agriculture

MDPI

Article

Framework for Apple Phenotype Feature Extraction Using Instance Segmentation and Edge Attention Mechanism

Zichong Wang [1], Weiyuan Cui [1,2], Chenjia Huang [1,3], Yuhao Zhou [1], Zihan Zhao [1,2], Yuchen Yue [1,2], Xinrui Dong [1,4] and Chunli Lv [1,*]

[1] China Agricultural University, Beijing 100083, China
[2] National School of Development, Peking University, Beijing 100871, China
[3] Faculty of Humanities, China University of Political Science and Law, Beijing 102249, China
[4] School of Economics and Management, University of Science and Technology Beijing, Beijing 100083, China
* Correspondence: lvcl@cau.edu.cn

Abstract: A method for apple phenotypic feature extraction and growth anomaly identification based on deep learning and natural language processing technologies is proposed in this paper, aiming to enhance the accuracy of apple quality detection and anomaly prediction in agricultural production. This method integrates instance segmentation, edge perception mechanisms, attention mechanisms, and multimodal data fusion to accurately extract an apple's phenotypic features, such as its shape, color, and surface condition, while identifying potential anomalies which may arise during the growth process. Specifically, the edge transformer segmentation network is employed to combine deep convolutional networks (CNNs) with the Transformer architecture, enhancing feature extraction and modeling long-range dependencies across different regions of an image. The edge perception mechanism improves segmentation accuracy by focusing on the boundary regions of the apple, particularly in the case of complex shapes or surface damage. Additionally, the natural language processing (NLP) module analyzes agricultural domain knowledge, such as planting records and meteorological data, providing insights into potential causes of growth anomalies and enabling more accurate predictions. The experimental results demonstrate that the proposed method significantly outperformed traditional models across multiple metrics. Specifically, in the apple phenotypic feature extraction task, the model achieved exceptional performance, with accuracy of 0.95, recall of 0.91, precision of 0.93, and mean intersection over union (mIoU) of 0.92. Furthermore, in the growth anomaly identification task, the model also performed excellently, with a precision of 0.93, recall of 0.90, accuracy of 0.91, and mIoU of 0.89, further validating its efficiency and robustness in handling complex growth anomaly scenarios. The method's integration of image data with agricultural knowledge provides a comprehensive approach to both apple quality detection and growth anomaly prediction, offering reliable decision support for agricultural production. The proposed method, by integrating image data with agricultural domain knowledge, provides precise decision support for agricultural production, not only improving the efficiency and accuracy of apple quality detection but also offering reliable technical assurance for agricultural economic analysis.

Keywords: apple phenotype feature extraction; growth anomaly recognition; deep learning; agricultural economic analysis

Academic Editors: Xiuguo Zou, Xiaochen Zhu, Wentian Zhang, Yan Qian and Yuhua Li

Received: 10 January 2025
Revised: 26 January 2025
Accepted: 27 January 2025
Published: 30 January 2025

Citation: Wang, Z.; Cui, W.; Huang, C.; Zhou, Y.; Zhao, Z.; Yue, Y.; Dong, X.; Lv, C. Framework for Apple Phenotype Feature Extraction Using Instance Segmentation and Edge Attention Mechanism. *Agriculture* **2025**, *15*, 305. https://doi.org/10.3390/agriculture15030305

1. Introduction

Apple phenotype characteristics are an important aspect of agricultural research and production, including the apple shape index (the ratio of longitudinal to transverse diameters), apple size (based on the transverse diameter), color (surface coloration rate), and surface condition (such as freshness and damage) [1]. These characteristics are not only key indicators for evaluating apple quality but also crucial for guiding agricultural production, improving cultivation techniques, and enhancing the commercial value of apples [2]. In recent years, with the development of agricultural automation and intelligence, computer vision and deep learning-based object detection and semantic segmentation technologies have demonstrated great potential in extracting apple phenotype data [3]. Traditional methods for apple phenotype analysis typically rely on manual measurement and empirical judgment, which are not only inefficient but also subject to subjective biases [4]. In contrast, instance segmentation-based techniques can precisely locate the boundaries of apples and simultaneously perform automatic extraction of various phenotype features, enabling large-scale apple quality detection and analysis [5]. However, apple growth is often influenced by environmental factors (such as climate change and soil conditions), leading to growth anomalies such as deformed apple shapes and surface damage, which significantly impact an apple's quality and market value [6].

To achieve precise extraction of apple phenotype characteristics and efficient identification of growth anomalies, a comprehensive method combining instance segmentation and natural language processing (NLP) is proposed in this study [7,8]. On the one hand, the edge transformer segmentation network based on instance segmentation can precisely extract various phenotype features of apples [9]. On the other hand, the NLP module parses and analyzes agricultural text data (such as expert notes, planting records, and meteorological data) to reveal potential causes of and development trends in growth anomalies from multiple dimensions [10]. This integration of image analysis and text parsing not only enhances the comprehensiveness and accuracy of apple phenotype data extraction but also provides new insights for apple quality management and anomaly prediction [11]. In recent years, object detection technologies such as the YOLO series and Mask R-CNN have been widely applied in apple recognition and classification, demonstrating excellent performance in both real-time detection and accuracy [12]. Additionally, semantic segmentation techniques, such as UNet and DeeplabV3+, have been proven to be highly effective in apple surface feature extraction and disease detection tasks [13,14]. Complementing these technologies, the rapid development of NLP has provided new tools for automating the processing of agricultural text data. For instance, Anand et al. proposed a deep learning framework, AgriSegNet, for multi-scale, attention-based semantic segmentation using drone-acquired images to automatically detect agricultural field anomalies [15]. Zhang et al. proposed a segmentation method which outperforms traditional PSO clustering methods in terms of stability and accuracy. It can accurately and effectively segment agricultural product images in various complex environments, facilitating automated agricultural product picking robots [16]. Su et al. introduced a novel data augmentation framework based on random image cropping and patching (RICAP), which effectively improves segmentation accuracy. The proposed framework boosts the average accuracy of deep neural networks from 91.01% to 94.02% by enhancing the original RICAP approach [17]. Zhang et al. developed a pruning inference method which automatically deactivates part of the network structure based on different conditions, reducing network parameters and operations and significantly increasing the network speed. The proposed model achieved accuracy, recall, and mAP rates of 90.01%, 98.79%, and 97.43% in detecting apple flowers, respectively [18]. These advancements demonstrate the promising potential of integrating image analysis with NLP technologies in agricultural production. Therefore, the method proposed in

this paper aims to address the limitations of single-image analysis methods and provides technical support for improving agricultural production efficiency and economic benefits through multimodal data fusion. The contributions of this paper are as follows:

- Integration of Instance Segmentation and Natural Language Processing: For the first time, instance segmentation technology is combined with natural language processing (NLP) to achieve multi-dimensional data fusion in apple phenotype feature extraction and anomaly identification. Instance segmentation ensures the precise extraction of critical phenotype features such as an apple's size, color, and surface condition by accurately delineating an apple's boundaries and surface features. Meanwhile, the NLP module analyzes agricultural text data (e.g., expert notes, planting records, and meteorological data) to reveal potential causes of growth anomalies, providing a comprehensive and accurate analysis which overcomes the limitations of traditional single-image analysis.
- Innovative Application of the Edge Transformer Segmentation Network: The edge transformer segmentation network introduced in this paper integrates Transformer mechanisms with edge-aware modules to better handle complex boundary information of apples. This innovation improves segmentation precision and robustness, especially when dealing with damaged or irregular apple shapes. The method shows excellent performance in extracting key phenotype features from apple surface characteristics and provides reliable support for apple quality assessment and growth anomaly monitoring.
- Multi-Modal Data Fusion for Anomaly Recognition: In contrast to traditional single-image analysis, this study proposes a method for multi-modal data fusion. By combining image data with agricultural text data (e.g., meteorological data and planting records), the NLP module conducts multi-dimensional analysis, leading to more accurate identification of growth anomalies (such as apple deformities and surface damage) and revealing potential causes from a broader context. This cross-modal data fusion offers new perspectives for anomaly prediction and apple quality management in agricultural production.

In the following sections, we provide a detailed overview of the proposed method and its components. Section 2 reviews the foundational methods in object detection, semantic segmentation, and related techniques, which serve as the building blocks for our approach. Section 3 introduces the materials and methods used in this study, including the dataset collection, preprocessing steps, and detailed architecture of the proposed model. Section 4 presents the experimental results and discusses the performance of our method in comparison to existing baseline models. Finally, Section 5 concludes this paper by summarizing the findings and discussing potential future research directions.

2. Related Work

Object detection, semantic segmentation, and natural language processing (NLP) are three critical technologies in modern computer science and artificial intelligence, with broad application prospects in agricultural fruit recognition and phenotype data analysis [19,20].

