Article

A Deep Residual U-Type Network for Semantic Segmentation of Orchard Environments

Gaogao Shang, Gang Liu *, Peng Zhu, Jiangyi Han, Changgao Xia and Kun Jiang

College of Automotive and Traffic Engineering, Jiangsu University, Zhenjiang 212013, China; shanggaogao@ujs.edu.cn (G.S.); 2221804068@stmail.ujs.edu.cn (P.Z.); hjy0306@ujs.edu.cn (J.H.); xiacg@ujs.edu.cn (C.X.); 2211904011@stmail.ujs.edu.cn (K.J.)

* Correspondence: 2211804018@stmail.ujs.edu.cn; Tel.: +86-188-5285-6872

Abstract: Recognition of the orchard environment is a prerequisite for realizing the autonomous operation of intelligent horticultural tractors. Due to the complexity of the environment and the algorithm's dependence on ambient light, the traditional recognition algorithm based on machine vision is limited and has low accuracy. However, the deep residual U-type network is more effective in this situation. In an orchard, the deep residual U-type network can perform semantic segmentation on trees, drivable roads, debris, etc. The basic structure of the network adopts a U-type network, and residual learning is added in the coding layer and bottleneck layer. Firstly, the residual module is used to improve the network depth, enhance the fusion of semantic information at different levels, and improve the feature expression capability and recognition accuracy. Secondly, the decoding layer uses up-sampling for feature mapping, which is convenient and fast. Thirdly, the semantic information of the coding layer is integrated by skip connection, which reduces the network parameters and accelerates the training. Finally, a network was built through the Pytorch Deep Learning Framework, which was implemented to train the data set and compare the network with the fully convolutional neural network, the U-type network, and the Front-end+Large network. The results show that the deep residual U-type network has the highest recognition accuracy, with an average of 85.95%, making it more suitable for environment recognition in orchards.

Keywords: machine vision; deep residual U-type network; semantic segmentation; information fusion

Citation: Shang, G.; Liu, G.; Zhu, P.; Han, J.; Xia, C.; Jiang, K. A Deep Residual U-Type Network for Semantic Segmentation of Orchard Environments. *Appl. Sci.* **2021**, *11*, 322. <https://doi.org/10.3390/app11010322>

Received: 5 December 2020

Accepted: 26 December 2020

Published: 31 December 2020

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The “Made in China 2025 strategy” puts forward new requirements for China’s agricultural equipment and requires continuous improvement of the ability of agricultural machinery intelligence and precision operation [1]. The horticultural tractor is an important piece of agricultural machinery working in orchards, forest gardens, and other environments. One of the basic tasks of realizing intelligent operation of the horticultural tractor is to identify the working environment. With the development of science and technology, more and more methods are being used for environmental recognition, such as lidar for scanning and recognizing the surrounding environment. However, due to the high cost of lidar, it is difficult to apply to agricultural products. In contrast, the use of ordinary cameras as sensors has many advantages, like comprehensive information collection and low price [2]. For vision-based environmental recognition, related algorithms based on the target of achieving rapid and accurate recognition have mainly been formulated.

In the research of environment recognition, Radcliffe [3] developed a small autonomous navigation vehicle in a peach orchard based on machine vision. The machine vision system was based on the use of a multispectral camera to capture real-time images and process the images to obtain trajectories for autonomous navigation. Lyu [4] used naïve Bayesian classification to detect the boundary between the trunks and the ground of an orchard and proposed an algorithm to determine the centerline of the orchard road

for automatic driving of the orchard's autonomous navigation vehicle. In order to enable agricultural robots to extract effective navigation information in a complex, open, non-structural farmland environment, and solve the instability of navigation information extraction algorithm caused by light changes, An [5] proposed to use the color constancy theory to solve lighting problems in machine vision navigation. Zhao [6] proposed a vision-based agricultural vehicle guidance system, using the Hough transform method to extract guidance parameters, and designed a path recognition method based on centerline detection and erosion algorithm.

