequence should be determined based on the relevance and order of its features. The transformer model uses position encoding and attention mechanisms to process sequence data (Vaswani et al., 2017). Position encoding is a method to describe the order of every feature in a sequence. The attention mechanism evaluates the relevance between the source feature and the target, enabling the model to utilize the most relevant features. The relevance is calculated using three representable matrices, Query( $Q$ ), Key( $K$ ), and Value( $V$ ). Each vector in  $Q$  presented a corresponding feature, and  $K$  presents every feature in a computable form. By calculating the dot product of  $Q$  and  $K$ , the model can determine the one-to-other-feature correlations. With  $QK^T$ , the transformer model could effectively utilize each feature from  $V$  based on its corresponding level of correlation.

The attention mechanisms focus on relevant parts of the input data while paying less attention to others, thus guiding the process of reasoning. This advantage benefited transformers highly suitable to be used in computer vision. Until then, CNN had been the essential architecture in computer vision, but the transformer model showed its competitiveness on several computer vision benchmarks (LeCun et al., 1998). The vision transformer (ViT) is the first pure transformer model for image classification, laying the foundation of transformer application in computer vision (Dosovitskiy et al., 2020). Unlike CNN, ViT splits and flattens an image into a sequence of patches to reduce



computational costs and performs the attention mechanism on the image. Inspired by ViT, data-efficient image transformers (DeiT) rely on a teacher-student strategy to ensure effective model training and achieve top-1 accuracy of 0.82 on ImageNet (Touvron et al., 2021). Additionally, Han et al. (2021) developed the transformer in transformer (TNT) to enhance the utilization of local patches. TNT further divides patches into smaller blocks, calculates the attention of each block in the given patch to maintain computation cost, and achieves a top-1 accuracy of 0.83 on ImageNet. In addition to image classification, transformer models have shown promising results in object detection. Detection transformer (DETR) introduced an anchor-free detection mode, providing a simple and extendable framework for object detection (Carion et al., 2020). Researchers could easily adapt DETR for different object detection tasks for its high flexibility. Since the introduction of DETR, several transformer models have been developed and benchmarked in object detection. Transformer-based set prediction with RCNN (TSP-RCNN) added alignment to suppress the instability problem, enhancing performance from 0.43 AP to 0.45 AP on COCO (Sun et al., 2021). Deformable DETR was proposed as a more efficient model, reaching an AP of 0.46 on COCO (Zhu et al., 2021).

## CHAPTER 3. MATERIALS AND METHODS

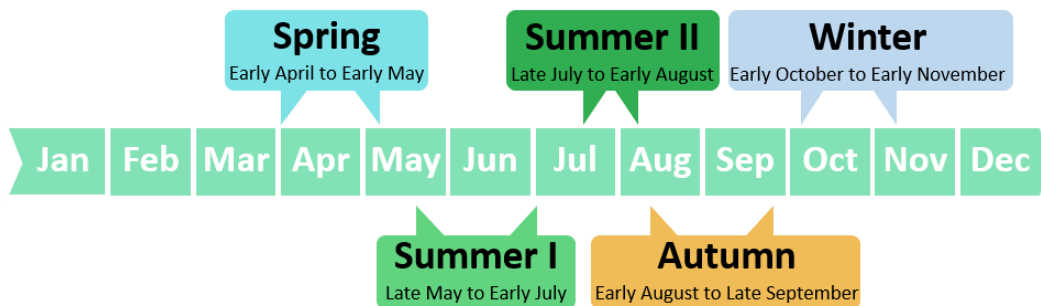


### 3.1 Data Acquisition

Tea shoot samples were from four cultivars, including “*Taiwan tea experiment station (TTES) No.1*”, “*TTES No.12*”, “*TTES No.17*”, and “*Chin-Shin-Dapan*”. These samples were collected between April 2020 and April 2023 from five harvest seasons: spring, summer I, summer II, autumn, and winter (Table 3.1). The five tea harvesting periods were defined by tea scientists throughout the year based on solar terms and the growth stage of the tea plants (Figure 3.1).

**Table 3.1** Number of tea shoots images and cultivars in five harvest seasons.

<b>Cultivar \ Season</b>	<b>Spring</b>	<b>Summer I</b>	<b>Summer II</b>	<b>Autumn</b>	<b>Winter</b>
TTES No.1	24	60	25	70	72
TTES No.12	30	40	24	75	32
TTES No.17	22	60	41	86	31
Chin-Shin-Dapan	34	0	16	57	26
<b>Total</b>	<b>825</b>				



**Figure 3.1** Five harvest seasons in a year.

These images were obtained from the tea plantations located in Yuchi Township, Nantou County, and eight tea factories in Taoyuan City and Hsinchu County, Taiwan. Each image consisted of 15 to 30 tea shoot samples, captured from a perspective view

using a digital camera or a smartphone. Images size ranged from 3280 x 2464 pixels to 5152 x 3864 pixels, with a resolution collected of at least seven megapixels. Two shooting backgrounds were used: a white sample board with grid patterns and a random floor background (Figure 3.2). Tea shoot samples were either neatly arranged or randomly scattered on both backgrounds. Samples placed on the sample board can be easily examined for their traits, while samples placed randomly on the floor provide higher complexity that enhanced the model's adaptability for various sample collection scenarios during the training stage.



**Figure 3.2** Examples of two different backgrounds.

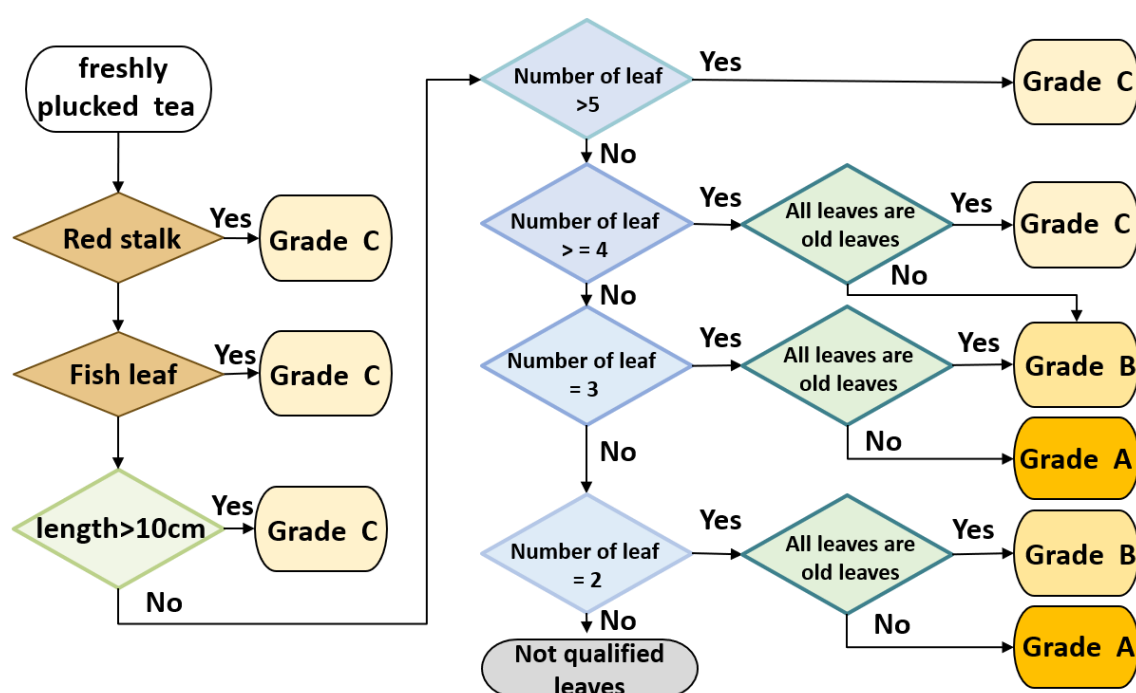
(a) Neatly arranged samples on a sample board. (b) samples randomly placed on floor.

### 3.2 Development of Grading Criteria

This study converted objective grading criteria into a decision tree, serving as a practical function to grade the tea shoot. The grading criteria were developed by experts from Tea Research and Extension Station (TRES), Executive Yuan, R.O.C.(Taiwan). They classified the tea shoot into three grades, A (excellent), B (ordinary), and C (defective) (Figure 3.3). The expert with extensive practical experience proposed the grading criteria by reviewing samples from four cultivars and several production seasons. Tea shoot meeting any of the following conditions would be classified as grade C:



containing red stalk or fish leaf, having a stem length longer than 10 cm, holding more than five leaves, or having more than four old leaves. While tea shoot with four or five leaves, and not all of them being old, were defined as grade B. Tea shoot with only two or three leaves, and all of the leaves were old, would also be defined as grade B. Grade A must have two or three leaves and not only old ones. Tea shoots containing only one leaf were regarded as disqualified samples and were not eligible for grading.



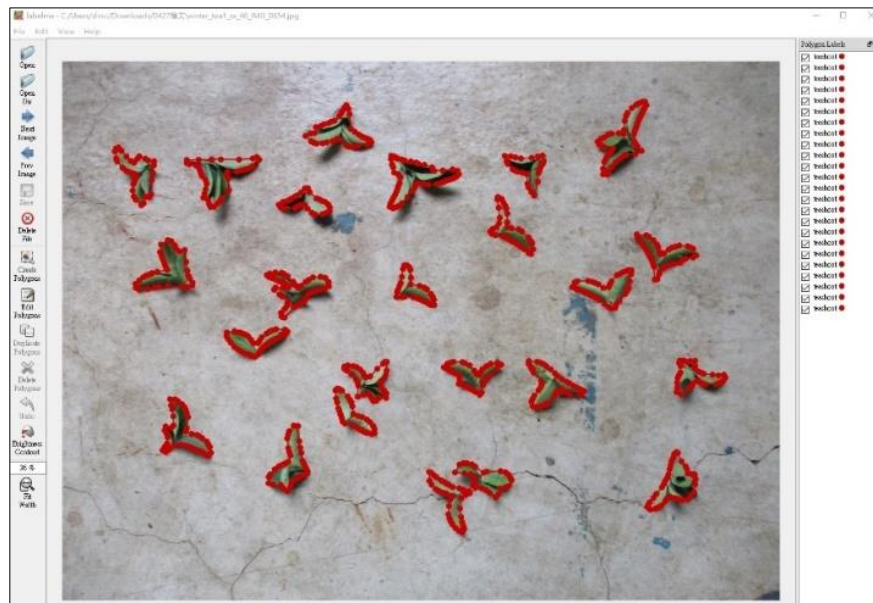
**Figure 3.3** Grading criteria for plucked tea shoot.



### 3.3 Datasets Preparation

#### 3.3.1 Single Tea Shoot Segmentation Dataset

The collected images were manually annotated to prove the ground truth, indicating the contours of the tea shoots. Labelme was used to perform sample annotation (Wada, 2016; Figure 3.4). In total, 825 images were annotated, with 784 images used as the training set and the remaining 41 images used as the testing set. The annotated training and testing sets included 13,680 and 770 tea shoot samples, respectively (Table 3.2).



**Figure 3.4** Tea shoot annotation in labelme.

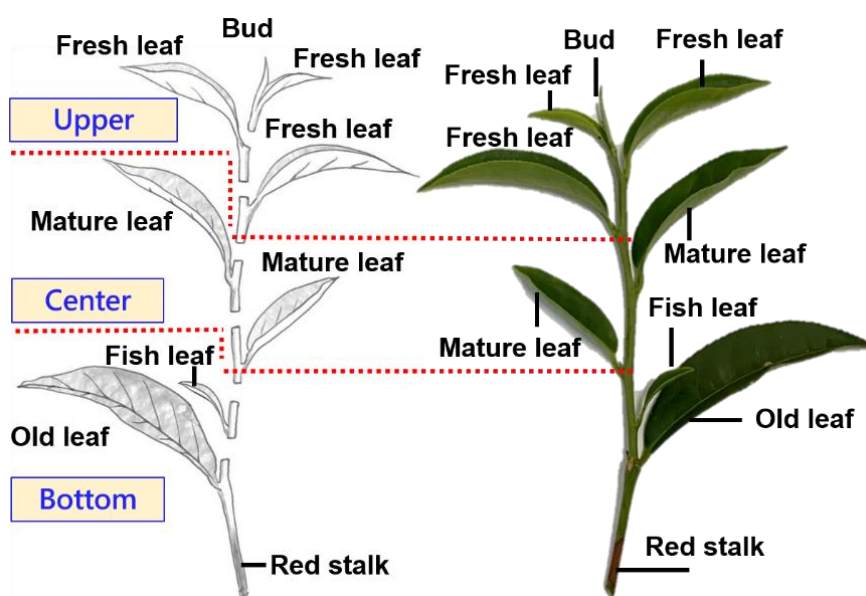
**Table 3.2** Number of images of single tea shoot segmentation dataset.