2.1. Object Detection

Object detection is a core task in computer vision, primarily aimed at locating and classifying target objects within an image [12,21]. The core of object detection lies in predicting the bounding box and category label for each target object [22]. The optimization objective of object detection generally consists of both classification loss and bounding box regression loss, with common loss functions including cross-entropy loss and intersection over union (IoU) loss [23]. The task of object detection can be divided into two main parts:

object localization and object classification. Object localization refers to accurately finding the location of the target in an image and determining its shape, usually by predicting the bounding box through a regression model. Object classification involves determining the category of the target based on its appearance and features. These two tasks are typically trained simultaneously by minimizing both the classification loss and localization loss. The difference between the predicted bounding box and the true values can be expressed as follows:

$$L_{\text{bbox}} = \sum_{i=1}^{N} \left(|x_i - \hat{x}_i| + |y_i - \hat{y}_i| + |w_i - \hat{w}_i| + \left| h_i - \hat{h}_i \right| \right), \tag{1}$$

where x_i, y_i, w_i, h_i are the predicted values, $\hat{x}_i, \hat{y}_i, \hat{w}_i, \hat{h}_i$ are the ground truth values, and N is the number of targets. Common methods for object detection include region proposal-based methods (such as the R-CNN series [24]) and regression-based methods (such as the YOLO series [25]). In agricultural fruit recognition tasks, object detection technology is widely applied to quickly locate the positions and categories of fruits [26]. These models learn the visual features of fruits in images, enabling efficient classification and localization of fruits amd allowing agricultural producers to monitor the growth status of crops in real time and optimize agricultural management. For instance, the Faster R-CNN model, by combining region proposal networks (RPNs) with convolutional neural networks (CNNs), enhances detection accuracy, particularly for smaller fruits like tomatoes and grapes [27].

2.2. Semantic Segmentation

Semantic segmentation is another crucial task in computer vision, aimed at classifying each pixel in an image to achieve fine-grained segmentation of the target objects [28]. Unlike object detection, which focuses on bounding box prediction, semantic segmentation is concerned with pixel-level prediction, requiring more complex model architectures to capture the fine-grained features of an image [29]. The UNet model, through its encoder-decoder structure, progressively extracts deep features from an image and gradually recovers the spatial resolution to achieve fine pixel-level segmentation [30]. DeepLabV3+ integrates dilated convolutions (atrous convolutions) to expand the receptive field, effectively capturing contextual information in the image while enhancing segmentation accuracy through multi-scale features [31]. The loss function for semantic segmentation is typically defined based on the pixel-level cross-entropy:

$$L_{\text{seg}} = -\frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} \sum_{c=1}^{C} y_{ijc} \log \hat{y}_{ijc}, \tag{2}$$

where H and W represent the height and width of the image, C is the number of categories, y_{ijc} is the true label indicating that the pixel (i, j) belongs to category c, and $\hat{y}_{ijc}$ is the predicted probability from the model. To address class imbalance, the weighted cross-entropy or Dice loss is often introduced to improve segmentation accuracy, particularly when handling smaller or less-common classes [32]. In the agricultural domain, semantic segmentation is widely applied in fruit phenotype analysis to extract the contours and surface features of fruits, such as measuring a fruit's size, shape, and surface damage [33–35].

2.3. Natural Language Processing

NLP is a core technology for processing and understanding textual data, with significant applications in agricultural data analysis [36,37]. In agricultural text data analysis, text classification tasks can be used to classify fruit planting records by variety or growth stage, assisting agricultural experts in taking appropriate management measures according to different stages. NER can extract specific key information from large volumes of

agricultural texts, such as climate conditions, soil types, and management practices, which significantly influence fruit growth and pest control [38–40]. Sentiment analysis can help agricultural managers understand the quality of crops, market demand, and other factors based on feedback from farmers or market evaluations, providing a foundation for subsequent agricultural planning and production. Text data embedding is the foundation of NLP, where word embeddings represent the semantic information of words [41]. Common text embedding technologies, such as Word2Vec and GloVe, capture the relationships between words through statistical methods. For example, word vectors **w** represent words, and contextual information is used to learn the semantic representation of each word in a specific context. The optimization objective can be expressed as follows:

$$L_{\text{NLP}} = -\sum_{i=1}^{N} \log p(w_i|\text{context}), \tag{3}$$

where $p(w_i|\text{context})$ represents the probability of predicting the word w_i given the context and N is the total number of words. As agricultural production increasingly relies on digital records, NLP technologies can also be used for automated agricultural log analysis, disease identification, and predictive analysis. For example, NLP-based models can automatically extract fruit growth patterns under different environmental conditions from historical planting records, providing data support for future agricultural decisions [10].

3. Materials and Methods

3.1. Image Construction

In this study, the collection of image datasets and image annotation are key steps in fruit phenotypic analysis and anomaly recognition tasks. To ensure the diversity and representativeness of the data, a large number of apple images were collected from multiple regions, covering different growth environments and climatic conditions. The image data were primarily collected from apple orchards in Changping District, Beijing, Qixia City, Yantai, Shandong Province from March 2023 to August 2024, with some images also sourced from the internet, as shown in Table 1, totaling 24,042 images. The image acquisition equipment and methods employed in this study are critical. The image acquisition method and samples are shown in Figure 1.

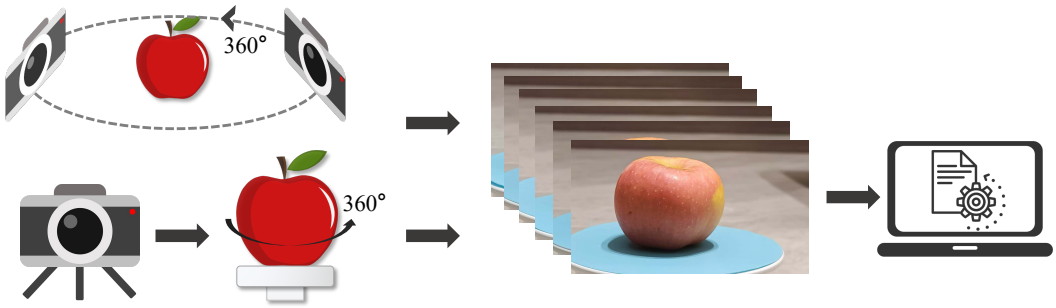

Figure 1. Image acquisition scheme and examples.

A Canon EOS 5D Mark IV camera, manufactured by Canon Inc., headquartered in Tokyo, Japan, was utilized due to its exceptional imaging quality and ability to capture fine details, meeting the requirements for precise acquisition of fruit details. This camera was paired with a Canon EF 100mm f/2.8L Macro IS USM macro lens, which is particularly suited for capturing high-precision images at close distances. To minimize the impact of shadows and reflections on the image quality, all images were captured under soft natural light conditions during early morning or late afternoon. However, in real-world agricultural settings, images may sometimes still contain shadows or overexposure due to fluctuating

lighting conditions. To handle such cases, we employed image preprocessing techniques such as histogram equalization and contrast adjustment to reduce the effects of shadows and overexposure. Additionally, images with excessive overexposure or shadows which significantly obscured fruit features would be identified and excluded from the dataset during the quality control process. This ensured that only high-quality images suitable for phenotypic analysis were included in the final dataset. During image acquisition, particular attention was paid to capturing fruits from multiple angles to document their morphological characteristics and surface abnormalities. To ensure the dataset's representativeness, images were collected covering various fruit types, maturity stages, varieties, shapes, colors, and symptoms of different diseases. The images were obtained from apple orchards in Changping District, Beijing, and Qixia City, Yantai, Shandong Province. These two regions differ in terms of climate and soil conditions, resulting in diverse image backgrounds and fruit states. For image annotation, a semi-automated approach was adopted using the LabelMe tool. Annotators first manually outlined the fruit positions and drew bounding boxes around them. Subsequently, detailed annotations were added regarding each fruit's shape index, size, color grading, and surface conditions.

Table 1. Number of apple images for different data types.

Index	Number
Fruit shape index	6794
Fruit size	5703
Color	6003
Surface condition	5542

Building on the large-scale image data acquisition, additional data were collected by scraping open-source datasets and agriculture-related websites. These sources included expert notes, cultivation records, and meteorological data which, combined with the image data, contributed to the construction of a multimodal dataset. This dataset serves as a valuable resource for fruit phenotypic analysis and anomaly detection while also establishing a foundation for agricultural economic analysis [42,43]. By integrating expert notes, cultivation records, and meteorological data, it is possible to analyze the influence of environmental factors, management practices, and production decisions on fruit quality and yield during the growth process. This provides scientific guidance for agricultural production, optimizing management strategies, and enhancing economic benefits. Expert notes represent a significant component of the dataset, encompassing 20,432 entries documenting common issues and practical experiences throughout the apple cultivation process. These entries span various stages from planting to harvesting and include information on climate, soil, pest management, and irrigation techniques. Cultivation records provide detailed data on apple cultivation processes, including planting times, fertilization practices, irrigation frequencies, and soil treatments, amounting to 19,267 entries. Additionally, meteorological data are another vital source, consisting of 22,803 entries. The growth of apples is closely associated with climatic conditions, as factors such as temperature, humidity, and precipitation directly impact growth cycles, fruit quality, and pest outbreaks. In-depth analysis of these data sources offers valuable references for agricultural economic studies [44,45].

3.2. Data Preprocessing

Data preprocessing is the process of cleaning and adjusting raw images to eliminate noise, correct color biases, and crop regions of interest (ROIs) to improve data effectiveness. Common preprocessing operations in image processing include image cropping, flipping, rotation, resizing, denoising, and white balance correction, as shown in Figure 2.

Figure 2. Data preprocessing: (**A**) original image, (**B**) horizontal flip, (**C**) perspective transformation, (**D**) rotation, (**E**) translation, and (**F**) center crop.