The environmental recognition method in the literature basically solves the problem of recognition in a specific environment and has specificity, which has certain limitations. In recent years, there has been rapid development of deep learning, which has shown excellent performance in environment recognition [7] and target detection [8]. Environment recognition algorithms based on deep learning have the advantages of strong robustness and high accuracy. In the research of environment recognition based on deep learning, Oliveira [9] used the high-order features in convolutional neural network learning scenarios to segment road scenes and generated training labels by training on image data sets, then used the new textures based on color layer fusion to obtain the maximum consistency of the road area, and finally combined the offline and online information to detect urban road areas. Badrinarayanan [10] proposed the Segnet network for road scene understanding. The network architecture is based on an encoding–decoding network, but the segmentation of road boundary details still needs to be improved. He [11] used the Spatial Pyramid Pooling Network (SPP-Net) to extract roads from remote sensing images and used structural similarity (SSIM) as the loss function for road extraction, which can extract the fuzzy prediction in the extraction results and improve the image quality of the extracted roads. Li [12] constructed a semantic segmentation model of field road scenes based on the hollow convolutional neural network for complex field road scenes in hilly and mountainous areas. The model includes a front-end module and a context module. The front-end module is an improved structure of VGG-16 (Visual Geometry Group Network) fusion cavity convolution, and the context module is a cascade of the convolutional layer with different expansion coefficients. The two-stage training method was adopted to train the model, and the model showed good adaptability to shadow interference. Wang used the YOLOV3 (You Only Look Once) convolutional neural network to extract the feature points on road images from orchards, generated the navigation line by the least square method [13], and conducted experiments in a variety of different natural environments, with a deviation of 3.5 cm. Zhang [14] combined the strengths of U-Net and residual learning to construct a seven-layer deep residual U-type network for road extraction from aerial images and achieved good results. Anita [15] used a deep residual U-Net convolutional nerve network for lung CT image segmentation in the field of biomedicine, and the segmentation effect was significantly improved, which effectively helped in the early diagnosis and treatment of lung diseases.

Environment recognition algorithms based on deep learning have greatly improved the accuracy of recognition [16], but they are mainly used in the recognition of structured roads or environments with few distractors, and the application scenarios are relatively singular. The orchard environment is complex and changeable, and the span between different objects is larger. Recognizing these objects requires a deeper network, but this will cause problems such as gradient disappearance, increased parameters, and difficulty in training. Therefore, this paper takes a U-type network as the main body of the network model; gives full play to its advantages, such as fewer parameters and high recognition accuracy; and adds residual blocks in the feature extraction process to deepen the network and improve the extraction ability of object edge information, finally building an orchard environment recognition algorithm based on a deep residual U-shaped network. Pixel-level semantic segmentation is used for the image of the orchard environment to obtain various information on the orchard environment, laying the foundation for autonomous operation of the gardening tractor in the future.

The main contributions of this paper are as follows: Section 1 introduces the current research status of environment identification and proposes an orchard environment identification algorithm based on the deep residual U-type network. Section 2 describes the acquisition and processing of orchard environment data sets. Section 3 presents the construction of the orchard environment model and constructs the deep residual U-type network segmentation mode by analyzing the characteristics of the residual network and the U-type network and combining the actual needs of the orchard environment identification. In the fourth section, experimental comparative analysis is carried out on the fully convolutional neural network, the U-type network, the Front-end+Large network, and the deep residual U-type network. Finally, the conclusions and future work are presented in Section 5.

2. Establishment of the Orchard Environmental Data Set

2.1. Data Set Acquisition

According to the different working conditions of the horticultural tractor in the orchard, the orchard environment can be divided into the transportation environment and the working environment, as shown in Figure 1. The working environment is characterized by horticultural tractor work between two rows of fruit trees, with minor environmental changes, mainly carrying out pesticide spraying, picking, and other work. The transportation environment is characterized by horticultural tractors driving on non-structural roads with unobvious road boundaries and many sundries.