Dataset \ Season	Spring	Summer I	Summer II	Autumn	Winter
<b>Training</b>					
Image	103	154	100	282	145
Tea shoot	1951	2418	1688	5147	2476
<b>Testing</b>					
Image	7	6	6	6	16
Tea shoot	143	109	121	128	269



### 3.3.2 Organ Identification Dataset

Identifying organs of tea shoots was critical as they serve as the key features for grading. Single tea shoot images were cropped from the previous single tea shoot segmentation dataset in section 3.3.1, and their organs were further annotated to create the dataset for the organ identification model. The annotation included seven types of classes: bud, fresh, mature, old, fish leaf, red stalk, and stem. The growth characteristics of organs, including the timing and relative position of their emergence, as well as changes in tissue organization, are all incorporated as reference criteria during annotation (Figure 3.5). For example, buds were a distinctive feature of growing tea plants and were typically found at the top of tea shoots. Tea plants grew from the ground and exhibited an upward arrangement of old, mature, and fresh leaves. Red stalks and fish leaves typically appear at the lower part of tea plants as traits of the senescent tea plants. Hence, the position information between organs was quite important in this dataset. A total of 1337 single tea shoot images were annotated for organ identification (Table 3.2).



**Figure 3.5** Organs of tea shoots.

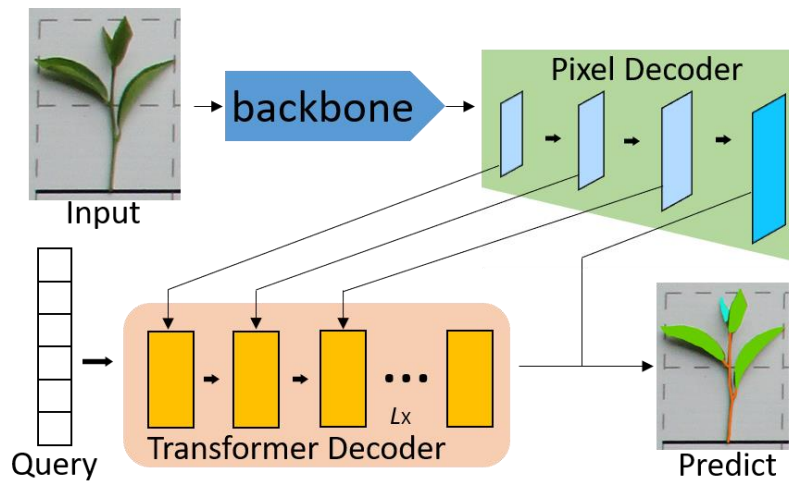
**Table 3.3** Number of instances of organ identification.

<b>Dataset \ Class</b>	<b>Bud</b>	<b>Fresh leaf</b>	<b>Mature leaf</b>	<b>Old leaf</b>	<b>Stem</b>	<b>Fish leaf</b>	<b>Red stalk</b>
<b>Training</b>	277	1019	1129	497	822	73	19
<b>Testing</b>	54	165	227	89	158	19	12

### 3.4 Models Architecture

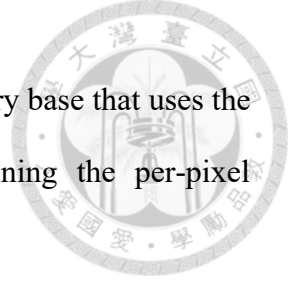
#### 3.4.1 Masked-attention Mask Transformer

TSGS applied masked-attention mask transformer (Mask2Former) to segment tea shoots and their organs from captured images (Cheng et al., 2022). Mask2Former is a transformer-based instance segmentation model. This model consists of a feature extraction backbone, a pixel-decoder, and a transformer-decoder. (Figure 3.6).



**Figure 3.6** The architecture of Mask2Former.

In the first stage, the backbone processes the input images for feature map extraction. After the backbone, the pixel-decoder generates high-resolution per-pixel embeddings by calculating attention across multiple scales of the feature map. Sinusoidal positional embeddings and scale-level embeddings are also applied to the per-pixel embeddings to present the spatial information and the scaling factor, respectively. Lastly, the



transformer-decoder calculates the probability of each class on a query base that uses the embeddings. The segmented image can be predicted by combining the per-pixel embeddings and query.

Mask2Former adopts a multi-scale strategy and mask-attention to enhance model performance. The mask-attention is a specialized mechanism to emphasize the foreground region and suppress the background interference. In masked attention, the area  $M_{l-1}$  containing detected target is considered as foreground, and the value of area  $M_{l-1}$  is set to zero (Eq. 3.1 and Eq. 3.2).

$$M_{l-1}(x, y) = \begin{cases} 0 & \text{if } M_{l-1}(x, y) = 1 \\ -\infty & \text{otherwise} \end{cases} \quad (3.1)$$

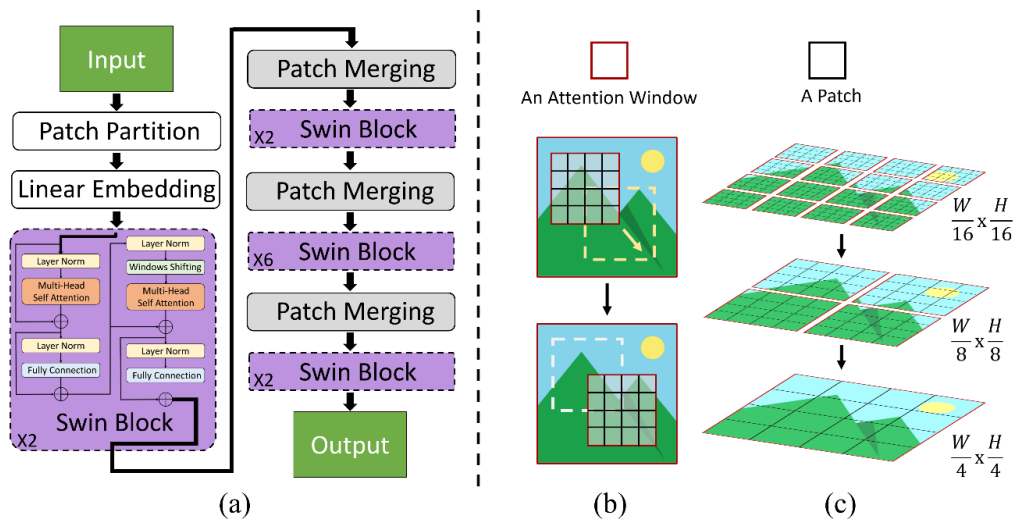
$$X_i = \text{softmax}(M_{l-1} + Q_l K_l^T) V_l + X_{l-1} \quad (3.2)$$

In contrast, if there are no targets in area  $M_{l-1}$ , its value is set to minus infinity, thereby reducing the multiplication of query and key  $Q_l K_l^T$  to zero. Therefore, area  $M_{l-1}$  does not contribute to evaluating value  $X_i$ , meaning the background area is completely neglected during attention calculation. A model trained with mask-attention could focus more on the foreground and its objects. Additionally, the multi-scale of feature map in the pixel-decoder could represent objects with a wide range of scaling, enhancing the model's ability to detect objects of different sizes.



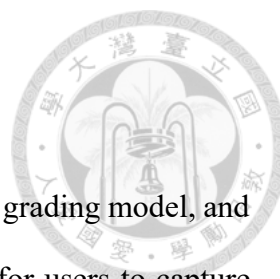
### 3.4.2 Shifted Windows Transformer

Shifted windows transformer (Swin) was applied for feature extraction from tea shoot images in TSGS (Liu et al., 2021). Swin is a pure transformer-based feature extracting network that adapts shifted windows and patch merging to address the high variation of scaling and the huge number of pixels (Figure 3.7). Swin splits images into patches with a fixed size, then partitioned the image with a set of non-overlapping windows. The attention mechanism is computed using patches located within the same window. Additionally, windows are shifted between consecutive attention layers, bridging the windows of the preceding layer, thus providing connections among patches located in different windows. Patch merging merges adjacent patches, gradually increasing their size. This process preserves the model's flexibility in detecting objects of various scales while ensuring feasible computational complexity. Swin has presented higher efficiency and more robust performance in several benchmarks, e.g., COCO, Cityscapes, ADE20K. This demonstrates its effectiveness as a powerful backbone for feature extraction.



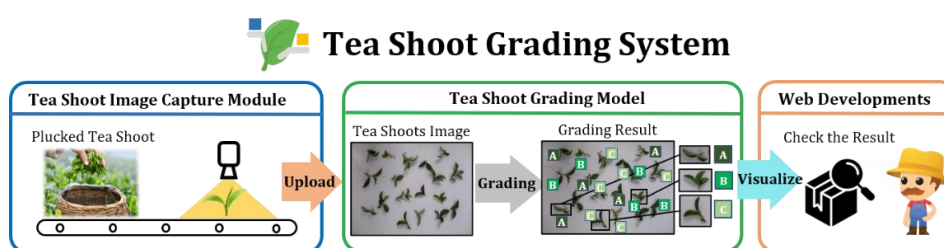
**Figure 3.7** The architecture of Swin transformer.

(a) Architecture of Swin transformer. (b) Shifted windows. (c) Patch merging.



### 3.5 Development of TSGS

TSGS consisted of a tea shoot image capture module, a tea shoot grading model, and web developments (Figure 3.8). The capture module was designed for users to capture images of purchased tea shoots and upload them to the grading model. Once uploaded, the tea shoot grading model processed the images and graded each tea shoot based on the quality of the tea shoot. The grading results and detailed information were then displayed on web developments, providing users with a convenient way to observe the grading results and obtain detailed identification information. The primary objective of the proposed TSGS was to offer a faster and more accurate grading method. Therefore, two performance evaluators were discussed: processing speed and grading accuracy. Both evaluators were used to compare the speed and accuracy of TSGS with manual grading.



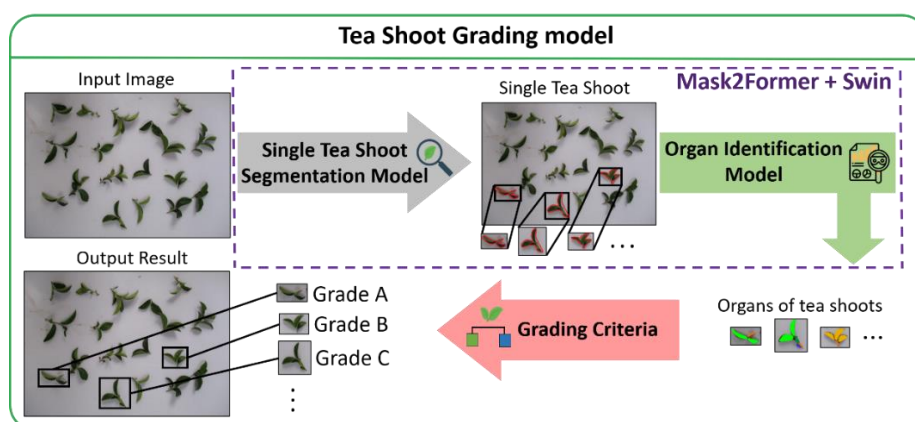
**Figure 3.8** The architecture of TSGS.