Image cropping involves selecting ROIs to remove irrelevant background information, thereby reducing computational redundancy and emphasizing the target object. This method is particularly important in fruit phenotype analysis as cropping eliminates background interference, allowing the model to focus more on the features of the fruit itself. Denoising is achieved through techniques such as filtering to reduce random noise in the image, thereby improving image quality. Gaussian filtering is a commonly used denoising method. White balance correction is used to correct color distortion in images and restore the true color information of the fruit. This process adjusts the mean values of the red, green, and blue (RGB) channels of the image such that they align with the target values.

3.3. Data Augmentation

Data augmentation is the process of applying various transformations to the original images to generate a more diverse set of training samples, thereby enhancing the robustness and generalization ability of the model. Common data augmentation techniques include Cutout, Mixup, and CutMix. Cutout involves randomly masking a rectangular region on an image to simulate scenarios where the target is partially occluded. This augmentation technique effectively improves the model's prediction ability under occlusion. Let the size of the image I be (H, W) and a region of a size $h \times w$ be randomly occluded at position (x_0, y_0). The augmented image A can be expressed as follows:

$$A(x,y) = \begin{cases} 0, & \text{if } x_0 \le x < x_0 + h,\ y_0 \le y < y_0 + w, \\ I(x,y), & \text{otherwise.} \end{cases} \tag{4}$$

Mixup involves linearly mixing two images at a certain ratio, with the aim of improving the model's smoothness and prediction ability for unseen samples. The augmentation formula for Mixup is

$$A = \lambda I_1 + (1-\lambda) I_2, \quad y = \lambda y_1 + (1-\lambda) y_2, \tag{5}$$

where $\lambda \sim \text{Beta}(\alpha, \alpha)$ is the mixing coefficient sampled from a Beta distribution, I_1 and I_2 are the two original images, and y_1 and y_2 are the corresponding labels. CutMix combines the ideas of Cutout and Mixup by pasting a portion of one image onto another and adjusting the labels to reflect the proportion of the mixed region. Let I_1 and I_2 be the two images. The augmented image A can be expressed as

$$A(x,y) = \begin{cases} I_1(x,y), & \text{if } (x,y) \in \text{Region1}, \\ I_2(x,y), & \text{otherwise.} \end{cases} \tag{6}$$

The label adjustment formula is

$$y = \lambda y_1 + (1-\lambda) y_2, \tag{7}$$

where λ is the proportion of Region1.

3.4. Hyperparameters

Dataset partitioning is a key step in training and validating machine learning models. The goal is to appropriately allocate data for model training, validation, and testing to ensure the scientific and representative evaluation of the model's performance. In this study, the dataset was divided into training and validation sets with a ratio of 8:2. Additionally, to enhance the stability and robustness of the model, K-fold cross-validation was employed. Specifically, the dataset was divided into K non-overlapping subsets, with $K-1$ subsets used for training and the remaining subset used for validation. After repeating this process K times, the average performance was computed. The formula for calculating the average performance of K-fold cross-validation is

$$\text{Performance} = \frac{1}{K}\sum_{k=1}^{K} P_k, \tag{8}$$

where P_k represents the performance metric (such as the accuracy or $mIoU$) for the kth validation. In fruit phenotype analysis, through proper data preprocessing, augmentation, and partitioning, a high-quality dataset was built to provide a reliable foundation for subsequent model training and performance evaluation. This method not only enhances the model's robustness but also provides important support for practical deployment.

3.5. Proposed Method

The method presented in this study, from its overall design to specific implementation, aims to extract fruit phenotype features and identify growth anomalies through a multimodal data fusion approach, combining instance segmentation and NLP techniques. The overall framework of the model is a continuous flow from input data to final output predictions, involving the collaborative work of multiple modules, as shown in Figure 3.

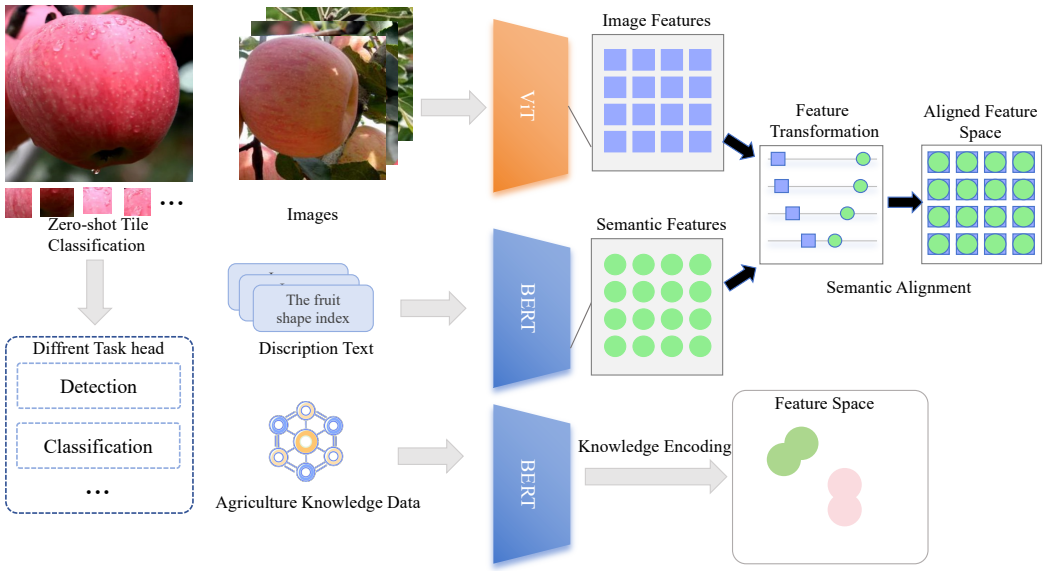

Figure 3. Flowchart of the whole process in the proposed method, where the Agriculture Knowledge Data block integrates domain-specific agricultural knowledge to enhance model performance and decision making.

First, after the input data were preprocessed, it entered the instance segmentation network for feature extraction from the images. The extracted image features were then processed by subsequent modules, such as the feature transformation and alignment stages,

to refine and integrate the information. Concurrently, the Agricultural Knowledge Data block, which includes textual data such as planting records, climate data, and expert notes, was parsed by the NLP module. This NLP process extracted meaningful insights from the textual data to complement the visual features extracted from the images. The integration of both image and agricultural knowledge data allowed for more accurate fruit phenotype recognition and the identification of growth anomalies. Finally, the combined results from both the image-based and text-based analyses were used to make precise predictions regarding the fruit phenotype and detect potential growth anomalies.

3.5.1. Edge Transformer Segmentation Network

In this study, the proposed edge transformer segmentation network is a key component for fruit phenotype analysis and growth anomaly identification, combining agricultural images with agricultural knowledge (text data).

As shown in Figure 4, the network design incorporates a deep fusion of instance segmentation and NLP techniques, aiming to precisely extract fruit phenotype features through image segmentation and text analysis, while also integrating agricultural text data (such as planting records and meteorological data) for anomaly detection. The network input consists of two parts: (1) agricultural images, which provide visual information about the fruits, and (2) agricultural knowledge (text data), which supplies multi-dimensional information regarding the fruit's growth environment, climate change, and management practices. The fusion of this multimodal data helps enhance the prediction accuracy of fruit growth anomalies and provides robust support for fruit quality management and agricultural production decision making. The network implementation explicitly leverages the Transformer architecture to model cross-modal interactions rather than applying simple rules to a 2D matrix. Specifically, both image and text features are projected into a shared feature space, where multi-head self-attention (MHSA) is employed to dynamically learn the dependencies between different modalities. First, the network extracts basic features from the image using convolutional layers, generating an initial feature map. This feature map then passes through the edge-aware module, which is designed to enhance the network's focus on the fruit's edge areas, especially when the fruit has a complex shape or surface damage. The edge-aware module further strengthens the edge features in the image by combining traditional edge detection algorithms (such as Sobel or Canny) with convolutional operations. The module computes the edge feature map of an image and merges it with the feature map generated by the convolution layers, enhancing the sensitivity to fruit contours, cracks, and other detailed regions and thereby improving the segmentation accuracy. Next, the enhanced feature map enters the Transformer module for global information modeling. Unlike conventional convolution-based approaches, which rely on local receptive fields, the Transformer module effectively models long-range dependencies within an image through self-attention mechanisms. Given an input feature representation F, the self-attention operation is computed as follows:

$$Z = \text{softmax}\left(\frac{(W_Q F)(W_K F)^T}{\sqrt{d_k}}\right) W_V F, \tag{9}$$

where W_Q, W_K, and W_V are the query, key, and value projection matrices, respectively, and d_k is the dimensionality of the key vectors. By leveraging this global attention mechanism, the network learns contextual relationships across different regions of an image, particularly addressing cases where fruit morphology spans multiple spatial regions. In addition to image data, the network's input also includes textual data related to fruit growth. The NLP module parses and analyzes agricultural text data (such as climate records, planting records, and expert notes), providing additional support for predicting fruit

growth anomalies. To ensure a seamless fusion of text and image data, textual information is encoded using a Transformer-based embedding model, such as BERT or a domain-specific language model. This process converts textual descriptions into dense feature vectors, which are then aligned with the visual embeddings via cross-attention layers. The NLP module employs a Transformer structure to process the text data, extracting key information related to fruit growth such as climate changes, fertilization management, and environmental factors. This information, combined with the image data, provides more comprehensive background knowledge for the network, helping the model better understand anomalies in the fruit's growth process while identifying its phenotype features. To further refine multimodal interactions, a cross-attention mechanism is incorporated, where text features serve as queries and image features serve as keys and values, thereby guiding the visual representation learning process based on domain knowledge. Ultimately, the edge transformer segmentation network, through the joint processing of image and text data, is capable of providing high-precision and robust segmentation results for fruit phenotype analysis and growth anomaly detection. Compared with traditional methods, the network design presented in this study fully accounts for the fruit's morphological features and surface conditions and the influence of external environments. This design, particularly when dealing with fruits with fuzzy edges, damage, or irregular shapes, demonstrates stronger accuracy and robustness. The multimodal data fusion approach not only enhances the accuracy of fruit quality assessment but also provides powerful technical support for intelligent decision making in agricultural production.