Figure 1. Orchard environment division: (a) The transportation environment; (b) The working environment.

To realize pixel-level segmentation of the orchard environment, the first step is to obtain a real image of the orchard environment. In this paper, a GoPro4 HD camera was used as the orchard environment acquisition tool, with pixel resolution of 4000×3000 . In order to improve the robustness and universality of the network model, we collected data under various conditions according to changes in weather and light. The collected data were saved in the form of a video. We used Python scripts to capture pictures from the video at a speed of 30 fps. In order to reduce the consumption of computer graphics memory, the picture pixels were adjusted to 1024×512 , and a total of 2337 images of the orchard environment were obtained after processing; the images were arranged according to the corresponding serial numbers.

2.2. Data Set Preprocessing

Semantic segmentation requires preprocessing of the obtained image data. Orchard environment recognition based on the deep residual U-type network is a type of supervised learning. Original pictures and labeled pictures need to be input during model training, so the collected images need to be labeled manually. According to the characteristics of the orchard environment and the autonomous operation requirements of horticultural tractors, the objects in the orchard environment were divided into four categories: background, road, peach trees, and debris. In this paper, the orchard environment data set was artificially labeled using the semantic split labeling tool Labelme [17], which labels different categories in different colors. Table 1 provides information on the orchard environment categories and Figure 2 shows a labeled map of the orchard environment.

Table 1. Orchard environment category label information.

Category	Number	R	G	B
Background	0	0	0	0
Road	1	0	0	255
Peach	2	0	255	0
Sundries	3	210	210	210

Figure 2. Labeled map of the orchard environment.

In order to solve the problem of insufficient training data scale, we used a data enhancement method [18] to expand the orchard environment data set, which is convenient, fast, and effective in solving the problem of inconsistent enhancement of original and labeled pictures. We expanded the data set to be 3 times as large by flipping, cropping, and scaling the images. After the above processing, the data set was divided into training and test sets in a ratio of 7:3 by the random selection method. There were 6214 training pictures and 2663 test pictures, and the serial number of each original picture corresponded to the serial number of the labeled picture.

3. Construction of the Orchard Environmental Identification Model

In recent years, with the application of fully convolutional networks (FCNs), convolutional neural networks (CNNs) have been used to generate semantic segmentation charts of any size on the basis of feature diagrams, which can split images at the pixel level [19]. Based on this method, many algorithms have been derived, such as the deep-splitting network framework DeepLab series [20,21], the scene resolution network PSPNet using a pyramid pooling module [22], and the U-Net network for medical image segmentation [23].

The U-Net network can still achieve better model segmentation accuracy with few samples, and for the first time, used the skip connection to add encoded features to decoding features, creating an information propagation path that allows signals to spread more easily between low-level and advanced features; this not only facilitates backpropagation during training but also improves model segmentation accuracy. However, the U-Net network has insufficient ability to obtain contextual information from images, especially for complex scene data with large differences in category scales. Multi-scale fusion is usually used to increase the depth of the network to improve the U-Net network's ability to obtain contextual information from images. However, with increasing network depth, the accuracy of model recognition decreases rapidly after saturation rapidly declines, and the recognition error increases. To solve this problem, He et al. [24] proposed a deep residual network that used identity mapping to obtain more contextual information. The error did not increase with the depth of the network, which solved the problem of training degradation.

3.1. Residual Network

In traditional convolutional neural networks, multi-layer features become more abundant with the superimposition of network layers, but simple superposition networks will cause the problem of gradient disappearance and hinder the convergence of the model. In order to improve the accuracy of environment recognition and prevent gradient disappearance, a residual network was considered for addition into the network structure.