#### 3.5.1 Tea Shoot Grading Model

As the core of TSGS, the tea shoot grading model extracted the location of each single tea shoot from the input image, identified its organs, and determined its grade (Figure 3.9). The tea shoot grading model was built on Python 3.6 and NVIDIA GeForce RTX 3090 (NVIDIA Corporation; Santa Clara, CA, USA). The grading model consisted of a single tea shoot segmentation model, an organ identification model and grading criteria. In the first stage, the single tea shoot segmentation model extracted each tea shoot based



on its contours. This model was based on Mask2Former, training with the single tea shoot segmentation dataset. Since the detection method was instance segmentation, the model could separately crop the tea shoot according to their contours without cropping neighboring objects. While instance segmentation required a higher computational complexity than detection using bounding boxes, it prevented errors in counting the number of leaves. A lightweight Swin-tiny was applied for feature extraction to balance computational complexity and image resolution. After the image was processed in the first stage, the organ identification model was used to identify seven types of organs from each extracted tea shoot. This model was established using Mask2Former and the organ identification dataset. However, tea shoots are not neatly arranged but randomly scattered in the production process. Data augmentation of random rotation was employed during model training to address this issue. This augmentation method randomly rotated each tea shoot positioning from 0° to 360° to simulate various degrees of placement. Furthermore, given the high correlation between organs and position information, this method could enhance the model's performance in identifying organs from various orientations. After previous models identified tea shoots, the grading criteria could classify each tea shoot into grades A, B, and C according to its organs.

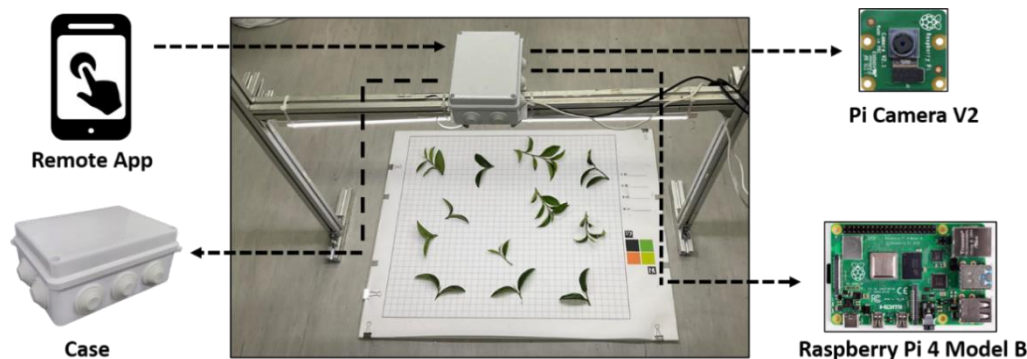


**Figure 3.9** The architecture of tea shoot grading model.



### 3.5.2 Design of Tea Shoot Image Capture Module

The tea shoot image capture module was designed for collecting and uploading the images of tea shoot. This module consisted of an RPI4 (Raspberry Pi 4 Model B; Raspberry Pi Foundation; Cambridge, UK), a camera (Raspberry Pi Camera V2; Sony; Tokyo, Japan), a remote app and a case (Figure 3.10). RPI4 was a small single-board computer, allowing developers to build compatible internet of things (IoTs) devices. Additionally, RPI4 was equipped with both Wi-Fi and Bluetooth, so it was suitable for field device development. The camera was assembled for image acquisition, and it was a camera with 3280 x 2464 resolution and was specialized for the raspberry pi series. The remote app was a Bluetooth controller software, allowing the user to conveniently control the capture module by a smartphone.

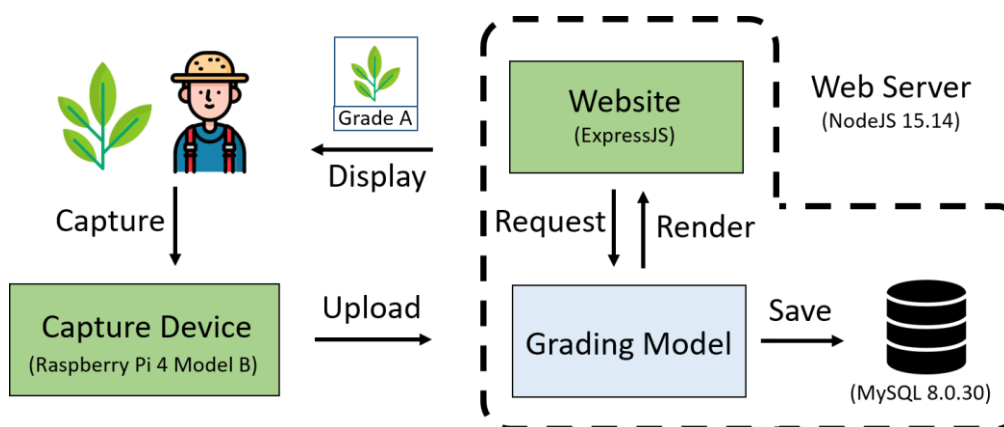


**Figure 3.10** Components for capture module.

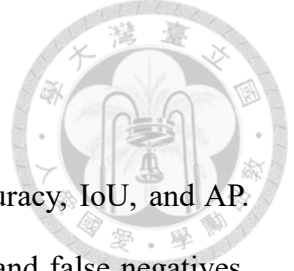


### 3.5.3 Design of Web Developments

The web developments of TSGS for result visualization and data storage was developed using ExpressJS framework and MySQL, running on the NodeJS platform (Figure 3.11). ExpressJS was a popular framework for web development in NodeJS, allowing dynamic responses for web pages and realizing fast communication with the database. NodeJS applied a single-thread design to manage tasks efficiently. Because NodeJS could handle numerous tasks without excessive resource overhead, this platform was suitable for the huge computational task. A database for data storage and access was essential for web development to provide comprehensive functionality. MySQL was a powerful database for storing and retrieving data efficiently and securely. This system could store the tea shoot images and their corresponding grading results, as well as retrieve and display these images and results on a webpage.

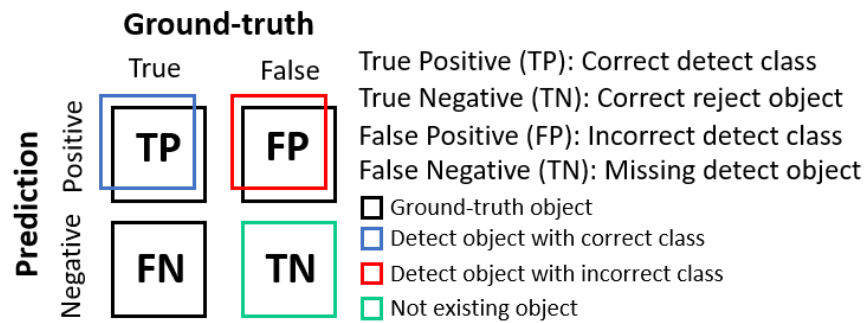


**Figure 3.11** Web developments architecture diagram of TSGS.



### 3.6 Evaluation Metrics

The main metrics for model performance evaluation were accuracy, IoU, and AP. Accuracy considered true positives, false positives, true negatives, and false negatives, giving the percentage of correct classifications (Figure 3.12) (Eq. 3.2). The IoU was given by the ratio of the intersection area between the prediction segment and the ground truth segment to the union area of the two segments, describing the localization accuracy (Eq. 3.3). The judgement of a model's predicting localization could be determined by IoU thresholding. Predictions with an IoU over the threshold were considered to be correct. With the combination of IoU, AP (Eq. 3.4) family were comprehensive metrics used to measure the performance of models on localization and classification by calculating the integral of precision (Eq. 3.5) and recall (Eq. 3.6). In the AP family, AP at IoU=0.5 (AP50), AP at IoU=0.75 (AP75), and mean AP (mAP) were three of the most popular metrics. AP50 and AP75 indicated the performance of predictions with an IoU threshold of 0.5 and 0.75, respectively. Additionally, mAP was calculated by averaging the AP values obtained at IoU thresholds ranging from 0.5 to 0.95.



**Figure 3.12** Definition of confusion matrix.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (3.2)$$

$$IoU = \frac{\text{area of intersection}}{\text{area of union}} \quad (3.3)$$

$$AP = \int PdR \quad (3.4)$$

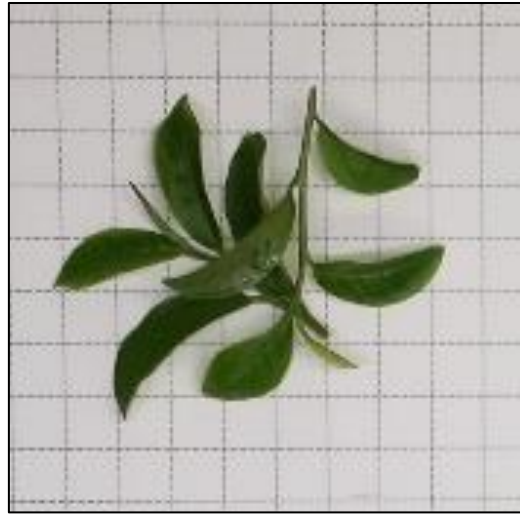
$$P = \frac{TP}{TP + FP} \quad (3.5)$$

$$R = \frac{TP}{TP + FN} \quad (3.6)$$



### 3.7 Tea Leaf Overlap Ratio Test

In the actual process of the tea factory, the tea shoot possibly overlapped with each other in images, and the overlapping case could negatively affect the recognition performance of models. Thus, it was necessary to estimate the model performance under different levels of overlap. In total, 414 images of randomly overlapping tea shoots were collected to evaluate the model performance in overlapping cases (Figure 3.12). The collected images contained 93, 104, 122, and 95 images of “TTES No.1”, “TTES No.12”, “TTES No.17”, and “Chin-Shin-Dapan”, respectively. Besides, an overlapping experiment was conducted to define overlap ratio that expresses the level of overlap (Algorithm 1). The overlapping experiment was carried out as follows: Each tea shoot was selected and paired with another tea shoot. The distance between paired tea shoots was gradually reduced and circled along with 0° to 360° until the distance was zero. As the distance decreased, the IoU of the paired tea shoot was continuously recorded. The distribution of overlap could be quantified based on the recorded IoU values, and these values could be used as the basis for defining the overlap ratio.



**Figure 3.13** An example of the overlapped tea shoots.

---

**Algorithm 1 - Overlapping tea shoots simulation**

---

**Input:** 702 tea shoot images  $T$   
**Output:** IoU distribution  $N$

```

1   $N, n$ : array[100]  $\leftarrow \{0\}$ ,  $P$ : image
2  for each  $a, b \in T$ 
3       $D \leftarrow \text{diagonal}(a + b) / 2$ 
4      Initialize ( $n$ )
5      for  $\theta \leftarrow 0$  to  $2\pi$ 
6          for  $d \leftarrow D$  to 0
7              Initialize ( $P$ )
8               $x = d \cos \theta, y = d \sin \theta$ 
9              paste  $a$  on  $P(0, 0)$ 
10             paste  $b$  on  $P(x, y)$ 
11              $n[\text{IoU}(P)] \leftarrow 1$ 
12         end for
13     end for
14      $N \leftarrow N + n$ 
15 end for
16 return  $N$ 

```

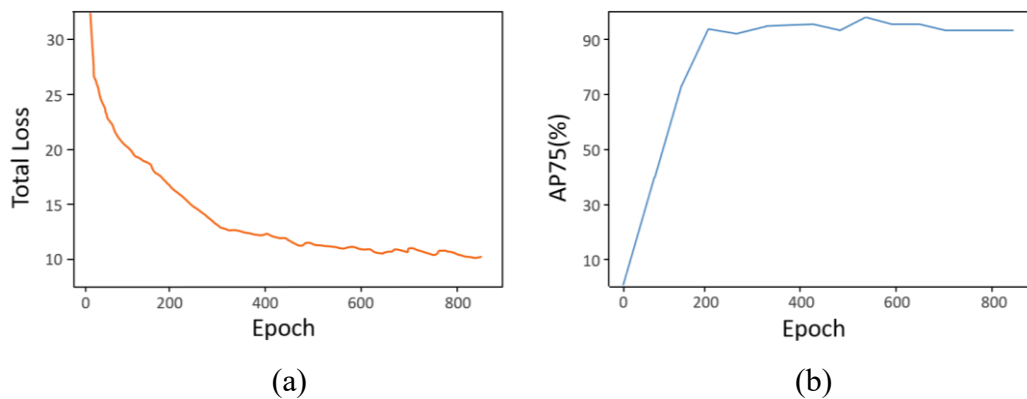
---

## CHAPTER 4. RESULTS AND DISCUSSION



### 4.1 Single Tea Shoot Segmentation Model

The single tea shoot segmentation model was developed based on Mask2Former architecture. Swin-tiny was applied to provide a relatively compact backbone structure for the model. The model performance was evaluated during the training process. Training loss achieved convergence at the 500 epoch (Figure 4.1).