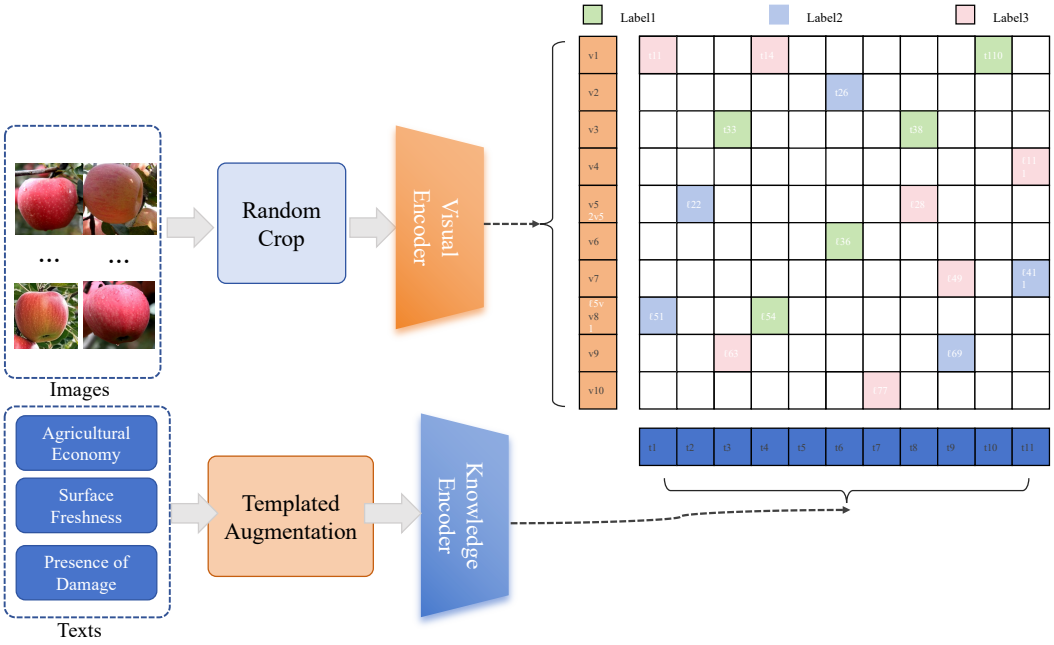

Figure 4. Architecture of edge transformer segmentation network.

3.5.2. Edge Attention Mechanism

The proposed edge attention mechanism module is an extension of the traditional self-attention mechanism, with a particular focus on edge information in the image to enhance the network's sensitivity to the boundary regions of an apple. The goal of this module is to introduce an edge-aware mechanism which increases the network's attention to the fruit's boundary regions, thereby optimizing segmentation results, particularly when fruit shapes are complex, the surface is damaged, or the boundaries are blurred. Compared with the traditional self-attention mechanism, the edge attention mechanism enables the network to

prioritize the learning and enhancement of edge features during training, ensuring accurate segmentation in complex fruit shapes and damaged regions, as shown in Figure 5.

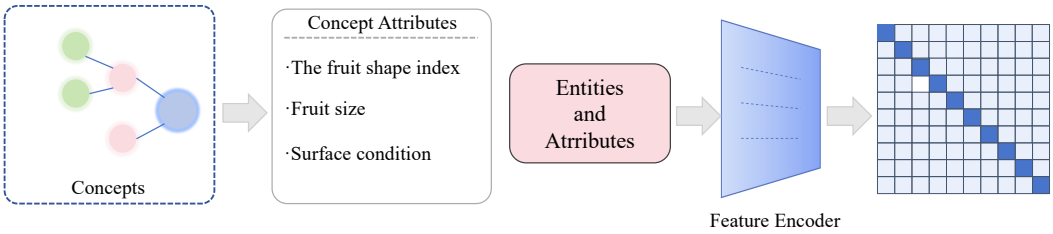

Figure 5. Flowchart of edge attention mechanism.

In the design of the edge attention mechanism, the core idea of the network architecture is to combine the advantages of the standard self-attention mechanism with the edge-aware module. Specifically, the network first uses traditional convolutional layers to extract feature maps and then employs edge detection algorithms to generate an edge map, followed by a weighted self-attention mechanism to increase the network's focus on the edge regions. The design of the number of layers in the edge attention mechanism includes multiple layers of self-attention mechanisms, with each layer containing a multi-head self-attention module and a feedforward neural network. The width and height of each layer remain consistent, ensuring spatial consistency of the feature maps across different layers. The specific design of the network is as follows. The input image size is $H \times W \times C$, where H is the image height, W is the image width, and C is the number of input channels. In the self-attention layers, the output feature map remains $H \times W \times C$ and is further enhanced in the edge-aware module, improving the edge regions of the feature map. To enhance the self-attention mechanism with edge awareness, an edge-weighting function $E(i, j)$ is introduced, dynamically adjusting the attention computation such that pixels in the edge regions receive higher weights. The edge-enhanced attention is formulated as follows:

$$Z_{\text{edge}} = \text{softmax}\left(\frac{(W_Q F)(W_K F)^T}{\sqrt{d_k}}\right) W_V (F \cdot E). \tag{10}$$

where E represents the edge feature map generated by the edge detection module, and its influence is adjusted through learnable weights. This formulation ensures that the network focuses more on fruit contours, even when dealing with irregular shapes or blurred edges, thereby improving segmentation precision. The parameters for each layer are as follows. The dimension of each self-attention head is C/N, where N is the number of heads, typically set to eight. The output from each layer is processed through a feedforward neural network, with the width and height of the feature map maintained at $H \times W$ and the number of channels remaining being C. To incorporate edge information, a weighting mechanism is designed by combining the edge feature map E with the feature map F to enhance the attention weights on the edge regions. The specific weighted calculation is given by

$$\text{Edge Attention} = \frac{\exp(\text{Edge Feature})}{\sum_i \exp(\text{Edge Feature}_i)}, \tag{11}$$

where Edge Feature refers to the edge features generated by the edge detection module and E represents the edge regions of the image. Through this mechanism, the edge attention mechanism prioritizes modeling the fruit's edges, ensuring segmentation precision, particularly in regions where the fruit shape is complex or damaged. The edge attention mechanism enhances the network's attention to the boundaries and surface damage of the fruit, while the edge transformer segmentation network handles extraction of the over-

all phenotype features from an image. These two modules work collaboratively in the segmentation process, improving segmentation precision. By enhancing the network's sensitivity to edge regions, the edge attention mechanism provides higher segmentation accuracy in detailed image parts, particularly in edges and damage areas. Meanwhile, the edge transformer segmentation network focuses on processing global information and fruit morphological features. This design ensures that when handling complex scenarios and fruit with detailed features, the model can perform segmentation with higher precision, especially when fruit surfaces exhibit cracks, spots, or rot, while still maintaining high recognition accuracy. Therefore, the edge attention mechanism not only improves the model's performance in segmentation tasks, but also, by combining edge-aware mechanisms with global context modeling, greatly enhances the model's adaptability and robustness in fruit phenotype analysis.

3.5.3. Edge Loss Function

The proposed edge loss function in this study is a novel loss function designed to optimize edge accuracy in image segmentation tasks, especially for fruit phenotype analysis. Traditional loss functions, such as cross-entropy loss and Dice loss, are effective in optimizing models in most cases. However, they typically overlook the details of the edge regions, especially when dealing with complex shapes, blurred boundaries, or damaged areas, leading to unclear segmentation results at the boundaries. To address this issue, the edge loss function introduces an edge-aware mechanism which assigns more weight to the edge regions of an image, enabling finer segmentation. The core idea of the edge loss function is to incorporate specific optimization for the edge regions on top of traditional loss functions. Traditional loss functions are usually optimized at the pixel level across the entire image, neglecting the importance of edge regions in image segmentation. In fruit phenotype analysis, accurate segmentation of the fruit's contours, surface damage, and deformed areas is crucial for the quality of the results. The edge loss function, by weighting the loss of the edge regions, ensures that the model focuses more on the boundary areas during training, thereby improving segmentation precision. The mathematical formula for edge loss function can be expressed as follows:

$$L_{\text{Edge}} = \sum_{i=1}^{N} \Big(|\hat{y}_i - y_i| \cdot \mathbb{1}_{\text{edge}}(i) \Big), \tag{12}$$

where $\hat{y}_i$ is the predicted value, y_i is the ground truth value, N is the total number of pixels, and $\mathbb{1}_{\text{edge}}(i)$ is the indicator function. When pixel i belongs to the edge region, $\mathbb{1}_{\text{edge}}(i) = 1$; otherwise, it is zero. This weighted loss allows the network to place more learning emphasis on the fruit's edge regions while reducing overemphasis on non-edge areas, resulting in more precise boundary segmentation. In traditional loss functions, such as cross-entropy loss, the formula is

$$L_{\text{CE}} = -\sum_{i=1}^{N} (y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)) \tag{13}$$

