agriculture

MDPI

Article

SwinLabNet: Jujube Orchard Drivable Area Segmentation Based on Lightweight CNN-Transformer Architecture

Mingxia Liang [1,2], Longpeng Ding [1,2], Jiangchun Chen [1,2], Liming Xu [3], Xinjie Wang [4], Jingbin Li [1,2,*] and Hongfei Yang [1,2,5,*]

1 College of Mechanical and Electrical Engineering, Shihezi University, Shihezi 832003, China; 20222109049@stu.shzu.edu.cn (M.L.)
2 Xinjiang Production and Construction Corps Key Laboratory of Modern Agricultural Machinery, Shihezi 832003, China
3 Mechanical Engineering and Power Engineering, Shanghai Jiao Tong University, Shanghai 200030, China; limingxu@sjtu.edu.cn
4 College of Economics and Management, Shihezi University, Shihezi 832099, China
5 College of Instrumentation and Electrical Engineering, Jilin University, Changchun 130061, China
* Correspondence: lijingbin@shzu.edu.cn (J.L.); yanghf20@mails.jlu.edu.cn (H.Y.)

Abstract: Identifying drivable areas between orchard rows is crucial for intelligent agricultural equipment. However, challenges remain in this field's accuracy, real-time performance, and generalization of deep learning models. This study proposed the SwinLabNet model in the context of jujube orchards, an innovative network model that utilized a lightweight CNN-transformer hybrid architecture. This approach optimized feature extraction and contextual information capture, effectively addressing long-range dependencies, global information acquisition, and detailed boundary processing. After training on the jujube orchard dataset, the SwinLabNet model demonstrated significant performance advantages: training accuracy reached 97.24%, the mean Intersection over Union (IoU) was 95.73%, and the recall rate was as high as 98.36%. Furthermore, the model performed exceptionally well on vegetable datasets, highlighting its generalization capability across different crop environments. This study successfully applied the SwinLabNet model in orchard environments, providing essential support for developing intelligent agricultural equipment, advancing the identification of drivable areas between rows, and laying a solid foundation for promoting and applying intelligent agrarian technologies.

Keywords: drivable area identification; SwinLabNet network model; jujube orchard environment; agricultural Intelligence

Citation: Liang, M.; Ding, L.; Chen, J.; Xu, L.; Wang, X.; Li, J.; Yang, H. SwinLabNet: Jujube Orchard Drivable Area Segmentation Based on Lightweight CNN-Transformer Architecture. *Agriculture* **2024**, *14*, 1760. https://doi.org/10.3390/agriculture14101760

Academic Editor: Bin Xie

Received: 5 September 2024
Revised: 27 September 2024
Accepted: 29 September 2024
Published: 5 October 2024

1. Introduction

Forestry and fruit industries play a crucial role in China's rural economy, with jujube being one of the main varieties, significantly boosting farmers' income. However, due to the complex orchard environment and variability in plant structure, the harvesting process faces challenges such as low mechanization, high labor intensity, and high costs [1]. Therefore, addressing these technical challenges and enhancing the applicability and efficiency of intelligent harvesting equipment in jujube production is an urgent issue. The intelligentization of agricultural machinery is essential for improving productivity and economic returns. In particular, visual navigation technology, representing the cutting-edge trend in intelligent agricultural machinery, guides robots to work efficiently through image processing, reducing labor intensity [2–5].

Currently, most research directions combine Global Navigation Satellite Systems (GNSS), Light Detection and Ranging (LiDAR), and visual sensors, along with multi-sensor fusion navigation methods [6]. The accuracy of existing GNSS positioning technology is

limited in obstructed environments such as orchards [7]. LiDAR-based navigation systems face limitations in agricultural applications due to high costs and the complexity of data processing [8]. Machine vision has become a key research focus, given visual sensors' cost-effectiveness, flexibility, and ability to mimic the human eye. Accurate segmentation of drivable areas is fundamental to achieving precise navigation in orchard environments [9,10]. Traditional image segmentation techniques first enhance features through image preprocessing, using color, texture, or shape to distinguish scenes. Algorithms are then applied to segment meaningful regions of the image [11], ultimately enabling the identification of navigation paths and understanding of the environment, which guides orchard machinery to move safely and efficiently in complex settings [12,13]. However, in the unstructured environment of jujube orchards, traditional methods struggle to achieve accurate segmentation due to factors such as the irregularity of jujube bands, varying lighting conditions, and complex backgrounds. In recent years, the rapid development of deep learning technology has led to its widespread application in navigation path recognition [14–16]. Deep convolutional neural networks have demonstrated exceptional performance in image classification, object detection, and semantic segmentation [17–20]. In particular, deep learning-based semantic segmentation has become one of the critical methods for analyzing and recognizing complex image scenes [21]. Semantic segmentation assigns each pixel in an image to a corresponding semantic category, enabling a fine-grained understanding of the image [22]. This approach not only extracts deep features of the image but also captures high-level semantic information, resulting in more accurate recognition in complex scenarios [23].

Yu et al. [24] studied five deep learning-based computer vision methods for field navigation line extraction across different field scenarios, achieving an average segmentation accuracy of 84.87%. Zhang et al. [25] proposed the Fast-Unet model, which effectively enhanced the recognition accuracy of multiscale features through an encoder-decoder structure and ASPP module. This model was trained on a peach dataset and successfully transferred to orange and kiwi datasets, with mean Intersection over Union (MIoU) values of 97.9%, 98.7%, and 95.6%, respectively. Zheng et al. [26] introduced an improved lightweight YOLOX-Nano architecture for detecting jujube tree root points, achieving an mAP of 84.08%, making it suitable for embedded deployment. Yang et al. [27] proposed a visual navigation path extraction method based on neural networks and pixel scanning. The study showed that segmentation accuracy exceeded 92% under different lighting conditions. Li et al. [28] proposed a fast U-net model, where the improved model reduced parameters by 65.86% and enhanced prediction performance by 97.39%. Cao et al. [29] introduced an improved Enet model that effectively utilized residual flow to extract low-dimensional boundary information, significantly improving boundary localization and segmentation accuracy between crop rows in fields. Zhang et al. [30] proposed an enhanced semantic segmentation method, utilizing a modified ResNet network as the backbone, combined with stripe pooling and hybrid pooling modules, achieving 95.6% accuracy and 77.6% MIoU. Baheti et al. [31] introduced an enhanced DeepLabV3+ method, using a low dilation rate ASPP module for dense flow prediction and expanding the Xception network as the backbone for feature extraction. On the Indian Driving Dataset, this approach achieved MIoU scores of 68.41 and 86.75 for unknown test data with Level 3 and Level 1 labels, respectively. Although these studies have achieved some success in image semantic segmentation at the pixel level using deep convolutional neural networks, they struggle to accurately segment the entire drivable area when dealing with long-distance jujube belt datasets with blurred boundaries and irregular shapes [32]. This presents challenges for the subsequent extraction of navigation lines.