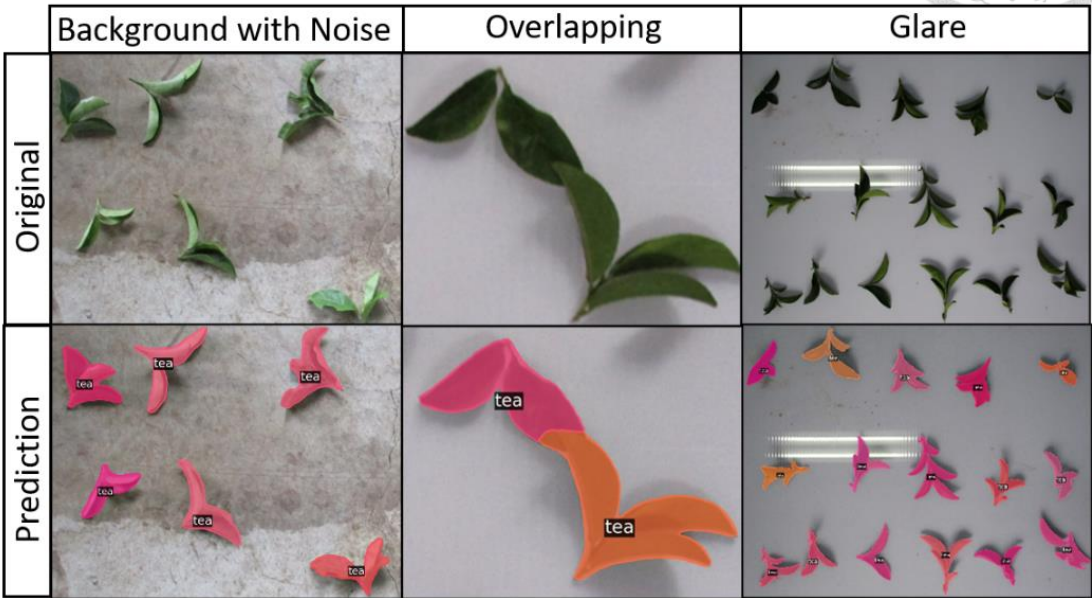
**Figure 4.1** Training curves of single tea shoot segmentation model.

(a) Loss. (b) AP75.

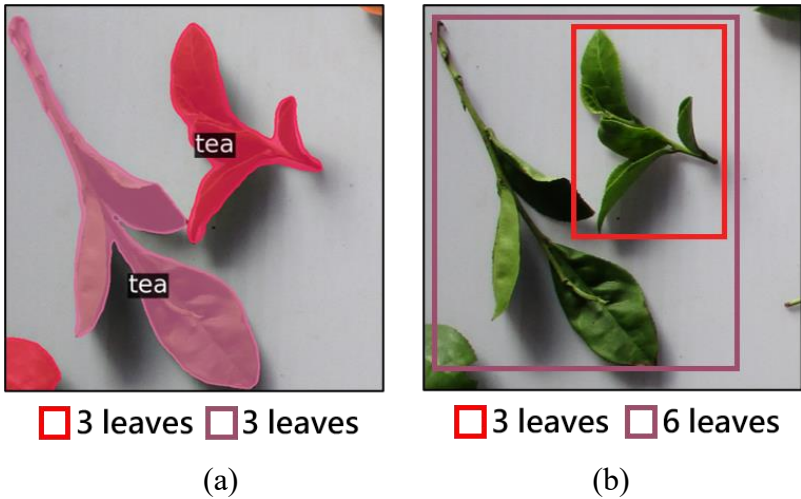
The finalized model reached AP50 of 0.99, AP75 of 0.98, and mAP of 0.88, respectively, in the testing dataset. The AP75 of 0.98 indicated the model could extract tea shoots with high accuracy. According to the demonstration images, the model successfully segmented the tea shoots in different shooting scenarios, e.g., dirty background, overlapping tea shoots, and glare case (Figure 4.2). Even for two very closely related adjacent tea shoots, this model can still identify them successfully. For example, when using detection box to identify tea shoots, adjacent tea shoots could be within the detection box along with the target tea shoot. These adjacent tea shoots could lead to errors in leaf number calculation or the misidentification of organs. In contrast, our model



identifies tea shoots individually based on their contours, thereby avoiding the impact of adjacent tea shoots (Figure 4.3).



**Figure 4.2** Successful tea shoot segmentation.

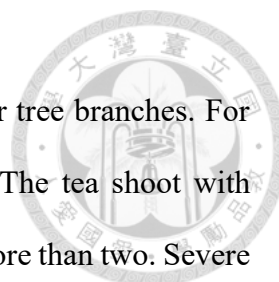


**Figure 4.3** Leaves calculation of tea shoot segmentation model.







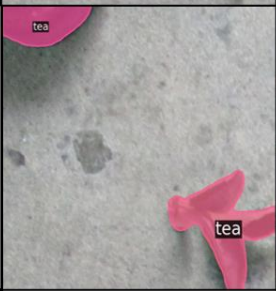


(a) Our instance segmentation model. (b) Traditional object detection model.

Although the model could correctly identify leaves in most cases, there were still some error cases. These errors could be categorized into three types, background noises, excessive length and severe overlap. In background noises, the model possibly



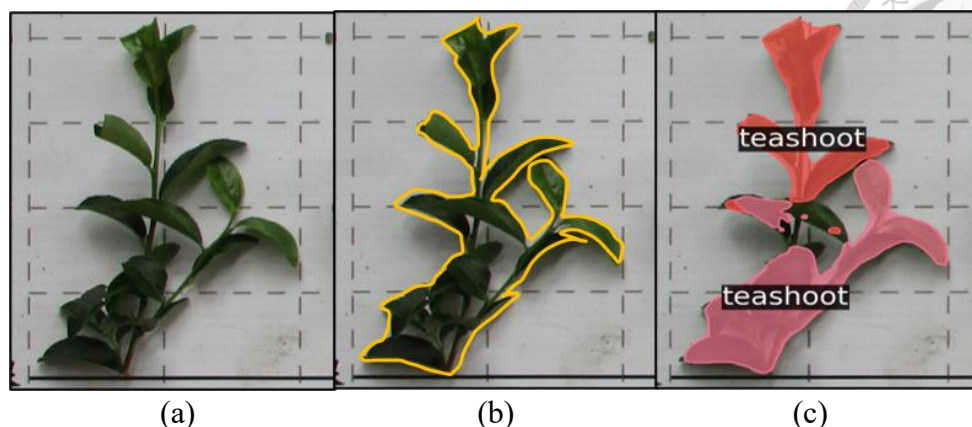


misidentified objects other than tea shoots, such as stones, leaves, or tree branches. For example, the model identified a shoe as a tea shoot (Figure 4.4). The tea shoot with excessive length could cause the model to identify one tea shoot as more than two. Severe overlap occurred when the leaves of two tea shoots are placed very close to each other, leading it difficult to correctly distinguish them. Additionally, overgrowth tea shoots tended to possess long stems and numerous leaves, leading to cases of both excessive length and severe overlap (Figure 4.5). Consequently, the model was more challenging to identify overgrowth tea shoots. These error cases potentially decreased the overall grading accuracy of the TSGS.

	Background noise	Excessive length	Severe overlap
Original			
Ground truth			
Prediction			

**Figure 4.4** Error cases of tea shoot segmentation.



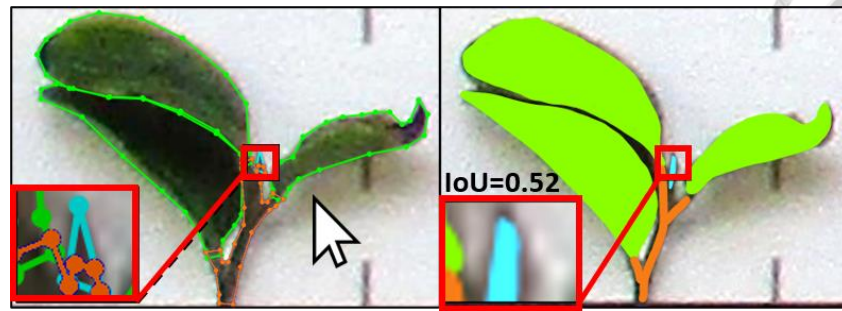


**Figure 4.5** Incorrectly identification of overgrowth tea shoot.  
 (a) Original image. (b) Ground truth. (c) Model prediction.

## 4.2 Organ Identification Model Performance

### 4.2.1 Metrics in Large Object and Small Object

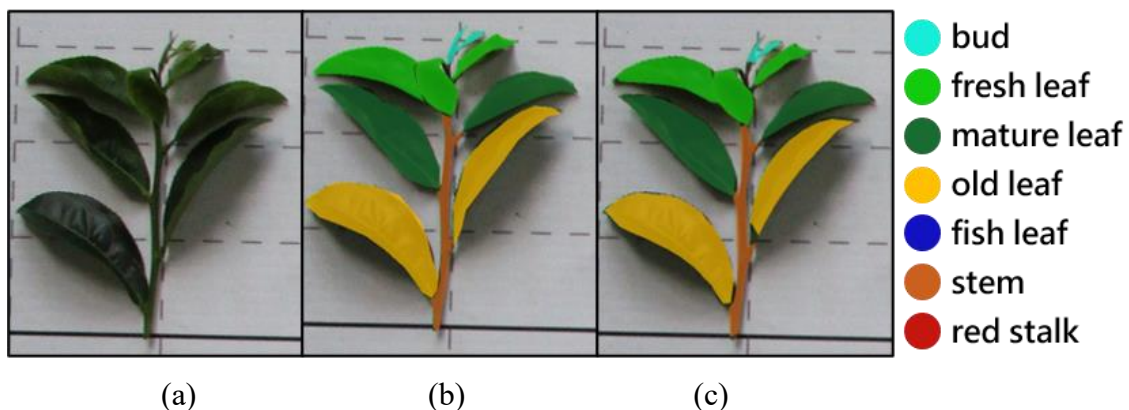
The identified organs could be categorized as large and small instances based on their size. Large organs included fresh, mature, and old leaves, while small organs included fish leaves, stems, red stalks, and buds. In small organs annotations, complete and precise annotation was very difficult due to their small size, resulting in inevitable difference between the annotated organ and the actual organ. This difference in terms of IoU led to the AP not accurately representing the actual performance of the model. For example, the model correctly identified the small tea bud in the image (Figure 4.6). However, when calculating its performance, the small bud led to pixel blurring, resulting in an IoU of only 0.52 and a decrease in AP. To minimize the impact of the gap between annotations and predictions, AP50 was utilized to estimate the organ identification model in small organ detection. Conversely, mAP was preserved to evaluate the performance of the model in large organ detection.



**Figure 4.6** Small organ annotation and prediction.  
(a) Manual annotation. (b) Model prediction.

#### 4.2.2 Backbone Substitution

Backbone substitution was implemented to achieve higher performance by applying different feature extraction models. Four trained model with different backbones, including Swin-base, Swin-small, Swin-tiny and Resnet-50 were compared. Swin-base achieved the best performance for large organ identification with mAP of 0.60, 0.71 and 0.53 for fresh leaf, mature leaf and old leaf, respectively; small organ identification with AP50 of 0.89, 0.51, 0.65 and 0.48 for stem, bud, fish leaf and red stalk respectively (Figure 4.7 and Table 4.1).