The residual network is mainly designed as a residual block with a shortcut connection, which is equivalent to adding a direct connection channel in the network, so that the network has a stronger identity mapping ability, thus expanding the network depth and improving the network performance without overfitting. The residual network consists of a series of stacked residual units. Each residual unit can be expressed in a general form:

$$y_i = F(x_i, w_i) + h(x_i), \quad (1)$$

$$x_{i+1} = f(y_i), \quad (2)$$

where x_i is the input of the residual unit of layer i ; w_i is the network parameters of the residual unit of layer i ; $F(\cdot)$ denotes the residual function; $h(\cdot)$ denotes the identity mapping function; and $f(\cdot)$ denotes the activation function.

The residual neural network unit consists of two parts: the identity mapping part and the residual part. Identity mapping mainly integrates the input with the output processed by the residual, which facilitates the fusion of subsequent feature information, and the residual part is generally composed of multiple converse neural networks, normalized layers, and activation functions. Through the superposition of identity mapping and residuals to realize information interaction, the problem of poor ability to extract underlying features in the residual part is compensated. Figure 3 shows the difference between the common neural network unit and the residual neural network unit. Figure 3a is the structure of the common neural network unit, and Figure 3b is the structure of the residual neural network unit.

Figure 3. The difference between the common neural network unit and the residual neural network unit: (a) Common neural network unit; (b) Residual neural network unit.

3.2. Construction of the Deep Residual U-Net Model

Based on the characteristics of the deep residual network and U-Net network, we propose the deep residual U-type network, which introduces a residual layer to deepen the U-Net structure and avoids the occurrence of excessive training time, too many training parameters, and overfitting. In semantic segmentation, it is necessary to use both low-level detailed information and high-level semantic information for better results. The deep residual U-type network can well retain the information of both. The deep residual U-type network has two specific benefits: (1) For complex environment recognition, adding

residual units will help the network training, improving recognition accuracy. (2) The long connection of low-level information and high-level information of the network and the skip connection of the residual unit are conducive to the dissemination of information, the parameter update distribution is more uniform, and the network model can have better performance.

In this paper, the nine-level architecture of the deep residual U-type network was applied to the identification of targets in the orchard environment, and the network consisted of three parts: a coding layer, bottleneck layer, and decoding layer. The first part of the coding layer extracts the features in the image, forming a feature map. The second part of the bottleneck layer connects the coding layer and decoding layer, equivalent to the bridge, to obtain low-frequency information in the image. The third part of the decoding layer restores the feature map to pixel-level classification, that is, semantic segmentation. Residual units were added to the coding and bottleneck layers to obtain contextual information, and the convolutional modules in the network contained a convolution layer, a batch normalization (BN) layer, and an activation function (Rectified Linear Unit, ReLU). Adopting the batch normalization layer can prevent instability of network performance caused by too many data before the activation function, which effectively solves the problem of gradient disappearance or gradient explosion [25]. Using the ReLU activation function can effectively reduce the amount of calculation and increase the nonlinear relationship between the various layers of the neural network [26]. The identity mapping connects the input and output of the residual neural network unit. Since the dimensionality of the input image changes during convolution, the corresponding dimension of the input image also needs to be changed during identity mapping. In this paper, a convolution kernel with a size of 1×1 , a step size of 1, and a batch normalization layer was used as the identity mapping function.

There are four residual units in the coding layer, each of which is activated by the ReLU function after the residual function and constant mapping function are added together, and then the feature graph size is halved by Maxpool, which can effectively reduce parameters, reduce overfitting, improve model performance, and save computational memory [27]. The decoding layer consists of four basic units, using bilinear upsampling and the convolutional module for decoding. Compared with deconvolution, the above method is easier to implement in engineering and does not involve too many hyperparameter settings [28]. At the same time, the feature information in the coding layer is fused with the feature information in the decoding layer by a jumping connection, which makes full use of the semantic information and improves the recognition accuracy. After the last layer of decoding, a 1×1 convolution and the Softmax activation function are used to achieve multi-classification identification of the orchard environment. The network model in this paper has 25 convolutional layers and 4 maximum pooled layers, and the structure of the network model is shown in Figure 4.