To address the issues identified in the earlier studies, this paper focuses on the jujube belt area (drivable area) in unstructured jujube orchards. It proposes a segmentation algorithm model based on a lightweight CNN-Transformer architecture for drivable area identification. The contributions of this paper are primarily concentrated in the following aspects:

(1) A lightweight CNN-Transformer hybrid architecture, SwinLabNet, is proposed, which includes both encoding and decoding structures.
(2) The SwinASPP feature extraction module is introduced to enhance the fine-grained segmentation of drivable areas in jujube belts by expanding the receptive field and capturing more contextual semantic information, effectively adapting to the complex orchard environment.
(3) The improved model effectively extracts drivable areas in jujube belts and demonstrates good generalization performance on vegetable datasets.

The structure of this paper is organized as follows: Section 1 is the Introduction, Section 2 covers the Materials and Methods; Section 3 describes the Experiment and Result Analysis; and Section 4 concludes the paper.

2. Materials and Methods

2.1. Experimental Equipment and Parameter ConFigureuration

The experimental environment used Windows 10, with a CPU of AMD Ryzen 7 5800H and a GPU of NVIDIA GeForce RTX 3050 (Lenovo's Legion series, Beijing, China). Based on the preliminary experimental results. The initial learning rate was set to 0.01, with a weight decay coefficient of 0.007, a batch size of 4, and 100 iterations. During training, the network model was updated using Stochastic Gradient Descent (S.G.D.) to learn and adjust network parameters with a cosine annealing learning rate decay. The Adam optimization algorithm was employed to optimize the weight updates.

2.2. Semantic Segmentation of Drivable Areas Based on Neural Network

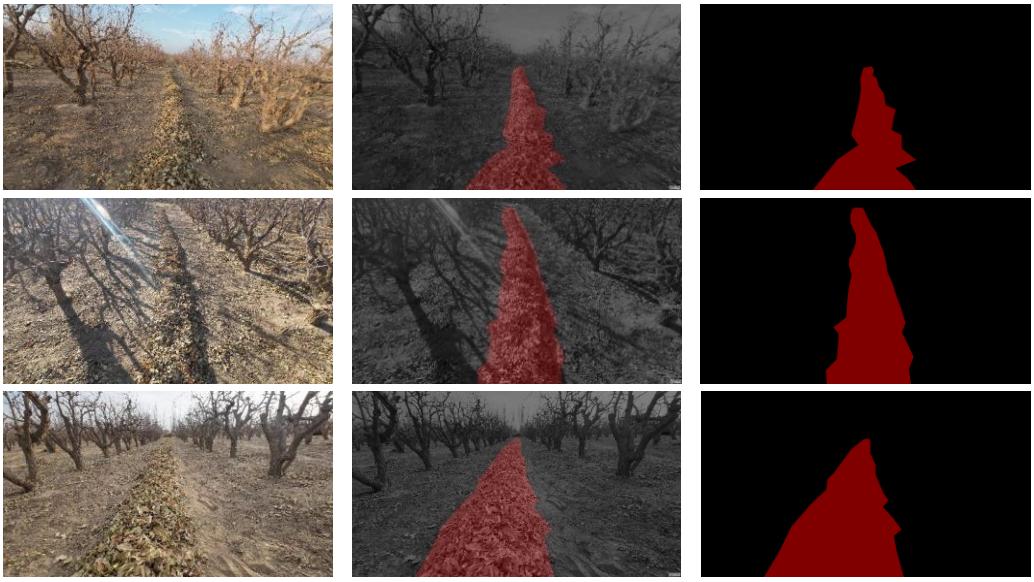
2.2.1. Dataset Acquisition

The experimental data were collected in November 2023 from a dwarf dense planting jujube orchard in Kunyu, Xinjiang. During the data collection, a stabilized camera was used to simulate the normal working state of a vehicle. To capture the multi-angle features of the jujube and ensure comprehensive and complete data collection, images were taken at angles of 10°, 30°, and 45°. The image resolution was 1920 × 1080, with a frame rate of 60 fps. The video was converted into images using OpenCV 4.3.3. The programming language was Python 3.6, and the code was edited in PyCharm Community Edition 2022.3.2. The program was executed on the Windows 10 operating system.

In line with the research objectives, this study collected only image data from the drivable areas between orchard rows, excluding the orchard headlands. To enhance sample diversity, we collected images of the jujube belt between orchard rows under various working conditions and lighting scenarios to better reflect natural environments. After processing, a total of 1550 jujube belt images were obtained.In the experiment, the drivable areas of the jujube belt in the orchard were annotated using Labelme software 3.16.7. The drivable areas were marked in red, with other regions labeled as background. Figure 1 shows the annotation results from different times and angles.

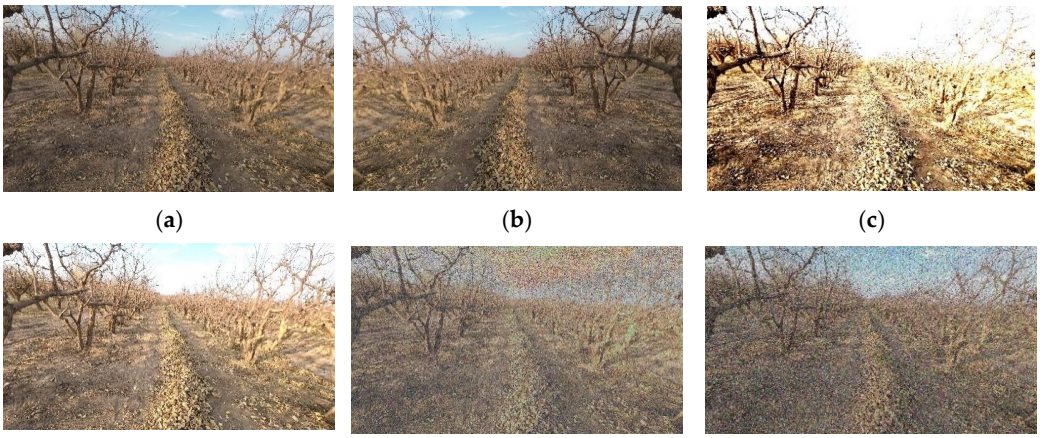
2.2.2. Dataset Augmentation

To validate the model's robustness and generalization ability and to reduce its sensitivity to noise, varying perspectives, and lighting changes, it is necessary to preprocess images to augment the dataset, enhance model performance, and prevent overfitting. The same jujube belt region exhibits variations in size, position, orientation, and brightness under different shooting conditions. To improve the generalization ability of the network model, data augmentation is applied to the dataset. The dataset was augmented by horizontal mirroring, color transformations (contrast, brightness), and adding Gaussian and salt-and-pepper noise to improve the model's robustness and accuracy, as shown in Figure 2. Mirroring aids in determining the navigation path between orchard rows, while random noise helps the model adapt to uneven natural lighting conditions. The augmented dataset contains a total of 3800 images, which were split into training and test sets in a 9:1 ratio.



(**a**) (**b**) (**c**)

Figure 1. Labeled Results of the Dataset at Different Periods. (**a**) Original Image, (**b**) Overlay of Mask Image and Real-World Image, (**c**) Mask Image.



(**a**) (**b**) (**c**)

(**d**) (**e**) (**f**)

Figure 2. Data Augmentation. (**a**) Original Image, (**b**) Mirroring, (**c**) Color Transformation, (**d**) Color Jitter, (**e**) Gaussian Noise, (**f**) Salt-and-Pepper Noise.