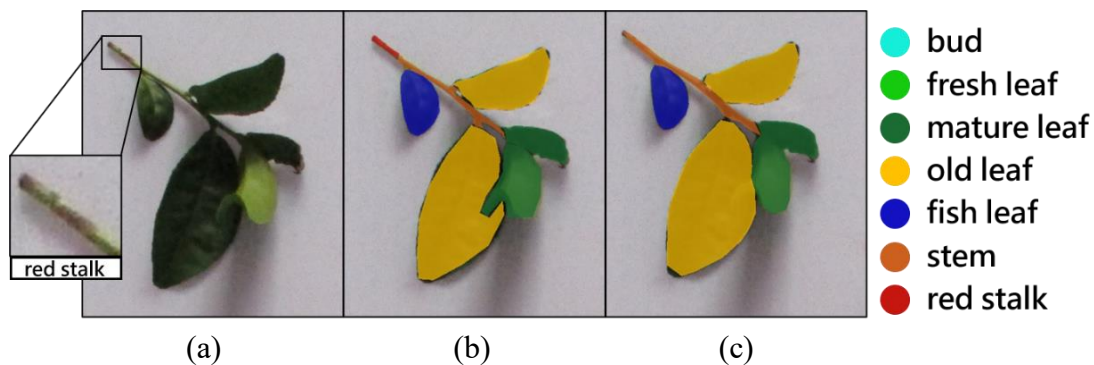


**Figure 4.7** Successful organ identification.  
(a) Original image. (b) Ground truth. (c) Model prediction.

**Table 4.1** Performance of organ identification model.

Backbone \ Organ	Large Organ			Small Organ			
	Fresh leaf	Mature leaf	Old leaf	Bud	Stem	Fish leaf	Red stalk
<b>Resnet-50</b>	0.53	0.59	0.50	0.39	0.86	0.28	0.22
<b>Swin-tiny</b>	0.61	0.64	0.46	0.33	0.86	0.48	0.28
<b>Swin-small</b>	0.63	0.71	0.61	0.45	0.91	0.42	0.24
<b>Swin-Base</b>	0.60	0.71	0.53	<b>0.51</b>	0.86	<b>0.65</b>	<b>0.48</b>

The color of the red stalk gradually turned darker red, making it difficult for the model to identify those red stalks undergoing this color transformation (Figure 4.8). Furthermore, the number of red stalks was lesser, leading to lower performance. In comparison, the overall performance of Swin family were higher than the Resnet-50, indicating that transformer-based backbones were more suitable in organ identification. In small organs, the Swin-base outperformed the Swin-small and Swin-tiny, showing that small organs, including red stalk, fish leaf, and bud, required a deeper backbone to be identified. Previous studies also discussed the appropriateness of utilizing deeper backbones for detecting small objects. A deeper backbone could generate feature maps at multiple resolutions, thereby extracting finer details of feature from small organs (Szegedy et al., 2015).

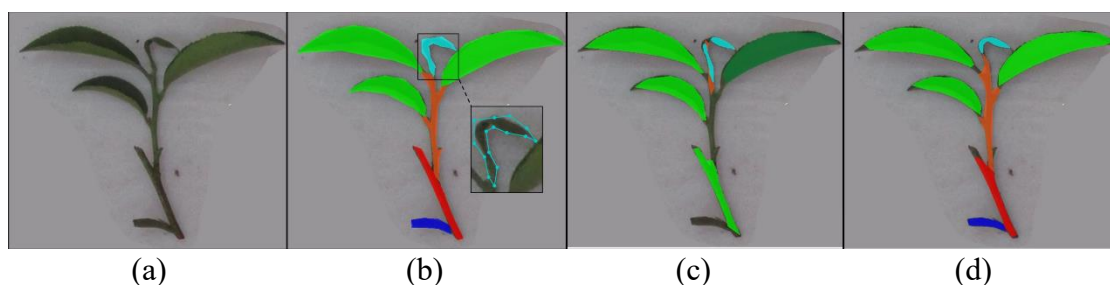


**Figure 4.8** Error case of red stalk identification.  
 (a) Original image. (b) Ground truth. (c) Model prediction.



### 4.2.3 Data Augmentation

Data augmentation was expected to enhance the performance of the organ identification model in identifying tea shoots from various angles. However, all tea shoots were positioned upright in the testing set, making it difficult to estimate the performance improvement of data augmentation on the model. Therefore, each tea shoot in the testing set was randomly rotated by  $90^\circ$ ,  $180^\circ$ , and  $270^\circ$  to evaluate performance improvement. Additionally, an augmented model and a base model were trained and compared to evaluate the effectiveness of data augmentation, where the augmented model was trained with data augmentation, and the base model was not. In the demonstration, the base model incorrectly classified a fresh leaf as mature and misidentified the stem, fish leaf, and red stalk (Figure 4.9). In contrast, the augmented model accurately classified the maturity of all leaves and successfully identified the stem, fish leaf, and red stalk.



**Figure 4.9** Model predictions in rotated case.

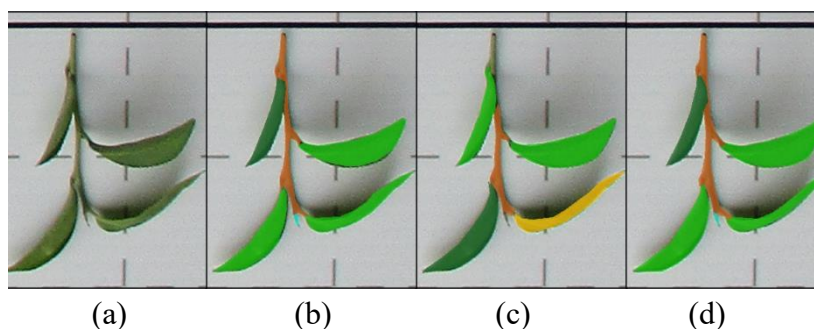
(a) Original image. (b) Ground truth. (c) Base model. (d) Augmented Model.

In the AP comparison of large organs, the base model obtained AP of 0.28, 0.37, and 0.06 for fresh leaf, mature leaf, and old leaf, respectively, while the augmented model achieved AP of 0.54, 0.65, and 0.53 (Table 4.2). For small organs, the base model obtained AP of 0.05, 0.34, 0.09, and 0.00 for bud, stem, fish leaf, and red stalk, respectively, while the augmented model achieved AP of 0.30, 0.87, 0.51, and 0.43. When inputting an inverted tea shoot image, the base model misclassified the leaves located in the lower

position of the tea shoot as fresh leaves and the leaves located in the upper position as old leaves, while the augmented model could correctly identify the types of leaves (Figure 4.10). The results demonstrated that randomly rotated data augmentation effectively improved the model performance in identifying organs from various orientations.

**Table 4.2** Performance of augmented model and base model.

Backbone \ Organ	Large Organ			Small Organ			
	Fresh leaf	Mature leaf	Old leaf	Bud	Stem	Fish leaf	Red stalk
<b>Base model</b>	0.28	0.37	0.06	0.05	0.34	0.09	0.00
<b>Augmented model</b>	<b>0.54</b>	<b>0.65</b>	<b>0.53</b>	<b>0.30</b>	<b>0.87</b>	<b>0.51</b>	<b>0.43</b>



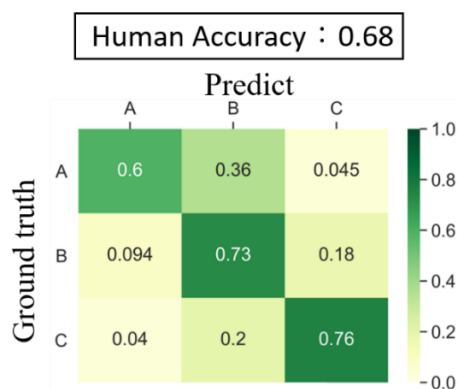
**Figure 4.10** Identification of augmented model and base model.

(a) Original image. (b) Ground truth. (c) Base model. (d) Augmented model.

### 4.3 TSGS Performance

The performance of TSGS was evaluated from two perspectives: processing speed and grading accuracy. Processing speed could be estimated by calculating the execution time required on the images. The grading accuracy of TSGS could be compared with the accuracy of manual grading. However, to evaluate the accuracy of the TSGS, additional tea shoot images with known grades were needed to refer to as ground truth. Therefore, additional 30 tea shoot batch images, containing 704 tea shoots in total, were collected individually. These 30 images were separated from the training and testing sets, and collected for a more objective evaluation of the accuracy of tea shoot grading. The

collected tea shoots were then graded based on the established grading criteria to create a grading dataset (Figure 3.3). This dataset contained 267, 235, and 202 tea shoots of grade A, B, and C, respectively. A thorough examination of organs was conducted, then graded each tea shoot based on the grading criteria as ground truth to ensure the correctness of the grades of the ground truth. In order to compare the TSGS and manual grading, each tea shoot would be graded by TSGS or manual grading by experienced personnel. Manual grading was conducted to simulate the tea shoot grading process in the tea factory. The experts from TRES with over six years of experience in tea research helped perform manual grading on the 30 tea shoot batch images. The experts were familiar with the grading criteria and memorized that in mind in general. To perform tea shoot grading, they looked over the tea shoots and determined the grade as soon as possible. Based on the record, the average processing time for one tea shoot batch image took approximately 40 s. Because manual grading did not completely examine the organs of each tea shoot, and the grading process relied more on experience rather than strictly following the grading criteria, there could be wrong grades. Therefore, a comparison with the ground truth was necessary for evaluating the grading accuracy of manual grading. Manual grading attained a grading accuracy of 0.68 (Figure 4.11). This result could be used to compare the accuracy of manual grading with TSGS.



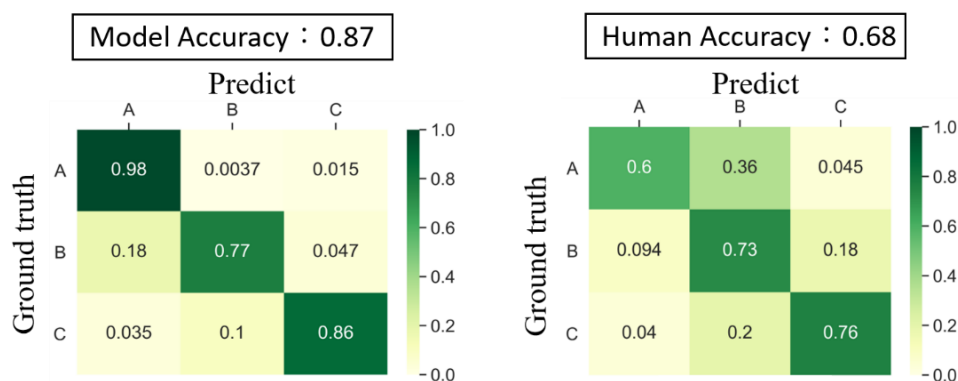
**Figure 4.11** The confusion matrix of manual grading.



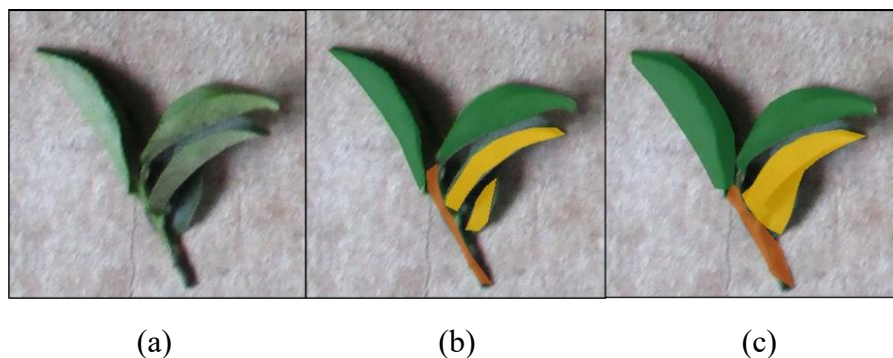


### 4.3.1 Grading Accuracy

In the evaluation of grading accuracy, TSGS achieved an overall grading accuracy of 0.87 on the grading dataset (Figure 4.12). The grading accuracy for Grades A, B, and C were 0.98, 0.77, and 0.86, respectively. The most common error made by the TSGS was misclassifying Grade B as Grade A for a misclassification rate of 0.18. Grading criteria classified Grade A and Grade B relied more on the leaf number of the tea shoot, but TSGS tended to miscalculate leaves due to the leaves in a tea shoot being closely positioned (Figure 4.13). These closely positioned leaves could be identified as a single leaf, leading to an underestimation of leaf number and contributing to grading inaccuracies.