These loss functions typically optimize all pixels with equal weight, ignoring the uniqueness of edge regions in segmentation. The edge loss function, on the other hand, addresses this by adding specialized weighting for the edge areas, allowing the network to focus more on precise segmentation of the edges, especially when handling complex fruit shapes and surface damage. This design significantly improves the recognition accuracy of edge parts, particularly when fruit surfaces exhibit cracks, spots, or rot, preventing misegmentation due to blurred edges. The integration of the edge attention mechanism and

edge loss function within the edge transformer segmentation network plays a critical role in fruit phenotype analysis and anomaly detection. The Transformer-based segmentation model captures both global context and fine-grained local details, while the edge-specific enhancements ensure that boundary information is accurately preserved and learned. Compared with traditional segmentation methods which rely solely on pixel-wise classification, this approach effectively models complex shape variations and enables precise contour extraction, making it particularly useful in agricultural applications where fruit shape and surface characteristics are crucial quality indicators. By incorporating both attention-based edge enhancement and an edge-sensitive loss function, the proposed approach not only improves the segmentation accuracy but also enhances robustness in real-world agricultural scenarios. The ability to distinguish subtle fruit defects and deformations makes this method valuable for automated quality assessment, early disease detection, and optimized resource management in smart agriculture.

3.6. Experimental Design

3.6.1. Hardware and Software Platforms

In this study, the choice and configuration of the hardware and software platform played a critical role in ensuring efficient execution of the experiments and the reliability of the results. On the hardware side, an NVIDIA A100 GPU was used, which is designed specifically for artificial intelligence and high-performance computing. Based on the Ampere architecture, the A100 supports multi-precision computations (including FP64, FP32, TF32, and FP16), with up to 6912 cores and 40 GB or 80 GB of memory, providing exceptional data processing capabilities. During large-scale data training and deep learning model execution, the A100 GPU significantly accelerates computation and supports parallel processing of multiple tasks. The experimental platform also includes a high-performance CPU, more than 256 GB of memory, and high-speed NVMe solid-state drives, ensuring efficient data loading, processing, and storage. For the software platform, the experiment was run on a Linux operating system, specifically Ubuntu 20.04 LTS, which is widely used for its stability and good support for deep learning frameworks. The deep learning models were developed and trained using the PyTorch framework version 1.12.0, with CUDA 11.6 and cuDNN 8.3 installed to fully leverage GPU acceleration. The Adam optimizer was chosen for its fast convergence and adaptability, making it one of the mainstream optimization algorithms in deep learning. The learning rate was set to 0.001, which was determined through several experimental adjustments to be the optimal value, ensuring quick convergence without oscillation. Additionally, the OpenCV and Albumentations libraries were used for efficient data preprocessing and augmentation, while model performance evaluation and visualization were performed using tools such as Matplotlib and Seaborn. The entire experimental environment was deployed through Docker containerization, which not only improved the reproducibility of the experiments but also facilitated cross-platform migration.

3.6.2. Baseline Models

To comprehensively assess the performance of the proposed method, several classic deep learning models were chosen as baseline models, including Tiny-Segformer [23], Mask R-CNN [33], UNet [30], UNet++ [46], and DeepLabV3+ [31]. These models represent different technological directions and architectural characteristics in the field of image segmentation. Tiny-Segformer is a lightweight Transformer architecture which combines an efficient self-attention mechanism with convolution operations, maintaining computational efficiency while offering strong feature extraction capabilities, making it particularly suitable for resource-constrained scenarios. Mask R-CNN is a dual-task model based on

object detection and segmentation capable of generating pixel-level segmentation masks in addition to bounding box detection. Its loss function includes the classification loss L_{cls}, bounding box regression loss L_{bbox}, and segmentation loss L_{mask}. Both UNet and its improved version, UNet++, use an encoder-decoder structure at their core, integrating multi-scale features through skip connections. These architectures are particularly suitable for fine-grained segmentation tasks in medical and agricultural imaging, with UNet++ further enhancing the network's expressive power by redesigning the skip connection modules. DeepLabV3+ uses dilated convolution (atrous convolution) and the Atrous Spatial Pyramid Pooling (ASPP) module to effectively capture multi-scale contextual information, with the loss function typically based on pixel-level cross-entropy.

To ensure a fair comparison, all baseline models were trained and tested on the same dataset, using the same image data for phenotype feature extraction and growth anomaly detection. However, since most of these models (such as UNet, UNet++, Mask R-CNN, and DeepLabV3+) are designed primarily for image-based segmentation tasks, they were not originally built to process multimodal information. Therefore, for these models, only the image input was utilized, without directly integrating meteorological data or other agricultural textual information. In contrast, the proposed method incorporates both image and textual data, leveraging a dedicated NLP module to process agricultural knowledge (such as planting records and meteorological data) and fusing it with visual features through an attention-based multimodal learning approach. To maintain a fair experimental set-up, Tiny-Segformer, which is a Transformer-based segmentation model, was extended with an NLP component similar to the one in the proposed system. However, due to its original lightweight design, its capacity for processing and integrating textual information remains more limited than the proposed method. These baseline models provide a reference standard for performance comparison in this study and help thoroughly verify the effectiveness of the proposed method.

3.6.3. Evaluation Metrics

In this study, several evaluation metrics were used to comprehensively assess the model's performance, including the precision, recall, accuracy, and mean intersection over union (mIoU). These metrics measure the model's performance in fruit phenotype analysis and anomaly detection from different dimensions. Precision measures the proportion of true positive samples among all samples predicted to be positive, focusing on the correctness of the predictions, which is particularly important in high-precision scenarios. Recall measures the proportion of true positive samples which were correctly predicted to be positive, reflecting the model's ability to capture positive samples and serving as an important metric for evaluating false negatives. Accuracy represents the proportion of correctly classified samples among all predictions, suitable for evaluating the overall performance of the model. The mIoU, commonly used in semantic segmentation tasks, calculates the ratio of the intersection to the union between the predicted and ground truth regions, averaging this ratio across all categories to measure the global consistency of the model's segmentation results. The mathematical definitions of these evaluation metrics are as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{14}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{15}$$

$$\text{IoU} = \frac{\text{Prediction} \cap \text{Ground Truth}}{\text{Prediction} \cup \text{Ground Truth}}, \tag{16}$$

$$\text{mIoU} = \frac{1}{C}\sum_{c=1}^{C} \text{IoU}_c, \tag{17}$$

where TP represents true positives, FP represents false positives, FN represents false negatives, TN represents true negatives, $\cap$ denotes the intersection, $\cup$ denotes the union, C is the total number of classes, and IoU_c is the intersection over union for class c.

4. Results and Discussion

4.1. Experimental Results of Phenotype Feature Extraction Models

The experimental design presented in this study aims to evaluate the performance of different deep learning models in the task of extracting fruit phenotype features. By comparing the precision, recall, accuracy, and mIoU metrics of various models, the study analyzed the advantages and shortcomings of each model in fine segmentation tasks, providing a theoretical foundation for subsequent model optimization. The models used in the experiments included UNet, Mask R-CNN, DeeplabV3+, UNet++, Tiny-Segformer, and the proposed method. Through these comparative experiments, the influence of the model architecture, loss functions, and optimization strategies on the accuracy of phenotype feature extraction could be deeply understood, offering theoretical support for intelligent fruit analysis in agricultural production.

As shown in Table 2, the experimental results demonstrate different levels of performance across all models in terms of precision, recall, accuracy, and the mIoU. The UNet model achieved a precision of 0.84, recall of 0.82, accuracy of 0.83, and mIoU of 0.80, indicating good performance in basic segmentation tasks but with room for improvement in handling finer details. Compared with UNet, Mask R-CNN showed improvements in its precision and recall, achieving scored of 0.86 and 0.83, respectively, with an accuracy and mIoU of 0.85 and 0.83, respectively, indicating better handling of object boundaries and details in the instance segmentation task. DeeplabV3+ introduced atrous convolution and spatial pyramid pooling modules, which enhanced the model's ability to capture multi-scale features, with precision, recall, and mIoU scores of 0.89, 0.86, and 0.86, respectively, demonstrating an advantage in processing multi-scale contextual information. UNet++ further improved the precision to 0.90, the recall to 0.88, and the mIoU to 0.87, showing that its enhanced skip connections improved detail recovery. Tiny-Segformer, a lightweight Transformer architecture, demonstrated strong feature extraction capabilities with efficient self-attention and convolution operations, achieving precision and recall scores of 0.92 and 0.89, respectively, and an mIoU of 0.89, indicating that the model could provide strong feature extraction while maintaining computational efficiency. The proposed method outperformed all other models, with a precision of 0.95, recall of 0.91, accuracy of 0.93, and mIoU of 0.92, demonstrating that the model, through combining edge perception mechanisms and global contextual modeling, provided higher precision and robustness in the complex task of fruit phenotype feature extraction. From a theoretical perspective, the architectures of the models significantly influenced the experimental results. Both UNet and UNet++ employ encoder-decoder structures and fuse multi-scale features through skip connections, but UNet++ further enhances network expressiveness through improved skip connections, leading to better performance in detail recovery. Mask R-CNN, a dual-task model for object detection and segmentation, not only provides bounding box detection but also generates pixel-level segmentation masks, which contribute to better segmentation results when handling object boundaries and complex structures. DeeplabV3+ effectively captures multi-scale contextual information through atrous convolution and spatial pyramid pooling modules, which is particularly advantageous when dealing with complex backgrounds and large objects. The lightweight Transformer architecture of Tiny-Segformer, which combines efficient self-attention mechanisms with convolution operations, enables

the model to extract strong global features while maintaining computational efficiency, which is why it achieves high precision. The proposed method combines the advantages of these models by incorporating edge perception mechanisms and global contextual information modeling, enabling more precise capture of fruit contours, surface damage, and other detailed features, which is why it outperformed the other models across all evaluation metrics. These results highlight the significant impact of the model structure, loss functions, and optimization strategies on segmentation performance, especially in complex scenarios. The models which incorporated edge information and a global context had greater robustness and accuracy.