2.3. Construction of a Model for Identifying Drivable Areas between Rows in Unstructured Jujube Orchards

The extraction of drivable jujube belts in unstructured jujube orchards involves two labels: jujube belts and background. Due to the characteristics of the jujube belt regions, such as long and blurred boundaries, complex and dispersed information, and irregular shapes and sizes, researchers have proposed several CNN-based neural network models to address the multiscale segmentation problem and to extract drivable jujube belt regions of varying sizes and shapes [33,34]. However, these models, which rely on local information from surrounding pixels, fail to establish dependencies between features or capture dense contextual information, making them ineffective in achieving satisfactory results when dealing with jujube belt regions that have extended, irregular boundaries and require global consistency or long-range contextual understanding for accurate classification in complex scenarios. This study proposes the SwinLabNet model based on a lightweight CNN-Transformer hybrid architecture. SwinLabNet enhances segmentation accuracy by combining high-level semantic information with low-level spatial details through feature

aggregation in the encoder-decoder structure, integrating upsampling and downsampling information. By incorporating the lightweight MobileNetV3 backbone, the innovative SwinASPP module, and focal loss optimization, SwinLabNet achieves efficient and high-precision semantic segmentation, excelling in capturing multiscale contextual information and detailed features. The model structure is shown in Figure 3.

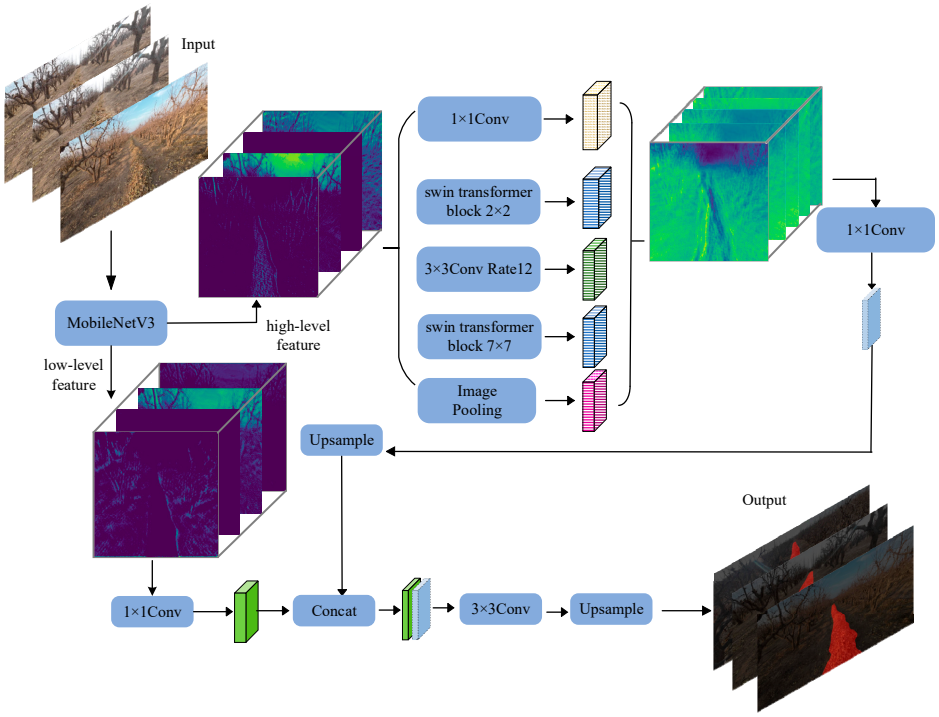


Figure 3. Proposed SwinLabNet Architecture.

The jujube belt images are first processed through the improved lightweight backbone, MobileNetV3-ECA, to extract high-level and low-level features. Subsequently, the high-level features are processed by the modified SwinASPP module: a Swin Transformer with a window size of 2 (W-MSA) is used to extract multiscale information and capture dependencies between distant pixels; dilated convolution with a dilation rate of 12 is employed for mid-scale feature extraction to enhance contextual information; and a Swin Transformer with a window size of 7 (SW-MSA) achieves cross-window information exchange through window shifting, further enriching the contextual information. Replacing traditional dilated convolution with Swin Transformers of window sizes 2 and 7 expands the receptive field. It enhances the model's ability to capture long-range dependencies and global semantic information, leading to more refined and accurate results in segmenting drivable areas in densely planted dwarf jujube orchards.

Finally, the high-level features processed by the improved SwinASPP module are fused with the low-level features, and the resulting feature maps are input into the prediction module to achieve the segmentation of drivable areas in the jujube belt. The following will introduce the structural framework and principles of each module separately.

2.3.1. Lightweight Backbone Network MobileNetV3-ECA Module

To meet the real-time navigation and subsequent edge deployment requirements in unstructured jujube orchards while ensuring segmentation accuracy and reducing model complexity and parameter count, this study employs a lightweight MobileNetV3 network as the backbone in the feature extraction module, as illustrated in Figure 4. Considering the blurred boundaries of the jujube belt images in the target area, the presence of a complex background, and detailed features, the Efficient Channel Attention (ECA) module was used to replace the Squeeze-and-Excitation (SE) module in the model. This improves the

network's ability to extract target feature maps and focus on continuous regional features. Particularly for images with high noise, significant viewpoint variations, and differing lighting conditions, this mechanism helps the model concentrate on key features of the jujube belt navigable area segmentation task, reducing the impact of interference.This allows the network to adaptively learn the inter-channel correlations during training and apply them to channel weighting in each feature map.

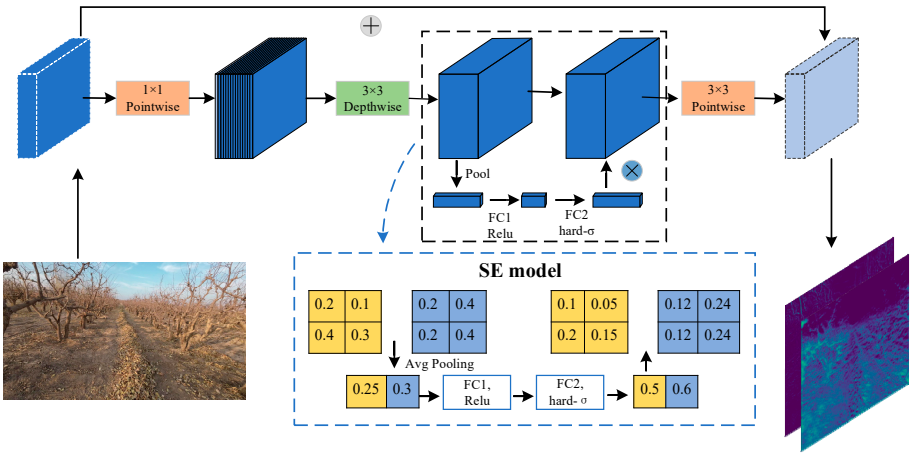


Figure 4. MobileNetV3 Unit.