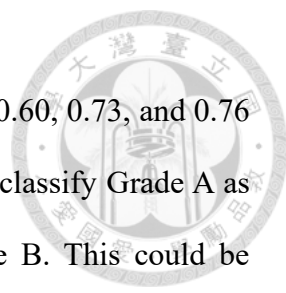


**Figure 4.12** Confusion matrix between model and human.



**Figure 4.13** Inaccurate leaf number calculations.

(a) Original image. (b) Truth: four leaves. (c) Model: three leaves.



The accuracy of manual grading was 0.68, giving accuracies of 0.60, 0.73, and 0.76 for Grades A, B, and C, respectively. Manual grading tended to misclassify Grade A as Grade B, with 0.36 of Grade A tea shoots being graded as Grade B. This could be attributed to the subjective inwardness of human judgment, leaning toward a stricter standard during the grading process. TSGS demonstrated a higher accuracy of 0.98 in Grade A compared to manual grading of 0.60, indicating a more reliable correctness. Additionally, the accuracy of TSGS in Grade C was 0.10 higher than manual grading. This improvement could be attributed to the better performance of TSGS in identifying small organs such as red stalks and fish leaves, while these organs were difficult to find with the naked eye. In conclusion, TSGS presented a higher accuracy of 0.87 compared to the manual grading accuracy of 0.68. The results suggested this system achieved better grading performance than the manual way.

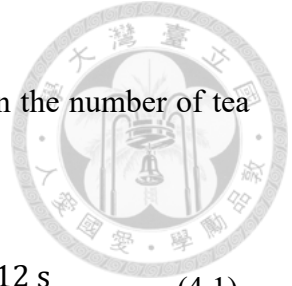
#### 4.3.2 Processing Speed

The grading process of TSGS involved tea shoot extraction, organ identification, the decision tree of the grading criteria, and the processing speed of each image was correlated with the number of tea shoots contained within the image. Thus, the actual processing time could be calculated by summing up the runtimes of tea shoot extraction, organ identification, and the decision tree of the grading criteria (Eq. 4.1). According to the performance evaluation based on the dataset, the tea shoot extraction spent 1.8 s to process an image. For the organ identification, it took 0.12 s to process one tea shoot. The number of tea shoots in a batch image was ranged from 15 to 30 generally. The execution time of the decision tree component was less than 1 ms and considered negligible in the calculation. Therefore, the actual execution time of an image is between 3.6 s to 5.4 s. It



could be concluded that the execution time of an image depended on the number of tea shoots, with more tea shoots requiring a longer execution time.

$$\text{runtime(s)/image} = 1.8 \text{ s} + (\text{number of tea shoots}) \times 0.12 \text{ s} \quad (4.1)$$



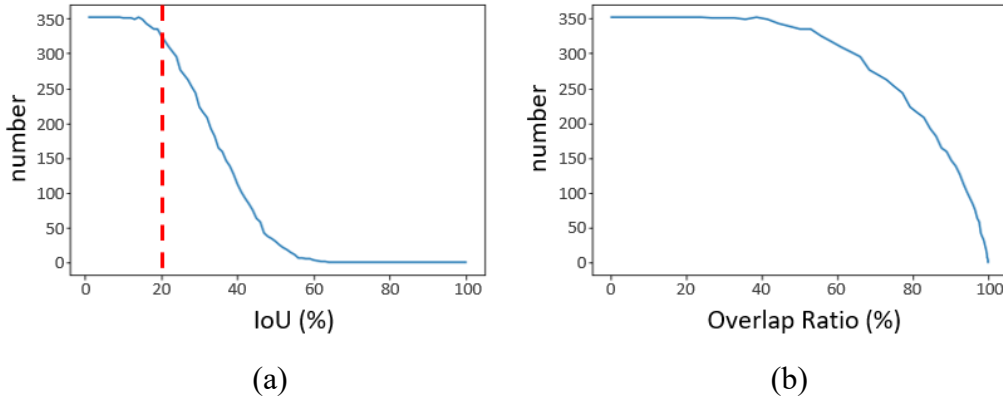
## 4.4 Discussion: Overlapping Leaves

### 4.4.1 Evaluation of Overlap Ratio Test

In section 4.1, it was observed that the leaves of overlapping tea shoots increased the difficulty of identification. However, in a randomly arranged scenario, overlapping becomes an unavoidable situation. Therefore, this study aimed to examine the performance of TSGS in identifying overlapping tea shoots and further proposed the acceptable level of overlap for TSGS, along with the corresponding grading accuracy. In order to evaluate the accuracy of TSGS in overlapping, 414 images of overlapping tea shoots were collected. Each image contained two randomly overlapped tea shoots. Besides, an overlapping experiment was carried out for overlap distribution to describe the overlap ratio.

In the overlapping experiment, the grading dataset was used to conduct the experiment, and the IoU of most overlapping images was below 20% (Figure 4.14a). As the IoU increased, there was a significant reduction in the number of overlapping images. When the IoU exceeded 60%, the number of generated overlapping images was less than five. The uneven distribution of overlapping images was attributed to the various contours of tea shoots. Thus, even if two tea shoots overlap completely, their IoU is rarely over 50%. An IoU of 100% in images indicated the contours and sizes of the two overlapping

tea shoots were identical and overlapped completely. The overlap ratio was defined as  $F(IoU)$  (Eq. 4.2) to represent the level of overlap.





**Figure 4.14** The distribution of overlapping tea shoots.

(a) IoU. (b) overlap ratio.

$$F(x) = \frac{1}{N} \sum_{i=0}^x IoU(i) \quad (4.2)$$

The function  $F(x)$  was the histogram mapping function obtained by applying histogram equalization to the experimented IoU distribution, and  $F(x)$  transformed the uneven distribution into a more balanced one (Figure 4.14b).  $F(IoU)$  provided an intuitive overlapping representation for overlap ratio, where an overlap ratio of 0% denotes no overlap, and an overlap ratio of 100% denotes complete overlap (Table 4.3). Based on the overlap ratio, the 414 collected overlapping images could be arranged from low overlap ratio to high overlap ratio. The majority of overlapping images had an overlap ratio below 20%. Once the overlap ratio reached 25%, the number of images decreased rapidly. The total count of images with an overlap ratio exceeding 32.5% was only 24. These images were divided into eight categories to create an overlapping dataset, ranging from overlap ratio of 0% to 35% in 5% intervals (Table 4.4). The overlapping dataset could be used to evaluate the performance of TSGS under different levels of overlap.

**Table 4.3** Scenarios demonstration: overlapped tea leaves under various IoU and overlap ratio.

<b>IoU</b>	<b>1%</b>	<b>5%</b>	<b>10%</b>	<b>15%</b>	<b>20%</b>
<b>Overlap ratio</b>	<b>3%</b>	<b>15%</b>	<b>30%</b>	<b>48%</b>	<b>58%</b>
<b>Image</b>					
<b>IoU</b>	<b>25%</b>	<b>30%</b>	<b>40%</b>	<b>50%</b>	
<b>Overlap ratio</b>	<b>71%</b>	<b>81%</b>	<b>95%</b>	<b>99%</b>	
<b>Image</b>					

**Table 4.4** Number of images of overlap ratio.

<b>Overlap ratio (%)</b>	<b>0</b>	<b>&lt;2.5</b>	<b>5</b>	<b>10</b>	<b>15</b>	<b>20</b>	<b>25</b>	<b>30</b>	<b>&gt;32.5</b>
<b>Images</b>	50	39	73	75	60	50	24	19	24

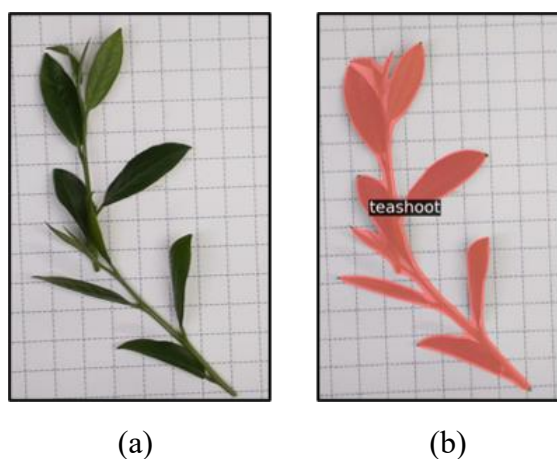


#### 4.4.2 Tea Shoot Segmentation in Overlapping

The overlapping tea shoots caused the single tea shoot segmentation model to separate tea shoots incorrectly. The model performance in the overlapping dataset decreased as the overlap ratio increased (Table 4.5). The performance in separation achieved accuracies of 0.93, 0.80, and 0.67 in the images with overlap ratio values of 5%, 15% and 25% respectively. Images with higher overlap ratio presented more challenges to segment. The vertical overlap between tea shoots resulted in more errors among all the overlapping cases. The possible reason was the similarity between vertically overlapping tea shoots and another tea shoot, as they shared the same growth direction (Figure 4.15).

**Table 4.5** Separation performance in overlapping.

Overlap ratio (%)	<2.5	5	10	15	20	25	30	>32.5
Accuracy	0.97	0.93	0.85	0.80	0.73	0.67	0.53	0.38



**Figure 4.15** Misclassification of identifying the overlapped tea shoots.  
(a) Original. (b) Prediction.



### 4.4.3 Grading Accuracy in Overlapping

The evaluation of grading accuracy for TSGS in overlapping was a comprehensive assessment. The overlapping condition of tea shoots could lead to different errors. For example, the failure to separate overlapping tea shoots resulted in a wrong calculation of leaves. Hence, the grading accuracy for correct separation and all tea shoots were calculated, respectively (Table 4.6). In the correct separation case, the grading accuracy of TSGS only considered the case of correctly identified tea shoots. While in the case of all tea shoots, the grading accuracy was calculated regardless of whether tea shoots were correctly identified.

**Table 4.6** Grading accuracy for overlapped tea shoots in two cases.

<b>Grading Accuracy in the case of Successful Separation</b>									
Overlap ratio (%)	0	<2.5	5	10	15	20	25	30	>32.5
Accuracy	0.94	0.96	0.94	0.88	0.82	0.80	0.81	0.85	0.67
<b>Grading Accuracy in the case All Tea Shoots</b>									
Overlap ratio (%)	0	<2.5	5	10	15	20	25	30	>32.5
Accuracy	0.94	0.94	0.88	0.78	0.68	0.60	0.58	0.45	0.30

Regarding correct separation, TSGS achieved a grading accuracy above 0.80 for overlap ratio ranging from 0% to 30%. Overlapping images with a 30% overlap ratio exhibited a higher grading accuracy of 0.85 due to the limited count of images for correct separation images in an overlap ratio of 30%. Because there were 19 images with an overlap ratio of 30%, tea shoots were successfully identified in only 10 of these images. Hence, 0.85 did not completely describe the grading accuracy in the overlap ratio of 30% for correct separation.

In the case of all tea shoots, TSGS acquired grading accuracies of 0.88, 0.78, 0.68,

0.60, 0.58, and 0.45 for overlap ratio values of 5%, 10%, 15%, 20%, 25%, and 30%, respectively. The images with an overlap ratio of around 10% were recommended for users to use TSGS, as the grading accuracy maintained a value of 0.78. Compared to the case of correct separation, the grading accuracy of TSGS decreased for all tea shoots as the overlap ratio increased. However, correct separation still maintained a grading accuracy above 0.80 for overlap ratios ranging from 5% to 30%. This result indicated the performance in tea shoot identification significantly affected the grading accuracy in overlapping cases.

## 4.5 Tea Shoot Image Capture Module

A capture module was developed for collecting tea shoot images on the production line (Figure 4.16). An app was created to operate the capture module. This app enabled connectivity to the capture module via Bluetooth and allowed the configuration of purchasing information, such as batch number, tea species, and date. The images were captured at a resolution of 3680\*2464 pixels and were then uploaded to TSGS for grading. Additionally, the captured images were backed up within the module. The module supported 4G and 5G frequency bands for image uploading to ensure a stable WiFi signal. The secure file transfer protocol (SFTP) was used to upload the images. SFTP applied data encryption to protect the transferred images and user credentials during transmission. The capture module consisted of two modes: manual mode and automatic mode. In manual mode, the module captured and uploaded an image whenever the user pressed the capture button. In automatic mode, the module continuously captured and uploaded images, synchronizing with the speed of the production line.