Table 2. Experimental results of phenotype feature extraction models.

Model	Precision	Recall	Accuracy	mIoU
UNet	0.84	0.82	0.83	0.80
Mask R-CNN	0.86	0.83	0.85	0.83
DeeplabV3+	0.89	0.86	0.87	0.86
UNet++	0.90	0.88	0.99	0.87
Tiny-Segformer	0.92	0.89	0.91	0.89
Proposed Method	0.95	0.91	0.93	0.92

4.2. Experimental Results of Growth Anomaly Recognition Models

The design of this experiment aims to evaluate the performance of various deep learning models in the task of growth anomaly recognition, particularly in identifying abnormal conditions during fruit growth, such as pest damage and cracks. As shown in Table 3, the experiment compared the performance of different models based on various metrics to analyze their advantages and limitations in handling growth anomalies and to provide a theoretical basis for practical applications in fruit quality monitoring and pest warning systems.

Table 3. Experimental results of growth anomaly recognition models.

Model	Precision	Recall	Accuracy	mIoU
UNet	0.82	0.80	0.81	0.79
Mask R-CNN	0.84	0.82	0.83	0.81
DeeplabV3+	0.87	0.84	0.85	0.82
UNet++	0.89	0.87	0.88	0.84
Tiny-Segformer	0.91	0.88	0.90	0.86
Proposed Method	0.93	0.90	0.91	0.89

From the experimental results, it is evident that all models exhibited varying degrees of performance in the growth anomaly recognition task. The UNet model achieved a precision of 0.82, recall of 0.80, accuracy of 0.81, and mIoU of 0.79, demonstrating basic performance. Compared with UNet, Mask R-CNN showed improvements in precision, recall, and mIoU, with values of 0.84, 0.82, and 0.81, respectively. This indicates that its dual-task structure (object detection and instance segmentation) played an active role in recognizing fruit growth anomalies. DeeplabV3+, with the incorporation of atrous convolution and spatial pyramid pooling modules, showed an advantage in handling multi-scale contextual information, achieving a precision of 0.87, recall of 0.84, and mIoU of 0.82. This suggests that DeeplabV3+ performs better than the previous models in complex scenarios. UNet++, with its improved skip connection module, demonstrated excellent performance in terms of both precision (0.89) and recall (0.87), with an mIoU of 0.84, confirming its advantage in detail recovery and multi-scale information fusion. Tiny-Segformer, a lightweight Transformer-based

architecture, combined self-attention mechanisms and convolution operations, achieving a precision of 0.91, recall of 0.88, and mIoU of 0.86, indicating its powerful feature extraction ability and global information modeling capabilities. The proposed method outperformed all other models in all metrics, with a precision of 0.93, recall of 0.90, accuracy of 0.91, and mIoU of 0.89, demonstrating that combining edge perception mechanisms and global information modeling significantly improves recognition accuracy and robustness in handling complex growth anomaly scenarios. From a theoretical analysis perspective, the architectural characteristics of the models directly influenced the experimental results. Both UNet and UNet++ use an encoder-decoder structure and fuse multi-scale features through skip connections. However, UNet++ further optimizes skip connections, improving its ability to recover details, which is why it performed better in growth anomaly recognition. Mask R-CNN, with its dual-task structure for object detection and instance segmentation, can simultaneously precisely segment a fruit's location and area, making it superior to UNet in handling anomalies with clear boundaries. DeeplabV3+ benefits from atrous convolution and spatial pyramid pooling modules, enabling it to capture richer contextual information in multi-scale contexts and thereby improving its ability to recognize complex backgrounds and irregular anomalies. Tiny-Segformer, through the combination of efficient self-attention mechanisms and convolution operations, excels in feature extraction and global information modeling, allowing it to better capture long-range dependencies, which contributed to its high precision and recall scores. The proposed method introduces an edge perception mechanism, enabling the network to focus more on fruit edges and anomaly regions. By combining this mechanism with global information modeling, the model effectively improves performance in complex growth anomaly scenarios. The edge perception mechanism strengthens the precise recognition of anomaly regions, while the Transformer architecture enhances the understanding of the global context, allowing the model to maintain high precision when dealing with complex fruit shapes and surface damage. These experimental results suggest that optimizing the model architecture, enhancing feature extraction capabilities, and introducing edge perception mechanisms are crucial for improving the accuracy and robustness of growth anomaly recognition.

4.3. Accuracy Results of Different Models for Various Phenotype Features

As shown in Table 4, the design of this experiment aimed to evaluate the performance of various deep learning models in extracting and recognizing fruit phenotype features such as the fruit shape index, fruit size, color, and surface state. By comparing the accuracy of different models on these phenotype features, the goal was to analyze the performance differences among models when handling various fruit features and to investigate the impact of different model architectures on feature extraction and recognition accuracy. The analysis of the experimental results provides theoretical support for selecting the optimal model in fruit phenotype analysis tasks and reveals the role of model architecture and optimization strategies in enhancing fruit feature extraction accuracy.

Table 4. Accuracy results of different models for various phenotype features.

Model	Fruit Shape Index	Fruit Size	Color	Surface State
UNet	0.81	0.82	0.83	0.84
Mask R-CNN	0.83	0.84	0.85	0.87
DeeplabV3+	0.85	0.87	0.87	0.89
UNet++	0.86	0.89	0.89	0.91
Tiny-Segformer	0.90	0.91	0.92	0.92
Proposed Method	0.92	0.93	0.94	0.95

The experimental results show significant differences in the performance of all models for different phenotype features. The UNet model exhibited relatively lower accuracy scores across all phenotype features, with values of 0.81, 0.82, 0.83, and 0.84, indicating its basic ability in fruit feature extraction but struggles in capturing complex feature relationships due to its relatively simple architecture. Mask R-CNN showed improvement over UNet, especially in recognizing fruit surface states, with an accuracy of 0.87 compared with 0.84 for UNet. This suggests that the model's dual-task structure (object detection and instance segmentation) enhances boundary detail extraction in the instance segmentation task. DeeplabV3+ incorporates atrous convolution and spatial pyramid pooling modules, which better handle multi-scale features, leading to improved recognition accuracy for the fruit size and surface state, with values of 0.87 and 0.89, respectively. UNet++ improves upon the skip connection in the encoder-decoder structure, further enhancing the accuracy of all phenotype features, especially the fruit size and surface state, with accuracy values of 0.89 and 0.91, respectively, indicating its advantage in handling complex structural features. Tiny-Segformer, a lightweight Transformer-based model, significantly improved the accuracy across all features, especially fruit color and surface state, reaching 0.92 for both, demonstrating the advantage of the self-attention mechanism in global feature modeling. The proposed method outperformed all other models in feature extraction, with accuracy values of 0.92, 0.93, 0.94, and 0.95 for the various phenotype features, especially in surface state recognition. By combining edge perception mechanisms with global contextual information modeling, the proposed model significantly enhanced its detail capture, allowing it to more precisely identify the surface features of the fruit. From a theoretical analysis perspective, the differences in the experimental results were directly influenced by the model architectures. Both UNet and UNet++ adopt an encoder-decoder structure and fuse multi-scale features through skip connections, but UNet++ further optimizes skip connections, enhancing its ability to integrate multi-scale features, which led to better performance in fruit phenotype feature extraction. Mask R-CNN, as a multi-task learning model, leverages object detection mechanisms to effectively extract object boundaries and generate accurate segmentation masks. This ability enabled the model to achieve better results when handling surface state and complex fruit shape features. DeeplabV3+ uses atrous convolution and spatial pyramid pooling modules, which expand the receptive field and capture multi-scale contextual information, giving it an advantage in handling large objects and complex backgrounds. Tiny-Segformer combines self-attention mechanisms with convolution operations, allowing it to better capture long-range dependencies and efficiently utilize global information during feature extraction, resulting in improved performance in terms of fruit color and shape recognition. The proposed method, with the introduction of an edge perception mechanism, not only improved the segmentation accuracy of the surface state but also effectively captured subtle changes in fruit shape, demonstrating strong robustness and high precision when identifying complex fruit shapes and surface damage. These results indicate that model architecture innovation, enhanced feature extraction capabilities, and effective utilization of edge information are crucial factors in improving the accuracy of phenotype feature extraction.