Unlike the S.E. attention mechanism, the core idea of the E.C.A. module is to effectively capture global correlations between channels to emphasize important channel information while avoiding high computational cost and increased model complexity. Figure 5 shows the structure of the E.C.A. module. This module uses a 1×1 convolution layer after the global average pooling layer, eliminating the fully connected layer, thereby avoiding dimensionality reduction and effectively capturing cross-channel interaction information. The E.C.A. module employs a 1D convolution layer followed by the σ activation function to enhance the model's ability to interact with cross-channel details to generate the attention weight vector A. The vector A is then multiplied element-wise with the original feature map χ to adjust the channel attention, resulting in the recalibrated feature map Y_{cij}, calculated as follows:

$$A = \sigma(\text{Conv1D}(\text{GPA}(\chi))) \tag{1}$$

$$\begin{gathered} B_{cij} = A_c \cdot \chi_{cij} \\ \forall c \in \{1, .., C\}, i \in \{1, .., H\}, j \in \{1, .., W\} \end{gathered} \tag{2}$$

In the equation, A represents the attention weight vector; Conv1D denotes the 1D convolution layer; GAP stands for global average pooling; χ represents the original feature vector; B_{cij} is the recalibrated feature map; A_c refers to the element in the attention vector A corresponding to the c-th channel; χ_{cij} represents the value of the original input feature map χ at the c-th channel and position (i,j); and B_{cij} represents the value at the same position after channel attention adjustment.

The convolution kernel size in the E.C.A. module adaptively changes based on a function, allowing layers with more channels to perform more extensive cross-channel interactions. The adaptive 1D convolution primarily adjusts the size of the input channel C through the parameter K. The specific functional relationship is as follows:

$$K = \varphi(C) = \left| \frac{\log_2 C + b}{\gamma} \right|_{odd} \tag{3}$$

In this equation, $|t|_{odd}$ represents the nearest bizarre number to t, and γ and b are user-defined parameters with default values of 2 and 1, respectively [35].

During the segmentation process, the network can better focus on continuous jujube belt region features while suppressing background interference, enabling adaptive feature mapping to the target channels. Consequently, even under varying lighting conditions and complex backgrounds, the network can effectively capture the visual features of the Jujube belt region, resulting in more transparent and distinct segmentation boundaries and improved accuracy in segmenting the Jujube belt area. The Hard-Swish activation function is also introduced to reduce computational load and enhance performance. The definition of this activation function is as follows:

$$h - swish[x] = x\frac{ReLU6(x+3)}{6} \tag{4}$$

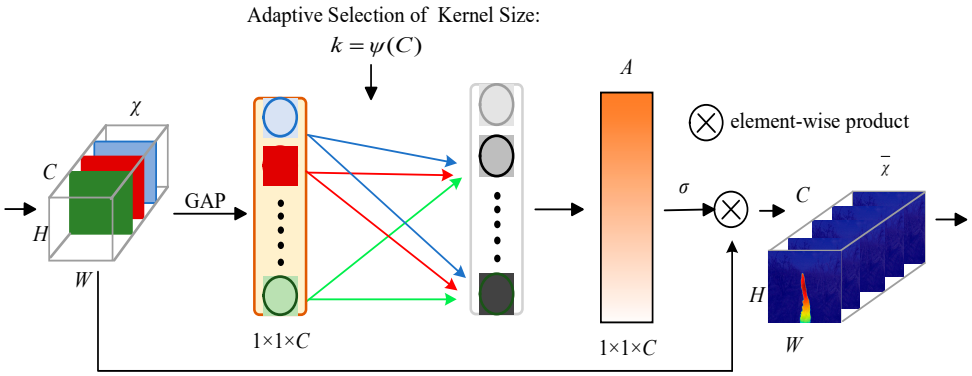


Figure 5. Structure diagram of the E.C.A. attention mechanism.

2.3.2. SwinASPP

For the segmentation task of jujube belt image datasets in dwarf densely planted jujube Orchards, this paper proposes a new SwinASPP module to address the challenges of low segmentation accuracy due to the small proportion, narrow shape, blurred boundaries, and complex internal information of the segmented regions in jujube belt images. In existing ASPP models, atrous convolution is limited to local areas when extracting contextual information, making it difficult to fully capture dense contextual semantic information, which affects the correlation of jujube belt image features and segmentation accuracy. Moreover, high dilation rates in atrous convolutions within higher-level network structures significantly reduce sampling density, potentially causing the "checkerboard effect". The dilated convolutions with different dilation rates lack dynamic interaction, making it difficult to effectively extract deep semantic content from low-resolution feature maps, which hinders the segmentation of complex jujube belt regions.

To address the abovementioned issues, this paper introduces an improved ASPP module by incorporating the Swin Transformer structure, as illustrated in Figure 6. The SwinASPP module includes a Swin Transformer block with a window size of 2 for extracting local details such as edges and textures. A dilated convolution layer with a dilation rate of 12 is used to extract medium- to large-scale contextual features through a larger receptive field, which aids in capturing the overall shape of objects and background information. Additionally, a Swin Transformer block with a window size of 7 captures the global structure and correlations in the jujube belt images.

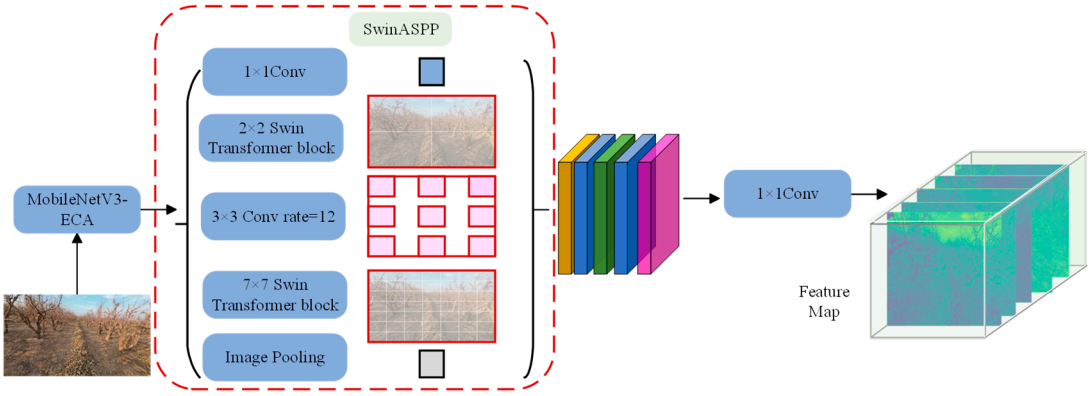


Figure 6. SwinASPP module.

The Swin Transformer block is a hierarchical structure that employs sliding windows to capture dependencies between distant pixels. This approach is efficient for images like those in the jujube belt dataset, where segmentation regions have extended, blurred boundaries and complex textures. By leveraging spatial information, the block improves segmentation accuracy. The specific structure is shown in Figure 7. The SwinASPP module refines the boundary features of the jujube belt region using W-MSA and SW-MSA self-attention mechanisms, enabling parallel extraction and cascading integration of features at different scales. Each module focuses on feature extraction at different scales, and by exchanging information across windows, it captures and integrates more extensive contextual information, leading to more accurate segmentation of the jujube belt region from the background. The model effectively leverages multiscale feature information in the jujube belt images by cascading or parallelly combining these modules within the network. The specific structure is shown in Figure 8. Based on this shifted window partitioning method, consecutive Swin Transformer blocks can be represented as follows:

$$\hat{Z}^{\iota} = W - MSA(LN(z^{\iota-1})) + z^{\iota-1}, \tag{5}$$

$$Z^{\iota} = MLP(LN(\hat{z}^{\iota})) + \hat{z}^{\iota}, \tag{6}$$

$$\hat{Z}^{\iota+1} = SW - MSA(LN(z^{\iota})) + z^{\iota}, \tag{7}$$

$$\hat{Z}^{\iota+1} = MLP(LN(\hat{z}^{\iota+1})) + \hat{z}^{\iota+1}, \tag{8}$$

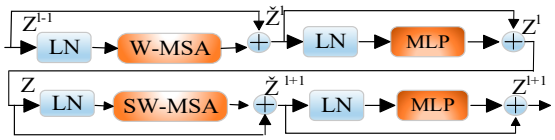


Figure 7. Swin Transformer block.