**Figure 4.16** Demonstration of the tea shoot image capture module.

## 4.6 Web Developments

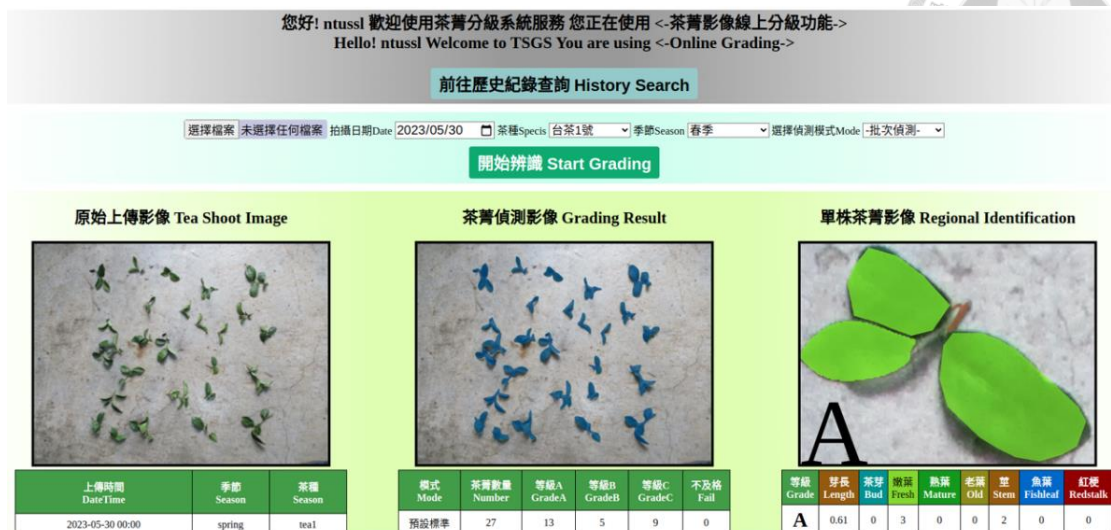
The tea service website offered three main webpages: online identification, database query, and grading criteria customization. This website integrated the tea shoot grading model and database from TSGS, providing users a user-friendly interface for convenient utilization. Users could enter their account and password on the login page to access the tea service webpage.

The online identification webpage offered two modes: general grading function mode and standalone function mode for the organ identification of individual tea shoots. The general grading function mode provided a comprehensive process for evaluating the grade of a batch of tea samples in typical grading scenarios. The other function was specifically developed for researchers to observe traits of tea shoots and to identify and record sample compositions during organs investigations.

In the general grading function mode, users could upload images containing multiple tea shoots (Figure 4.17). Once the uploaded image was processed, the webpage displayed information on the grades of tea shoots and its organs in real time. This mode allowed users to evaluate the overall quality of the tea shoots. In the standalone function mode, users could upload individual tea shoot images for organ identification (Figure 4.18). The webpage presented detailed information about the identified organs, providing users with specific visualization.







**Figure 4.17** General grading function mode for tea shoot grading.



**Figure 4.18** Standalone function mode for organ investigation.

The database query webpage consisted of three main areas: the search area, visualization area, and statistics area (Figure 4.19). Users could specify search conditions in the search area to query the stored tea shoot images in the database. The search conditions included date, batch, and tea species. The dropdown menu in the search area listed all queried images that matched the specified search conditions. By selecting an image from the dropdown menu, the selected image could be shown in the left block of

the visualization area. The middle block presented the identified tea shoot. Users could click on a tea shoot to zoom in, and the right block displayed the identification of organs for the selected tea shoot. The statistics area included three tables. The left table provided detailed information about all images that met the search conditions. It displayed the total number of images, the total number of tea shoots, and the quantities of grades. The middle table presented specific information about the selected image, such as the date, batch number, and species. The right table presented the identified organs for the clicked tea shoot.

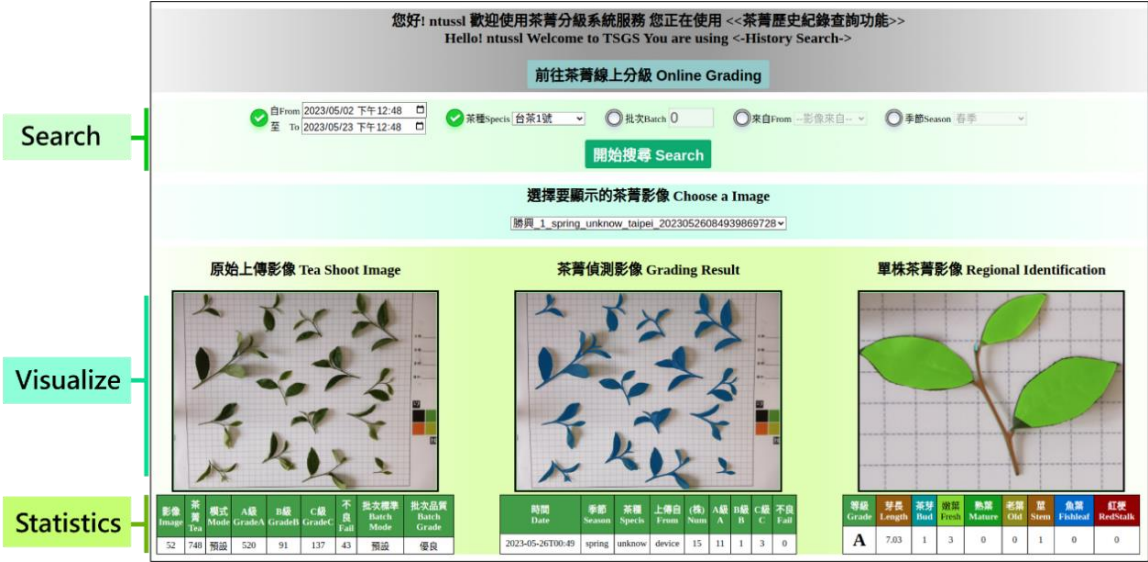


Figure 4.19 Database query webpage.

The grading criteria customization feature allowed users to create their own exclusive grading criteria (Figure 4.20). The criteria were presented on a canvas, where users could add, delete, and edit nodes within the criteria (Figure 4.21). Once the modifications were done, the customized criteria could be transformed into a decision tree for tea shoot grading. This development enabled users without programming skills to build a decision tree based on their demands. Users could flexibly adjust grading criteria based on the harvesting season, tea species, or purchasing strategy.

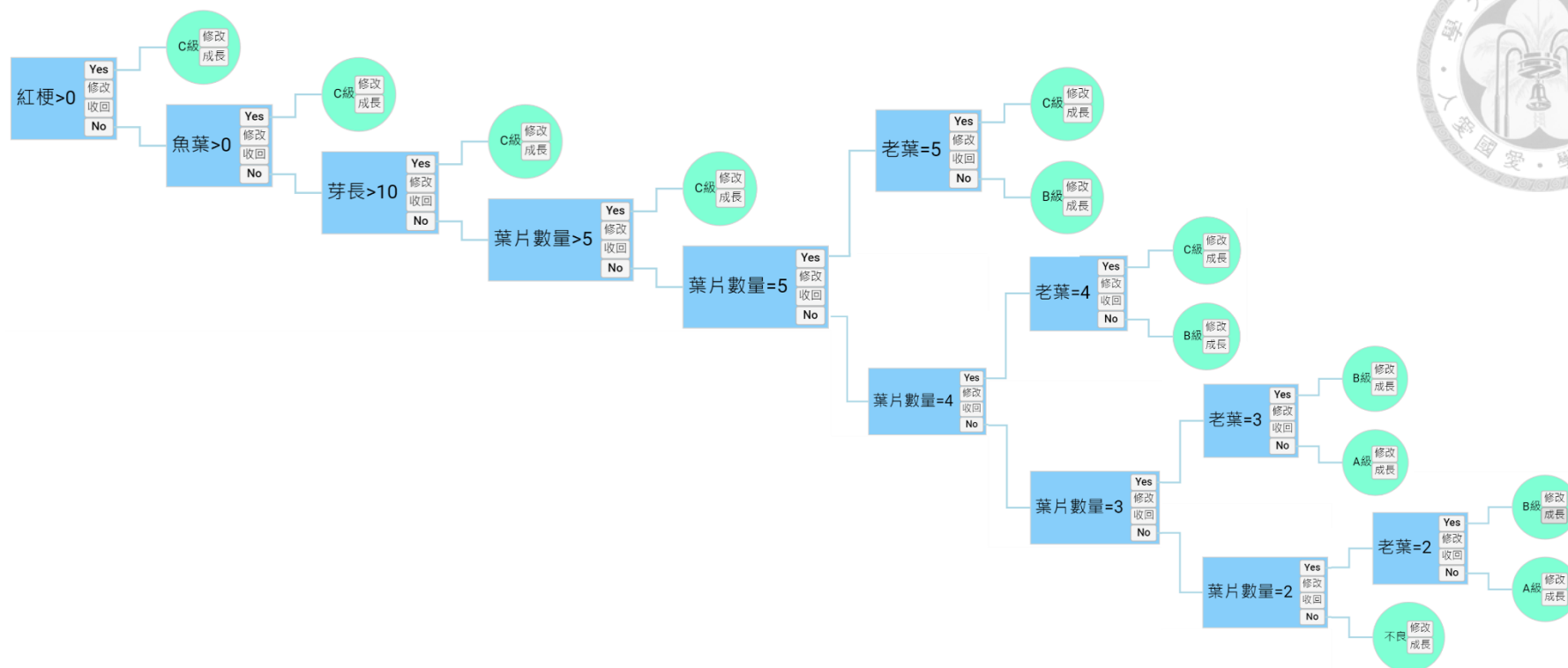
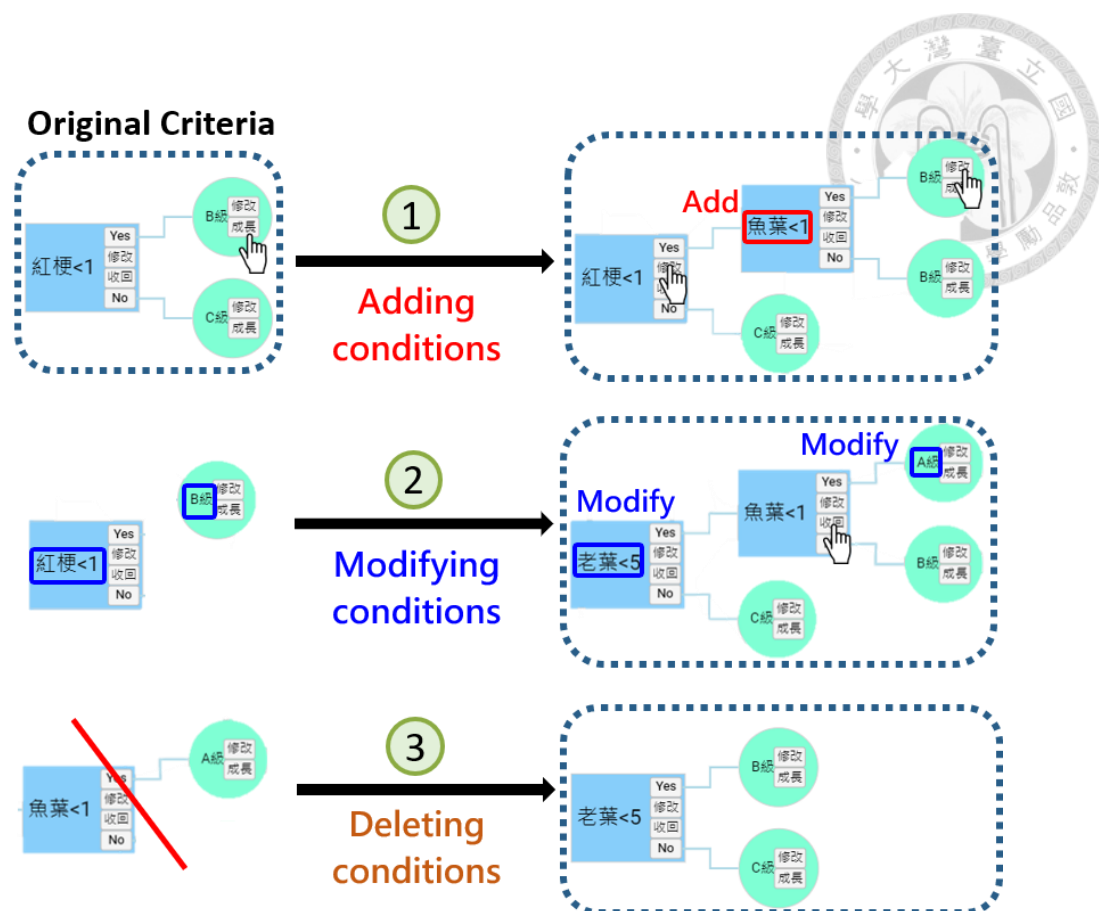


Figure 4.20 Decision tree for grading criteria.