4.4. Ablation Experiment with Different Attention for Phenotype Features

As shown in Table 5, the design of this experiment aimed to evaluate the impact of different attention mechanisms on model performance, particularly in the precise extraction of features and recognition of anomalies in fruit phenotype analysis tasks. Specifically, the experiment compared the performance of the standard self-attention mechanism, the channel and spatial attention mechanism (CBAM), and the proposed improved attention mechanism across various metrics. The objective of these comparisons was to clarify the ef-

fectiveness of different attention mechanisms in capturing image features, thereby verifying whether the proposed method could effectively enhance model performance, particularly in the tasks of fruit phenotype feature extraction and growth anomaly recognition.

Table 5. Ablation experiment with Different attention for phenotype feature.

Model	Precision	Recall	Accuracy	mIoU
Standard Self-Attention	0.76	0.72	0.74	0.71
CBAM	0.85	0.81	0.83	0.80
Proposed Method	0.95	0.91	0.93	0.92

From the experimental results, it can be observed that the standard self-attention mechanism exhibited relatively worse performance, with a precision of 0.76, recall of 0.72, accuracy of 0.74, and mIoU of 0.71. This suggests that while the standard self-attention mechanism can capture global information, its ability to focus on local features and details is limited. In comparison, the CBAM showed significant improvements across all metrics, with a precision of 0.85, recall of 0.81, accuracy of 0.83, and mIoU of 0.80. This indicates that the combination of channel and spatial attention mechanisms effectively enhanced the model's focus on different regions and features of an image, especially improving sensitivity to fruit features and anomaly areas. The proposed method, which combines edge perception mechanisms and global contextual information modeling, achieved the best results in terms of precision (0.95), recall (0.91), accuracy (0.93), and mIoU (0.92), demonstrating that the method effectively enhances feature capture in edge and complex regions, improving the precision of fruit phenotype feature extraction and growth anomaly recognition. From a theoretical perspective, the standard self-attention mechanism models global information by calculating the similarity between image features, but it lacks specialized attention to key local regions. As a result, it may overlook edge information and details when processing complex image features. The CBAM improves upon the standard self-attention mechanism by introducing attention mechanisms in both the channel and spatial dimensions, enhancing the model's attention to different channels and spatial regions. This improvement effectively boosts the model's ability to extract features, particularly in handling the details of fruit shapes and surface states. The proposed method further refines this approach by introducing an edge perception mechanism, which enables the network to focus on the edge regions of a fruit during training. This is especially crucial when dealing with surface damage or deformities, as it allows for precise identification of these complex regions. Mathematically, the standard self-attention mechanism typically models global information by calculating the relationships between each pixel in the input feature map, but its expression of local details, particularly the surface details of a fruit, is weak. The CBAM, by introducing attention mechanisms in both the channel and spatial dimensions, not only allows the model to focus on global features but also enables it to weight the important regions of an image, enhancing its ability to recognize details. The proposed method improves this further by incorporating an edge perception mechanism, which focuses more attention on edge regions, maintaining high segmentation accuracy even in cases where a fruit shape is complex, the surface is damaged, or boundaries are unclear. This design illustrates the advantage of the self-attention mechanism in combining both local and global features, significantly enhancing the model's robustness and precision, especially when dealing with complex image tasks.

4.5. Application in Agricultural Economics

The model proposed in this study, by integrating agricultural knowledge with deep learning techniques, has made significant contributions to the agricultural economy.

Through the precise extraction of fruit phenotypic features and the effective identification of growth anomalies, the model not only enhanced the ability to monitor fruit quality but also provided data support for quantifiable economic assessments in agricultural production. In the context of agricultural economics, the quality of fruit directly influences the market value and production efficiency. Therefore, accurate fruit quality assessment is crucial for improving agricultural productivity and economic benefits. By comprehensively analyzing the shape, surface condition, size, and color of a fruit, the model provides agricultural producers with detailed fruit quality information, enabling farmers to monitor crop growth in real time and take timely measures to optimize production processes and reduce losses. This model further contributes to economic optimization by reducing post-harvest losses through precise defect detection and classification, allowing for better sorting and market positioning of agricultural products.

Additionally, the incorporation of agricultural knowledge for anomaly detection allows the model to address the impact of environmental factors such as climate change and soil conditions on fruit growth, offering targeted warnings and recommendations. This not only helps improve the sustainability of agricultural production but also reduces the occurrence of pests and diseases, thus minimizing pesticide usage and promoting the development of green agriculture. By leveraging real-time phenotypic analysis, producers can adopt data-driven strategies to adjust cultivation practices dynamically, aligning resource investment with the predicted yield and market demand. Practical application of the model in agricultural economics has facilitated the optimization of resource allocation. By accurately identifying anomalies in fruit growth and pest occurrences, producers can precisely deploy fertilization, irrigation, and pest control measures, avoiding excessive resource use and further reducing production costs. The model's ability to integrate phenotypic analysis with environmental and economic parameters provides a foundation for cost-benefit assessments, allowing for more informed decision making regarding investment in precision farming technologies. In large-scale agricultural production, this intelligent fruit monitoring technology significantly increases labor productivity, reduces labor costs, and enhances the digitalization and automation levels of the agricultural supply chain, promoting the modernization of agriculture.

4.6. Future Work in Smart Agriculture

In future work, the focus will be placed on the practical application of the proposed model in real-world agricultural production. To achieve this, efforts will be directed toward optimizing the model for deployment on edge computing devices such as the Jetson Nano, enabling real-time processing with reduced computational complexity, as shown in Table 6.

Table 6. Performance on different hardware platforms.

Hardware Platform	Precision	Recall	Accuracy	mIoU	FPS
GPU platform (baseline)	0.95	0.91	0.93	0.92	47.1
Jetson Nano (lightweight version)	0.91	0.89	0.90	0.89	30.5
Huawei P40 pro (lightweight version)	0.88	0.85	0.85	0.83	18.3

By designing a lightweight version of the proposed edge transformer segmentation network, the feasibility of implementing the model in resource-constrained environments, such as automated agricultural machinery and intelligent monitoring systems, will be explored. This will allow for real-time fruit phenotype analysis and anomaly detection in the field, providing immediate feedback for decision making in agricultural management. Additionally, the integration of the model into unmanned aerial vehicles (UAVs) and robotic harvesting systems will be investigated to enhance precision agriculture practices.

These developments will further bridge the gap between theoretical advancements and practical applications, ensuring that the proposed methodology contributes to improving agricultural efficiency, reducing resource waste, and supporting intelligent agricultural decision making in real-world scenarios.

5. Conclusions

With the rapid development of intelligent agriculture, efficiently and accurately assessing fruit growth status and quality in apple production has become a critical factor in enhancing agricultural productivity and economic benefits. This study aimed to propose a deep learning-based approach for apple phenotype analysis and growth anomaly recognition by integrating instance segmentation, NLP, and innovative attention mechanisms. The method addresses the limitations of traditional fruit quality detection and anomaly recognition techniques, providing robust technical support for intelligent agriculture and agricultural economic analysis.

The proposed approach introduces several innovations. First, a comprehensive method combining instance segmentation and NLP is presented. By integrating image analysis and text parsing, the method accurately extracts fruit phenotype features, such as the fruit shape, color, and surface condition, while simultaneously utilizing agricultural textual data, including meteorological information and cultivation records, to identify growth anomalies. This multimodal data fusion overcomes the limitations of traditional image-based methods, enabling a holistic improvement in the accuracy of fruit quality detection and anomaly prediction from multiple perspectives. Moreover, this study introduced innovative edge attention modules and edge loss mechanisms, which enhance the model's focus on fruit edge regions and refine its handling of abnormal areas. These advancements significantly improve the model's performance in scenarios involving complex fruit morphology, surface damage, and growth anomalies. Through these innovative designs, the proposed method not only enhances the precision of fruit phenotype feature extraction but also provides more accurate and reliable data support for agricultural production decision making. The experimental results demonstrate that the proposed approach achieved significant improvements in accuracy and offers new perspectives and technological support for agricultural economic analysis.

Author Contributions: Conceptualization, Z.W., W.C., C.H. and C.L.; data curation, Y.Z. and Z.Z.; formal analysis, Y.Z., Y.Y. and X.D.; funding acquisition, C.L.; investigation, Y.Y.; methodology, Z.W., W.C. and C.H.; project administration, C.L.; resources, Y.Z. and Z.Z.; software, Z.W., W.C., C.H. and X.D.; supervision, C.L.; validation, Z.Z.; visualization, Y.Y. and X.D.; writing—original draft, Z.W., W.C., C.H., Y.Z., Z.Z., Y.Y., X.D. and C.L. Z.W., W.C. and C.H. contributed evenly to this work. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by The National Key Research and Development Program of China (2024YFC2607600).