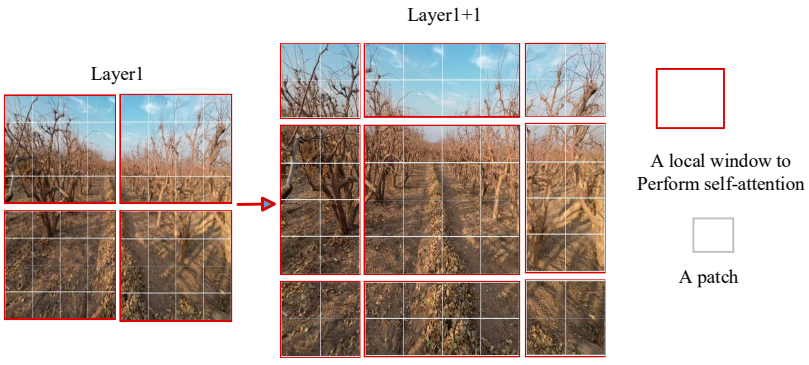


Figure 8. Shifted Window in Swin Transformer.

The window partitioning mechanism enables the SwinASPP module to handle complex segmentation tasks in unstructured jujube orchards, including drivable areas and background segmentation. This approach enhances segmentation accuracy and boundary clarity. It demonstrates strong perception and generalization capabilities, even when dealing with challenges such as extended, blurred boundaries, complex information, dispersed distribution, and irregular shapes of the jujube belt regions.

2.3.3. Loss Function Design

The dataset contains significantly more background pixels than drivable area pixels in this study. This imbalance biases the loss function toward the background pixels, leading the model to misclassify the drivable area as background, thus reducing the accuracy of the jujube belt segmentation.

Considering the characteristics of the jujube belt dataset in dwarf dense jujube orchards, this study is based on the multi-head attention mechanism. It improves the loss function by introducing a Combined Loss (C.L). This Combined Loss integrates Focal Loss (F.L.) [36] and Cross-Entropy Loss (CE Loss) [37], aiming to enhance the model's classification performance on imbalanced datasets. Focal Loss, based on cross-entropy, uses a dynamic scaling factor to reduce the weight of easily classified samples, thereby focusing attention on harder-to-classify samples. The specific function is expressed as follows:

$$L_{CE} = -\sum_{i=1}^{N} y_i \log(p_i) \tag{9}$$

$$L_{Focal} = -\alpha(1 - p_i)^{\gamma} \log(p_i) \tag{10}$$

$$L_{Combined} = -\lambda_{CE} \cdot L_{CE} + \lambda_{Focal} \cdot L_{Focal} \tag{11}$$

L_{CE} represents the Cross-Entropy Loss, L_{Focal} denotes the Focal Loss, and $L_{Combined}$ indicates the Combined Loss. In this context, y_i represents the accurate label, p_i is the model's predicted probability, α is the balancing factor used to balance positive and negative samples, γ is the modulating factor used to control the weight of easy and hard samples, λ_{CE} is the weight of the Cross-Entropy Loss, and λ_{Focal} is the weight of the Focal Loss.

The improved hybrid loss function combines Cross Entropy Loss and Focal Loss. It preserves the stability of Cross-Entropy Loss across the dataset while using Focal Loss to enhance focus on hard-to-classify samples during training. Cross-Entropy Loss applies a higher loss value to easily distinguishable samples to ensure correct classification, while Focal Loss assigns a lower weight to these samples, minimizing their impact on the overall Loss. For difficult samples, Cross-Entropy Loss generates a loss value, while Focal Loss further amplifies this loss, increasing its weight in the overall loss. This helps the model to better handle challenging samples.

3. Experiment and Result Analysis

3.1. Evaluation Metrucs

This experiment aims to evaluate the model's performance in the jujube belt segmentation task, using the following metrics for assessment and comparison: Mean Intersection over Union (MIoU), Precision (Pr), and Recall (Re). The specific formulas for these calculations are shown below:

$$MIoU = \frac{1}{k+1} \sum_{i=0}^{k} \frac{TP}{FN + FP + TP} \tag{12}$$

$$Pr = \frac{TP}{TP + FP} \tag{13}$$

$$Re = \frac{TP}{TP + FN} \tag{14}$$

3.2. Influence of Different Loss Functions on Experimental Results in an Improved Model

To investigate the impact of the improved loss function on the performance of the SwinLabNet network in the navigable area segmentation task, we compared the improved Combined Loss with CE Loss and Focal Loss. Under the same training conditions, the model was trained for 100 epochs with only the loss function changed. The comparison results are shown in Figure 9. As shown in the figure, with the increase in iterations, the loss function of the training set gradually decreases. The Combined Loss demonstrates a faster convergence speed and smoother convergence range compared to CE Loss and Focal Loss. Furthermore, the final convergence loss values of the three loss functions indicate that, in cases of class imbalance, the Combined Loss is better suited as the loss function for the model.

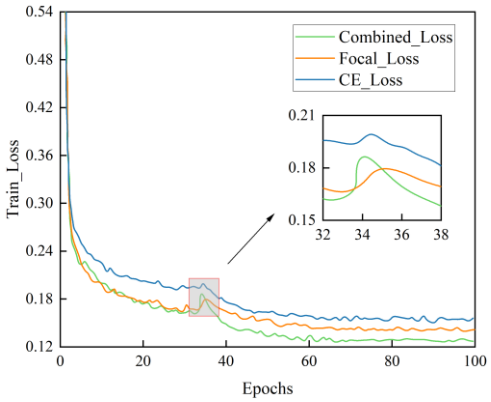


Figure 9. Comparison of loss values between three loss functions in the same conditions.

3.3. Comparison of Lightweight Backbone Network Performance

This study employs MobileNetV3 to replace the original backbone network, Xception, to achieve lightweight model improvements. To validate the effectiveness of the method, comparative experiments were conducted using four representative backbone networks: Xception, GhostNet, MobileNetV3, and MobileNetV3-ECA, under the same conditions. Aside from the changes in the backbone network, other parameters remained constant, and the original loss function was utilized for testing. Detailed experimental results are presented in Table 1.

As shown in the table, compared to the MobileNetV3 backbone network, the improved MobileNetV3-ECA backbone network achieved increases of 3.3 percentage points in MIoU and 2.22 percentage points in Recall. At the same time, its parameter count is only one-tenth that of the Xception network. Although GhostNet has slightly fewer parameters than the improved backbone network, MobileNetV3-ECA demonstrates superior segmentation performance from a comprehensive perspective. Considering the complexity of jujube belt image shape characteristics, such as blurred boundaries, uneven spatial distribution, and diverse and irregular forms and sizes, MobileNetV3-ECA is selected as the optimal choice for feature extraction due to its superior performance.