**Figure 4.21** Flexible database design for user-definable grading criteria.


## CHAPTER 5. CONCLUSION AND FUTURE WORK



### 5.1 Conclusion

This study proposed a tea shoot grading system to objectively estimate the quality of plucked tea shoot. The proposed system consisted of a tea shoot image capture module, a tea shoot grading model, and a web development. The capture module was established for collecting tea shoot images. Once the collected image was uploaded to the system, the grading model could identify and classify each tea shoot into one of three grades. The grading model incorporated a combination of a single tea shoot segmentation model, an organ identification model, and grading criteria proposed by an expert from the Taiwan Tea Research and Extension Station. Both models were built on Mask2Former. The single tea shoot segmentation model was trained with 825 tea shoot images, achieving a mean AP of 0.88. The organ identification model was trained with 1337 single tea shoot images, achieving an AP50 of 0.63 for small organs and an AP50 of 0.62 for large organs. The accuracy of the grading model was 0.87, indicating it was a practical method for estimating the quality of plucked tea shoot. Furthermore, web development allowed storing images and displaying identification results on a user-friendly interface. Users could log in to their accounts to access detailed information about plucked tea shoot, including the purchasing date, the distribution of grades, and the identified organs for each tea shoot. Additionally, users could customize their grading criteria according to their specific requirements. The promising results were expected to mitigate the procurement conflict and facilitate the labor-saving tea shoot grading process.

## 5.2 Future work



Future work will focus on enhancing the generalization of the TSGS. By applying a high-speed camera to the capture module, the module could work more stable and speedily, adapting to the production line. A specific device that could scatter tea shoots sparsely should be employed to suppress the level of overlapping in the image. Moreover, incorporating a filtering mechanism to assist in removing broken leaves and twigs from the purchased tea shoot could create a less disruptive identification environment for the TSGS. The single tea shoot segmentation model could be improved by collecting more various images to represent different shooting cases, including challenging scenarios such as low light or blurry conditions. The organ identification model required a more diverse range of cultivars of plucked tea shoots to improve its accuracy and robustness. Different processes can be employed based on the proportion of Grades A, B, and C to achieve more competitive tea products. The transparency and clarity of the grading process could be established by publicly expanding the TSGS to all tea factories and farmers. This system could gradually establish a credible method for determining the grade of plucked tea shoot. Additionally, the system will provide clear and objective grading criteria and information, providing valuable feedback for tea plantation management.

## REFERENCES



- Borah, S., & Bhuyan, M. (2003). Non-destructive testing of tea fermentation using image processing. *Insight-Non-Destructive Testing and Condition Monitoring*, 45(1), 55-58.
- Borah, S., Hines, E. L., & Bhuyan, M. (2007). Wavelet transform based image texture analysis for size estimation applied to the sorting of tea granules. *Journal of Food Engineering*, 79(2), 629-639.
- Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. *Computer Vision–ECCV 2020: 16th European Conference*, Glasgow, U.K. (pp. 213-229).
- Chen, C., Liu, B., Song, F., Jiang, J., Li, Z., Song, C., Li, J., Jin, G., & Wu, J. (2022). An adaptive fuzzy logic control of green tea fixation process based on image processing technology. *Biosystems Engineering*, 215, 1-20.
- Cheng, B., Misra, I., Schwing, A. G., Kirillov, A., & Girdhar, R. (2022). Masked-attention mask transformer for universal image segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. LA, USA. (pp. 1290-1299).
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Hounsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Food and Agriculture Organization Statistical Database. (2022). Available at: <https://www.fao.org/3/cc0238en/cc0238en.pdf>. (Accessed 11 June 2023).



Gayathri, S., Wise, D. J. W., Shamini, P. B., & Muthukumaran, N. (2020). Image analysis and detection of tea leaf disease using deep learning. *International Conference on Electronics and Sustainable Communication Systems*. Coimbatore, India. (pp. 398-403).

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

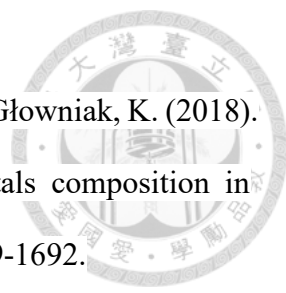
Han, K., Xiao, A., Wu, E., Guo, J., Xu, C., & Wang, Y. (2021). Transformer in transformer. *Advances in Neural Information Processing Systems*, 34, 15908-15919.

Hazra, A., Saha, S., Dasgupta, N., Kumar, R., Sengupta, C., & Das, S. (2021). Ecophysiological traits differentially modulate secondary metabolite accumulation and antioxidant properties of tea plant [*Camellia sinensis* (L.) O. Kuntze]. *Scientific Reports*, 11(1), 2795.

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

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



- 
- Koch, W., Kukula-Koch, W., Komsta, Ł., Marzec, Z., Szwerc, W., & Głowniak, K. (2018). Green tea quality evaluation based on its catechins and metals composition in combination with chemometric analysis. *Molecules*, 23(7), 1689-1692.
- Kamrul, M. H., Rahman, M., Robin, M. R. I., Hossain, M. S., Hasan, M. H., & Paul, P. (2020). A deep learning based approach on categorization of tea leaf. *Proceedings of the International Conference on Computing Advancements*. Kaohsiung, R.O.C.(Taiwan). (pp. 1-8).
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6), 84-90.
- Kumar, S. & Deshmukh, R. (2022). Tea market report. Available at: <https://www.alliedmarketresearch.com/tea-market> (Accessed 11 June 2023).
- LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278-2324.
- Lee, S. H., Lin, S. R., & Chen, S. F. (2020). Identification of tea foliar diseases and pest damage under practical field conditions using a convolutional neural network. *Plant Pathology*, 69(9), 1731-1739.
- Lin, Y. K., & Chen, S. F. (2019). Development of navigation system for tea field machine using semantic segmentation. *IFAC-PapersOnLine*, 52(30), 108-113.
- Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision*. Virtual, Online. (pp. 10012-10022).

Paszke, A., Chaurasia, A., Kim, S., & Culurciello, E. (2017). Enet: A deep neural network architecture for real-time semantic segmentation. *arXiv preprint arXiv:1606.02147*.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas, USA. (pp. 779-788).

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sun, Z., Cao, S., Yang, Y., & Kitani, K. M. (2021). Rethinking transformer-based set prediction for object detection. *Proceedings of the IEEE/CVF International Conference on Computer Vision*. Virtual, Online. (pp. 3611-3620).

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Boston, USA. (pp. 1-9).

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

Wada, K. (2016). Labelme: Image Polygonal Annotation with Python. Available at: <https://github.com/wkentaro/labelme> (Accessed 10 June 2023).

Werbos, P. J. (1990). Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10), 1550-1560.



- Wijeratne, M. A. (2012). Pros and cons of mechanical harvesting: a review of experience on tea harvesters tested. *Tea Bull.* 21(2), 1-9.
- Wu, C. C. (2015). Developing situation of tea harvesting machines in Taiwan. *Engineering, Technology & Applied Science Research*, 5(6), 871-875.
- Wu, D., Chen, X., & He, Y. (2007). Application of image texture for discrimination of tea categories using multi-spectral imaging technique and support vector machine. *International Conference on Computational Intelligence and Security Workshops*. Harbin, China, (pp. 291-294).
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998-6008.
- Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated residual transformations for deep neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu, USA. (pp. 1492-1500).
- Xu, C., Liang, L., Li, Y., Yang, T., Fan, Y., Mao, X., & Wang, Y. (2021). Studies of quality development and major chemical composition of green tea processed from tea with different shoot maturity. *LWT – Food Science and Technology*, 142, 111055.
- Xu, W., Zhao, L., Li, J., Shang, S., Ding, X., & Wang, T. (2022). Detection and classification of tea buds based on deep learning. *Computers and Electronics in Agriculture*, 192, 106547.

Zhang, C., Wang, J., Lu, G., Fei, S., Zheng, T., & Huang, B. (2023). Automated tea quality identification based on deep convolutional neural networks and transfer learning. *Journal of Food Process Engineering*, 46(4), e14303.

Zhu, H., Ye, Y., He, H., & Dong, C. (2017). Evaluation of green tea sensory quality via process characteristics and image information. *Food and Bioprocess Processing*, 102, 116-122.

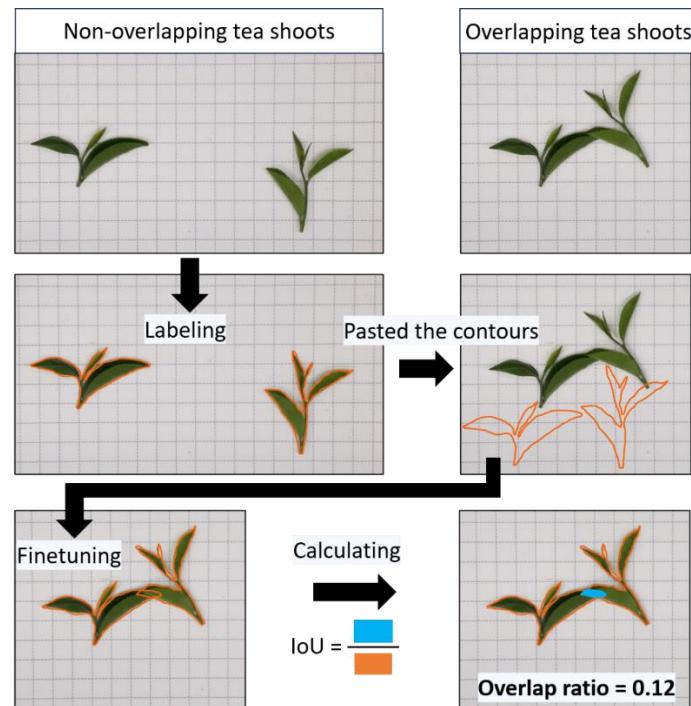
Zhu, X., Su, W., Lu, L., Li, B., Wang, X., & Dai, J. (2021). Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*.

## APPENDIX



### Appendix A. Tea Shoot Labeling for Overlap Ratio Calculation

The overlap ratio of tea shoots required separate calculations of the total areas of the two tea shoots before and after overlapping. The overlap ratio is then determined based on the change in areas between the overlapping and non-overlapping. The contours of overlapping and non-overlapping tea shoots were labeled to calculate their areas. Based on these labeled contours, the overlap ratio of tea shoots could be defined. However, manually labeling the same two tea shoots in different images potentially introduced human errors. Besides, tea shoots are complex-shaped objects, so human errors substantially impact the accuracy of area calculations. These errors caused measured areas to float between plus or minus 10%. This study employed a fixed-contour method to avoid the influence of human errors on the area calculations (Figure A.1).



**Figure A.1** Overlapping tea shoot labeling using fixed-contour method.

This method grouped all tea shoot images, and each group consisted of several images of the same two tea shoots in different degrees of overlap. Then, the two non-overlapping tea shoots in each group were labeled, and the contours of these labeled tea shoots were extracted. Next, the contours of the two tea shoots were pasted on the remaining images within the same group. Finally, the positions and angles of the two contours were adjusted to achieve complete coverage of the two overlapping tea shoots. By utilizing the fixed-contour method, the change in the tea shoots area was only influenced by overlapping. Hence, this method ensured an accurate overlap ratio calculation.