Data Availability Statement: The data presented in this study are available on Github at https://github.com/user837498178/apple-agriculture (accessed on 26 January 2025).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zhang, Y.; Cao, Y.F.; Huo, H.L.; Xu, J.Y.; Tian, L.M.; Dong, X.G.; Dan, Q.; Chao, L. An assessment of the genetic diversity of pear (*Pyrus* L.) Germplasm Resour. Based Fruit Phenotypic Trait. *J. Integr. Agric.* **2022**, *21*, 2275–2290. [CrossRef]
2. Medda, S.; Mulas, M. Fruit quality characters of myrtle (*Myrtus communis* L.) selections: Review of a domestication process. *Sustainability* **2021**, *13*, 8785. [CrossRef]
3. Zhang, Y.; Wa, S.; Sun, P.; Wang, Y. Pear defect detection method based on resnet and dcgan. *Information* **2021**, *12*, 397. [CrossRef]

4. Zhang, Y.; Lv, C. TinySegformer: A lightweight visual segmentation model for real-time agricultural pest detection. *Comput. Electron. Agric.* **2024**, *218*, 108740. [CrossRef]
5. Zhang, Y.; Liu, X.; Wa, S.; Chen, S.; Ma, Q. GANsformer: A detection network for aerial images with high performance combining convolutional network and transformer. *Remote Sens.* **2022**, *14*, 923. [CrossRef]
6. Cui, M.; Pham, M.D.; Hwang, H.; Chun, C. Flower development and fruit malformation in strawberries after short-term exposure to high or low temperature. *Sci. Hortic.* **2021**, *288*, 110308. [CrossRef]
7. Gu, W.; Bai, S.; Kong, L. A review on 2D instance segmentation based on deep neural networks. *Image Vis. Comput.* **2022**, *120*, 104401. [CrossRef]
8. Kazakova, M.A.; Sultanova, A.P. Analysis of natural language processing technology: Modern problems and approaches. *Adv. Eng. Res.* **2022**, *22*, 169–176. [CrossRef]
9. Bergen, L.; O'Donnell, T.; Bahdanau, D. Systematic generalization with edge transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 1390–1402.
10. Zhao, P.; Wang, W.; Liu, H.; Han, M. Recognition of the agricultural named entities with multifeature fusion based on albert. *IEEE Access* **2022**, *10*, 98936–98943. [CrossRef]
11. Andriyanov, N. Combining Text and Image Analysis Methods for Solving Multimodal Classification Problems. *Pattern Recognit. Image Anal.* **2022**, *32*, 489–494. [CrossRef]
12. Zhang, Y.; Wa, S.; Liu, Y.; Zhou, X.; Sun, P.; Ma, Q. High-accuracy detection of maize leaf diseases CNN based on multi-pathway activation function module. *Remote Sens.* **2021**, *13*, 4218. [CrossRef]
13. Zhang, Y.; Wa, S.; Zhang, L.; Lv, C. Automatic plant disease detection based on tranvolution detection network with GAN modules using leaf images. *Front. Plant Sci.* **2022**, *13*, 875693. [CrossRef] [PubMed]
14. Luo, Z.; Yang, W.; Yuan, Y.; Gou, R.; Li, X. Semantic segmentation of agricultural images: A survey. *Inf. Process. Agric.* **2024**, *11*, 172–186. [CrossRef]
15. Anand, T.; Sinha, S.; Mandal, M.; Chamola, V.; Yu, F.R. AgriSegNet: Deep aerial semantic segmentation framework for IoT-assisted precision agriculture. *IEEE Sensors J.* **2021**, *21*, 17581–17590. [CrossRef]
16. Zhang, H.; Peng, Q. PSO and K-means-based semantic segmentation toward agricultural products. *Future Gener. Comput. Syst.* **2022**, *126*, 82–87. [CrossRef]
17. Su, D.; Kong, H.; Qiao, Y.; Sukkarieh, S. Data augmentation for deep learning based semantic segmentation and crop-weed classification in agricultural robotics. *Comput. Electron. Agric.* **2021**, *190*, 106418. [CrossRef]
18. Zhang, Y.; He, S.; Wa, S.; Zong, Z.; Liu, Y. Using generative module and pruning inference for the fast and accurate detection of apple flower in natural environments. *Information* **2021**, *12*, 495. [CrossRef]
19. Lin, X.; Wa, S.; Zhang, Y.; Ma, Q. A dilated segmentation network with the morphological correction method in farming area image Series. *Remote Sens.* **2022**, *14*, 1771. [CrossRef]
20. Zhou, X.; Chen, S.; Ren, Y.; Zhang, Y.; Fu, J.; Fan, D.; Lin, J.; Wang, Q. Atrous Pyramid GAN Segmentation Network for Fish Images with High Performance. *Electronics* **2022**, *11*, 911. [CrossRef]
21. Kaur, R.; Singh, S. A comprehensive review of object detection with deep learning. *Digit. Signal Process.* **2023**, *132*, 103812. [CrossRef]
22. Diwan, T.; Anirudh, G.; Tembhurne, J.V. Object detection using YOLO: Challenges, architectural successors, datasets and applications. *Multimed. Tools Appl.* **2023**, *82*, 9243–9275. [CrossRef]
23. Amit, Y.; Felzenszwalb, P.; Girshick, R. Object Detection. In *Computer Vision: A Reference Guide*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 875–883.
24. Bharati, P.; Pramanik, A. Deep Learning Techniques—R-CNN to Mask R-CNN: A Survey. In *Computational Intelligence in Pattern Recognition: Proceedings of CIPR 2019*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 657–668.
25. Jiang, P.; Ergu, D.; Liu, F.; Cai, Y.; Ma, B. A Review of Yolo algorithm developments. *Procedia Comput. Sci.* **2022**, *199*, 1066–1073. [CrossRef]
26. Khan, S.; AlSuwaidan, L. Agricultural monitoring system in video surveillance object detection using feature extraction and classification by deep learning techniques. *Comput. Electr. Eng.* **2022**, *102*, 108201. [CrossRef]
27. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1137–1149. [CrossRef]
28. Mo, Y.; Wu, Y.; Yang, X.; Liu, F.; Liao, Y. Review the state-of-the-art technologies of semantic segmentation based on deep learning. *Neurocomputing* **2022**, *493*, 626–646. [CrossRef]
29. Strudel, R.; Garcia, R.; Laptev, I.; Schmid, C. Segmenter: Transformer for Semantic Segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 7262–7272.
30. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th international Conference, Proceedings, Part III 18, Munich, Germany, 5–9 October 2015; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.

31. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
32. De Boer, P.T.; Kroese, D.P.; Mannor, S.; Rubinstein, R.Y. A tutorial on the cross-entropy method. *Ann. Oper. Res.* **2005**, *134*, 19–67. [CrossRef]
33. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
34. Xu, S.; Shen, J.; Wei, Y.; Li, Y.; He, Y.; Hu, H.; Feng, X. Automatic plant phenotyping analysis of Melon (*Cucumis melo* L.) Germplasm Resour. Using Deep Learn. Methods Comput. Vision. *Plant Methods* **2024**, *20*, 166. [CrossRef] [PubMed]
35. Wang, C.; Du, P.; Wu, H.; Li, J.; Zhao, C.; Zhu, H. A cucumber leaf disease severity classification method based on the fusion of DeepLabV3+ and U-Net. *Comput. Electron. Agric.* **2021**, *189*, 106373. [CrossRef]
36. Min, B.; Ross, H.; Sulem, E.; Veyseh, A.P.B.; Nguyen, T.H.; Sainz, O.; Agirre, E.; Heintz, I.; Roth, D. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Comput. Surv.* **2023**, *56*, 1–40. [CrossRef]
37. Rezayi, S.; Liu, Z.; Wu, Z.; Dhakal, C.; Ge, B.; Dai, H.; Mai, G.; Liu, N.; Zhen, C.; Liu, T.; et al. Exploring new frontiers in agricultural nlp: Investigating the potential of large language models for food applications. *IEEE Trans. Big Data* **2024**. . [CrossRef]
38. Espinoza, S.; Aguilera, C.; Rojas, L.; Campos, P.G. Analysis of Fruit Images With Deep Learning: A Systematic Literature Review and Future Directions. *IEEE Access* **2023**, *12*, 3837–3859. [CrossRef]
39. Duan, J.L.; Lai, L.Q.; Yang, Z.; Luo, Z.J.; Yuan, H.T. Multi-feature language-image model for fruit quality image classification. *Comput. Electron. Agric.* **2024**, *227*, 109462. [CrossRef]
40. Apostolopoulos, I.D.; Tzani, M.; Aznaouridis, S.I. A general machine learning model for assessing fruit quality using deep image features. *AI* **2023**, *4*, 812–830. [CrossRef]
41. Ibtissam, B. Automatic Date Fruit Sorting System Based on Machine Learning and Visual Features. Ph.D. Thesis, University of Biskra, Biskra, Algeria, 2024.
42. Ahmed, A.A.; Reddy, G.H. A mobile-based system for detecting plant leaf diseases using deep learning. *AgriEngineering* **2021**, *3*, 478–493. [CrossRef]
43. Vishnoi, V.K.; Kumar, K.; Kumar, B.; Mohan, S.; Khan, A.A. Detection of apple plant diseases using leaf images through convolutional neural network. *IEEE Access* **2022**, *11*, 6594–6609. [CrossRef]
44. Khan, A.I.; Quadri, S.; Banday, S.; Shah, J.L. Deep diagnosis: A real-time apple leaf disease detection system based on deep learning. *Comput. Electron. Agric.* **2022**, *198*, 107093. [CrossRef]
45. Rahman, S.U.; Alam, F.; Ahmad, N.; Arshad, S. Image processing based system for the detection, identification and treatment of tomato leaf diseases. *Multimed. Tools Appl.* **2023**, *82*, 9431–9445. [CrossRef]
46. Zhou, Z.; Rahman Siddiquee, M.M.; Tajbakhsh, N.; Liang, J. Unet++: A Nested U-Net Architecture for Medical Image Segmentation. In Proceedings of the Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Proceedings 4, Granada, Spain, 20 September 2018; Springer: Berlin/Heidelberg, Germany, 2018; pp. 3–11.