Table 1. Performance comparison of backbone networks.

Method	MIoU/%	Pr/%	Re/%	Size/M
Xception	81.51	83.62	86.32	209
GhostNet	84.27	86.47	84.37	24
MobileNetV3	86.36	90.33	87.81	31.2
Our	89.66	90.22	90.03	31.2

3.4. Ablation Study

An ablation study was conducted to validate the impact of the proposed Mobilenetv3-ECA and SwinASPP modules at different positions on the model's segmentation per-

formance. The baseline model used DeepLabV3+ with the original Xception and ASPP modules. Due to the class imbalance issue in the jujube belt dataset, MIoU was used as the evaluation metric in the ablation study. As shown in Table 2, replacing the backbone network with Mobilenetv3-ECA increased the dataset's MIoU by 5.99 percentage points compared to the baseline model. After replacing the ASPP module with SwinASPP, the dataset's MIoU increased by 7.67 percentage points compared to the baseline model. When both modules were replaced simultaneously, the dataset's MIoU improved by 11.06 percentage points relative to the baseline model. In summary, replacing both modules significantly enhanced the segmentation performance on the jujube belt dataset.

Table 2. Ablation study results SwinLabNet.

Structure	MIoU/%
Xception + ASPP	84.67
Mobilenetv3-ECA + ASPP	90.66
Xception + SwinASPP	92.34
Mobilenetv3-ECA + SwinASPP	95.73

3.5. Comparative Analysis of Different Model Performances

To ensure the fairness of the experiments, all experiments in this chapter utilized the same training, validation, and test sets. The parameter settings for the comparative experiments were based on the optimal results from their respective preliminary experiments. To validate the effectiveness of the proposed method for semantic segmentation of jujube belt images, we conducted training comparisons under the same conditions using representative mainstream semantic segmentation models, including Enet [38], Bisenetv2 [39], IRASPP [40], U-Net [41], PSPNet [42], FCN [43], and DeepLabV3+ [44] (with MobileNetv2 as the backbone). Performance comparisons were made using the Mean Intersection over Union (MIoU) metric, which effectively balances precision and recall, making it suitable for assessing the overall performance of segmentation models.

As shown in the Table 3, under the same experimental conditions, the improved SwinLabNet model proposed in this paper outperforms the original and other mainstream segmentation models in performance metrics. The MIoU improved by 5.3% compared to the original DeepLabV3+ model. The experiments demonstrate the effectiveness and superiority of this method, which better accomplishes the segmentation of drivable areas in the jujube belt, providing a foundation for precise navigation in future applications (Table 3).

Table 3. Results of different semantic segmentation models.

Model	MIoU/%
Enet	84.90
Bisenetv2	84.36
IRASPP	88.84
U-Net	92.58
PSPNet	90.83
FCN	70.3
DeepLabV3+	90.57
Our	95.73

As shown in Figure 10, the segmentation results of various semantic segmentation methods selected in this study are displayed on test images, covering multiple scenarios at different times of the day. The figure shows that the ENet and BiseNetV2 methods can only detect partial jujube belt targets in the images. The IRASPP method performs well in the morning but struggles with long-distance jujube belt images and under low-light conditions in the evening. U-Net, PSPNet, and DeepLabV3+ can accurately detect the

jujube belt target regions with good performance, but they struggle with long-distance detection and are relatively large models. FCN. has the poorest detection performance, failing to adequately detect the jujube belt regions. In contrast, the proposed model achieves the best segmentation across different times of the day and long distances, outperforming DeepLabV3+ while significantly reducing the number of parameters.

Figure 10. Visualization of Results from Different Segmentation Models. (**a**) Morning, (**b**) Midday, (**c**) Afternoon.

3.6. Analysis of Visualization Results

To validate the superiority of the proposed model in long-distance jujube belt segmentation, it is compared with the previously effective models, Unet and Deeplabv3+.

This paper employed the Grad-CAM [45] heatmap visualization method to highlight the regions in the images that contribute significantly to classification. Figure 10 presents some of the results. In the heatmaps of the jujube belt dataset, red indicates the location and intensity of the target, with higher intensity signifying greater impact on the model's detection outcomes. As shown in the figure, under strong lighting and background interference conditions (Figure 11a,b), the comparison models focus on the jujube belt area, but due to the strong lighting, the texture distinction of the jujube belt becomes unclear, and shadows from trees and leaves cause misclassification. Under low-light conditions (Figure 11c), the low contrast of the jujube belt images results in blurred edge and texture features, causing the comparison models to lose some detail information during training and fail to capture the overall context effectively. In contrast, the SwinLabNet model is able to resist irrelevant background interference and focus more accurately on the navigable area. The experimental results demonstrate that the improved model focuses more effectively on the jujube belt area. The darker regions nearly cover the entire jujube belt, with higher intensity at the corresponding pixels and fewer misclassifications.

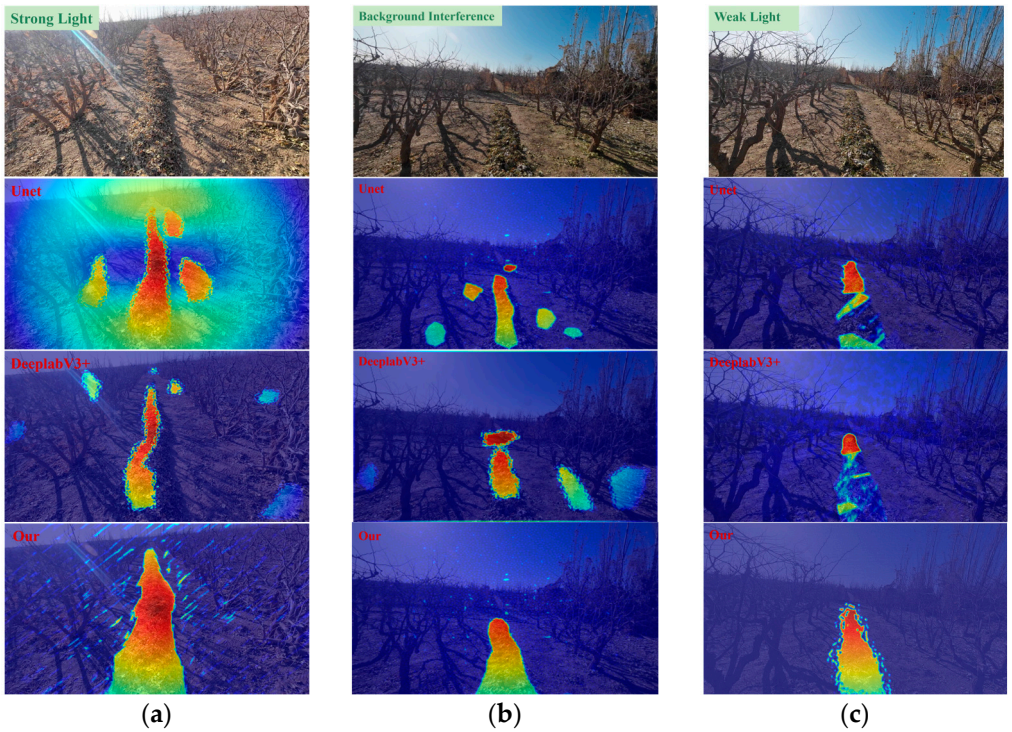


Figure 11. Grad-CAM Visualization Analysis. (**a**) Strong light; (**b**) Background Interference; (**c**) Weak Light.

3.7. Model Generalization

This section uses the vegetable dataset [24] for testing to validate the model's robustness and generalization performance. As shown in Figure 12, the improved model demonstrated strong robustness and stability across five complex field road scenarios, including a greenhouse strawberry garden, a mulched vegetable field, and environments with shadows, darkness, and intense light.

Figure 12. Segmentation Results on the Vegetable Dataset.

4. Conclusions

This paper presents SwinLabNet, an algorithm to identify drivable areas in jujube orchards in unstructured environments. It addresses the challenges of long and blurred jujube belt boundaries, complex and dispersed information, and highly irregular shapes and sizes. The models Enet, Bisenetv2, IRASPP, U-Net, PSPNet, FCN, and DeepLabV3+ were selected for comparative testing under the same training conditions, leading to the following conclusions:

(1) First, MobileNetV3-ECA was used in the feature extraction stage, significantly reducing the model's parameters. Second, the Swin Transformer was introduced to enhance the model's ability to capture contextual semantic information, addressing the issue of weak correlations between long-distance features. Finally, a mixed loss function was employed to handle the class imbalance problem, enabling the efficient extraction of abundant semantic information with a simple training method and fewer parameters.

(2) Regarding accuracy, the experimental results show that the improved model achieved an MIoU of 95.73%, a precision of 97.24%, and a recall of 98.36%. Compared to the original DeepLabV3+ network, these metrics improved by 5.22%, 3.62%, and 2.04%, respectively. When handling the jujube belt dataset, characterized by long and blurred boundaries, complex information, and discrete distribution, the proposed method demonstrated superior segmentation performance compared to other mainstream models. It also shows strong robustness and stability on vegetable datasets.

(3) Regarding lightweight design, this model uses MobileNetV3-ECA as the backbone network, with the number of parameters reduced to less than one-tenth of the original model. This provides better adaptability for deployment on edge devices.

Future research directions include optimizing the proposed MobileNetV3-ECA and SwinASPP modules and applying them to other neural networks. Further improvements will be made to the jujube belt image dataset in dwarf densely planted jujube orchards under varying light intensities and environmental conditions.

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

Funding: Supported by the Bingtuan Agriculture and Rural Affairs Bureau Project (2023AA402), the Tianshan Talents Program (2022TSYCCX0117), High-Level Talent Program of Shihezi University, Project (RCZK202441), the High-Level Talent Project (RCZK2021B15), and the Shanghai Municipal Science and Technology Innovation Action Plan' Domestic Science and Technology Cooperation Project (23015820400).

Institutional Review Board Statement: No applicable.

Data Availability Statement: The data are available within the article.

Acknowledgments: The authors would like to thank their schools and colleges, as well as the funding of the project. All supports and assistance are sincerely appreciated. Additionally, we sincerely appreciate the work of the editor and the reviewers of the present paper.

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

References

1. Dou, H.; Chen, Z.; Zhai, C.; Zou, W.; Song, J.; Feng, F.; Zhang, Y. Research Progress on Autonomous Navigation Technology for Intelligent Orchard Operation Equipment. *Trans. Chin. Soc. Agric. Mach.* **2024**, *55*, 891–898.
2. Meng, Z.; Wang, H.; Fu, W.; Liu, M.; Yin, Y.; Zhao, C. Research Status and Prospects of Agricultural Machinery Autonomous Driving. *Trans. Chin. Soc. Agric.* **2023**, *54*, 1–24.
3. Han, L.; He, X.; Wang, C.; Liu, Y.; Song, J.; Qi, P.; Liu, L.; Li, T.; Zheng, Y.; Lin, G.; et al. Key Technologies and Equipment for Smart Orchard Construction and Prospects. *Smart Agric.* **2022**, *4*, 1–11.
4. Zhou, H.; Wang, X.; Au, W.; Kang, H.; Chen, C. Intelligent robots for fruit harvesting: Recent developments and future challenges. *Precis. Agric.* **2022**, *23*, 1856–1907. [CrossRef]
5. Xie, B.; Jin, Y.; Faheem, M.; Gao, W.; Liu, J.; Jiang, H.; Cai, L.; Li, Y. Research progress of autonomous navigation technology for multi-agricultural scenes. *Comput. Electron. Agric.* **2023**, *211*, 107963. [CrossRef]
6. Gao, X.; Li, J.; Fan, L.; Zhou, Q.; Yin, K.; Wang, J.; Song, C.; Huang, L.; Wang, Z. Review of Wheeled Mobile Robots' Navigation Problems and Application Prospects in Agriculture. *IEEE Access* **2018**, *6*, 49248–49268. [CrossRef]
7. Bai, Y.; Zhang, B.; Xu, N.; Zhou, J.; Shi, J.; Diao, Z. Vision-based navigation and guidance for agricultural autonomous vehicles and robots: A review. *Comput. Electron. Agric.* **2023**, *205*, 107584. [CrossRef]
8. Jin, Y.; Liu, J.; Xu, Z.; Yan, S.; Li, P.; Wang, J. Development status and trend of agricultural robot technology. *Int. J. Agric. Biol. Eng.* **2021**, *14*, 1–19. [CrossRef]
9. Shi, J.; Bai, Y.; Diao, Z.; Zhou, J.; Yao, X.; Zhang, B. Row detection BASED navigation and guidance for agricultural robots and autonomous vehicles in row-crop fields: Methods and applications. *Agronomy* **2023**, *13*, 1780. [CrossRef]
10. Pham, T. Semantic Road Segmentation using Deep Learning. In Proceedings of the 2020 Applying New Technology in Green Buildings (ATiGB), Da Nang, Vietnam, 12–13 March 2021; pp. 45–48.
11. Wang, L.; Chen, X.; Hu, L.; Li, H. Overview of Image Semantic Segmentation Technology. In Proceedings of the 2020 IEEE 9th Joint International Information Technology and Artificial Intelligence Conference (ITAIC), Chongqing, China, 11–13 December 2020; pp. 19–26.
12. Kheradmandi, N.; Mehranfar, V. A critical review and comparative study on image segmentation-based techniques for pavement crack detection. *Constr. Build. Mater.* **2022**, *321*, 126162. [CrossRef]
13. Jing, J.; Liu, S.; Wang, G.; Zhang, W.; Sun, C. Recent advances on image edge detection: A comprehensive review. *Neurocomputing* **2022**, *503*, 259–271. [CrossRef]
14. Barhate, D.; Nemade, V. Comprehensive Study on Automated Image Detection by Robotics for Agriculture Applications. In Proceedings of the 2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 12–14 June 2019; pp. 637–641.
15. Kamilaris, A.; Prenafeta-Boldú, F.X. A review of the use of convolutional neural networks in agriculture. *J. Agric. Sci.* **2018**, *156*, 312–322. [CrossRef]
16. Saleem, M.H.; Potgieter, J.; Arif, K.M. Automation in agriculture by machine and deep learning techniques: A review of recent developments. *Precis. Agric.* **2021**, *22*, 2053–2091. [CrossRef]

17. Pally, R.J.; Samadi, S. Application of image processing and convolutional neural networks for flood image classification and semantic segmentation. *Environ. Model. Softw.* **2022**, *148*, 105285. [CrossRef]
18. Moazzam, S.I.; Khan, U.S.; Tiwana, M.I.; Iqbal, J.; Qureshi, W.S.; Shah, S.I. A Review of Application of Deep Learning for Weeds and Crops Classification in Agriculture. In Proceedings of the 2019 International Conference on Robotics and Automation in Industry (ICRAI), Rawalpindi, Pakistan, 21–22 October 2019; pp. 1–6.
19. Wang, J.-J.; Liu, Y.-F.; Nie, X.; Mo, Y.L. Deep convolutional neural networks for semantic segmentation of cracks. *Struct. Control Health Monit.* **2022**, *29*, e2850. [CrossRef]
20. Turay, T.; Vladimirova, T. Toward performing image classification and object detection with convolutional neural networks in autonomous driving systems: A survey. *IEEE Access* **2022**, *10*, 14076–14119. [CrossRef]
21. Kar, M.K.; Nath, M.K.; Neog, D.R. A review on progress in semantic image segmentation and its application to medical images. *S.N. Comput. Sci.* **2021**, *2*, 397. [CrossRef]
22. 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]
23. Thisanke, H.; Deshan, C.; Chamith, K.; Senviratne, S.; Vidanaarachchi, R. Semantic segmentation using Vision Transformers: A survey. *Eng. Appl. Artif. Intell.* **2023**, *126*, 106669. [CrossRef]
24. Yu, J.; Zhang, J.; Shu, A.; Chen, Y.; Chen, J.; Yang, Y.; Tang, W.; Zhang, Y. Study of convolutional neural network-based semantic segmentation methods on edge intelligence devices for field agricultural robot navigation line extraction. *Comput. Electron. Agric.* **2023**, *209*, 107811. [CrossRef]
25. Zhang, L.; Li, M.; Zhu, X.; Chen, Y.; Huang, J.; Wang, Z.; Hu, T.; Wang, Z.; Fang, K. Navigation path recognition between rows of fruit trees based on semantic segmentation. *Comput. Electron. Agric.* **2024**, *216*, 108511. [CrossRef]
26. Zheng, Z.; Hu, Y.; Li, X.; Huang, Y. Autonomous navigation method of jujube catch-and-shake harvesting robot based on convolutional neural networks. *Comput. Electron. Agric.* **2023**, *215*, 108469. [CrossRef]
27. Yang, Z.; Ouyang, L.; Zhang, Z.; Duan, J.; Yu, J.; Wang, H. Visual navigation path extraction of orchard hard pavement based on scanning method and neural network. *Comput. Electron. Agric.* **2022**, *197*, 106964. [CrossRef]
28. Li, X.; Su, J.; Yue, Z.; Duan, F. Adaptive multi-ROI agricultural robot navigation line extraction based on image semantic segmentation. *Sensors* **2022**, *22*, 7707. [CrossRef]
29. Cao, M.; Tang, F.; Ji, P.; Ma, F. Improved real-time semantic segmentation network model for crop vision navigation line detection. *Front. Plant Sci.* **2022**, *13*, 898131. [CrossRef]
30. Zhang, X.; Yang, Y.; Li, Z.; Ning, X.; Qin, Y.; Cai, W. An improved encoder-decoder network based on strip pool method applied to segmentation of farmland vacancy field. *Entropy* **2021**, *23*, 435. [CrossRef]
31. Baheti, B.; Innani, S.; Gajre, S.; Talbar, S. Semantic scene segmentation in unstructured environment with modified DeepLabV3+. *Pattern Recognit. Lett.* **2020**, *138*, 223–229. [CrossRef]
32. Bai, H.; Cheng, J.; Huang, X.; Liu, S.; Deng, C. HCANet: A Hierarchical Context Aggregation Network for Semantic Segmentation of High-Resolution Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 6002105. [CrossRef]
33. Dais, D.; Bal, I.E.; Smyrou, E.; Sarhosis, V. Automatic crack classification and segmentation on masonry surfaces using convolutional neural networks and transfer learning. *Autom. Constr.* **2021**, *125*, 103606. [CrossRef]
34. Deng, J.; Lu, Y.; Lee, V.C.-S. Concrete crack detection with handwriting script interferences using faster region-based convolutional neural network. *Comput. Aided Civ. Infrastruct. Eng.* **2020**, *35*, 373–388. [CrossRef]
35. Zhou, Z.; Zhang, J.; Gong, C. Hybrid semantic segmentation for tunnel lining cracks based on Swin Transformer and convolutional neural network. *Comput.-Aided Civ. Infrastruct. Eng.* **2023**, *38*, 2491–2510. [CrossRef]
36. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 318–327. [CrossRef] [PubMed]
37. Qu, Z.; Mei, J.; Liu, L.; Zhou, D.-Y. Crack Detection of Concrete Pavement With Cross-Entropy Loss Function and Improved VGG16 Network Model. *IEEE Access* **2020**, *8*, 54564–54573. [CrossRef]
38. Pratik, V.; Vedhapriyavadhana, R.; Chidambaranathan, S. Polyp Segmentation Using UNet and Enet. In Proceedings of the 2023 6th International Conference on Recent Trends in Advance Computing (ICRTAC), Chennai, India, 14–15 December 2023; pp. 516–522.
39. Hu, X.; Ren, H. A Road Scene Semantic Segmentation Algorithm Based on Improved BiSeNetV2. In Proceedings of the 5th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT 2024), Nanjing, China, 29–31 March 2024; pp. 649–652.
40. Sola, D.; Scott, K.A. Efficient Shallow Network for River Ice Segmentation. *Remote Sens.* **2022**, *14*, 2378. [CrossRef]
41. Lavrynenko, R.; Ryabova, N. Transforming Semantic Segmentation into Instance Segmentation with a Guided U-Net. In Proceedings of the 2023 IEEE 18th International Conference on Computer Science and Information Technologies (CSIT), Lviv, Ukraine, 19–21 October 2023; pp. 1–4.
42. Zhang, C.; Zhao, J.; Feng, Y. Research on Semantic Segmentation Based on Improved PSPNet. In Proceedings of the 2023 International Conference on Intelligent Perception and Computer Vision (CIPCV), Xi'an, China, 19–21 May 2023; pp. 1–6.
43. Farhangfar, S.; Rezaeian, M. Semantic Segmentation of Aerial Images using FCN-based Network. In Proceedings of the 2019 27th Iranian Conference on Electrical Engineering (ICEE), Yazd, Iran, 30 April–2 May 2019; pp. 1864–1868.

44. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 834–848. [CrossRef]
45. Inbaraj, X.; Villavicencio, C.; Macrohon, J.; Jeng, J.-H.; Hsieh, J.-G. Object identification and localization using Grad-CAM++ with mask regional convolution neural network. *Electronics* **2021**, *10*, 1541. [CrossRef]