3.3. Loss Function

When training the segmentation network, the images that need to be trained are split through the segmentation network to get segmentation images $S(X_n)$. $S(\cdot)$ is the segmentation network model, and X_n is the input image. The segmented image $S(X_n)$ is compared with the corresponding label image Y_n , and the loss function is minimized to make the segmented image close to the original labeled image, which ensures that the segmentation network can produce accurate predictions and has good robustness. In this paper, the standard cross-entropy loss function was used as a loss function to detect the difference between the segmentation image $S(X_n)$ and the labeled image Y_n [29]. The expression of the cross-entropy loss function is

$$L_{ce} = -\sum_{h,w} \sum_c Y_n^{(h,w,c)} \log(S(X_n)^{(h,w,c)}), \quad (3)$$

where L_{ce} is the cross-entropy loss function; Y_n is the ground truth (GT); X_n is the input image; $S(\cdot)$ is the segmentation image; and h, w, c denote the height, width, and number of channels of the image, respectively.

Figure 4. The structural of the deep residual U-type network model.

4. Orchard Environment Identification Test

The fully convolutional neural network, the U-type network, the Front-end+Large network, and the deep residual U-type network can all achieve image pixel-level semantic segmentation, but the four networks have their own unique characteristics. Among them, the fully convolutional neural network uses deconvolution to achieve semantic segmentation without a bottleneck layer, and it uses the force pooling method to enter the decoding layer from the encoding layer directly. The U-type network adds a bottleneck layer between the encoding and decoding layers to realize a smooth transition, and it adopts upsampling and skip connection for decoding to achieve semantic segmentation. The Front-end+Large network has high recognition accuracy and high adaptability to different images, so it is widely used in farmland road recognition. The deep residual U-type network adopts a U-type network structure. By adding residual blocks in the coding layer and the bottleneck layer, the image context information and multi-layer network are fully utilized to realize detailed processing of the image. We conducted an experimental comparative analysis of the above four networks, as detailed in the following.

4.1. Test Implementation Details

Based on the deep residual U-type network model proposed above, the deep learning framework Pytorch was used to build the orchard environment recognition and segmentation model. There were a total of 6214 training images with a size of 1024×512 . The hardware environment of the experiment was an Intel Core I7 9700K 8-core processor with a GeForce RTX 2070 and an 8 GB memory capacity.

With the deepening of training, the model is easily caught in the local minimum problem. In order to solve this problem, we adopted the RMSProp algorithm [30], which, according to the principle of minimization of the loss function, constantly dynamically adjusts the network model parameters to make the objective function converge faster. The initial learning rate in the RMSProp algorithm was set to 0.4, and the weight attenuation coefficient was 10^{-8} . During model training, the batch size of data loading was 8, the number of iterations was 300, and the loss function value was recorded for each iteration.

4.2. Evaluation Indicators

There are three evaluation criteria for semantic segmentation, namely, execution time, memory footprint, and accuracy. The accuracy is based on the manually annotated image as the basic standard, compared with the prediction image from the segmentation network, and judged by calculating the pixel error between the prediction image and the real annotated image.

Suppose there are $k + 1$ categories (k target categories and one background category), P_{ii} represents a correct prediction, and both P_{ij} and P_{ji} represent incorrect predictions. The general evaluation criteria are as follows:

- (1) Pixel Accuracy (PA): The ratio of the number of correctly classified pixels to the total number of pixels, and the formula is

$$PA = \frac{\sum_{i=0}^k p_{ii}}{\sum_{i=0}^k \sum_{j=0}^k p_{ij}}. \quad (4)$$

- (2) Mean Pixel Accuracy (MPA): The average value of the ratio of the number of correct pixels in each category to the total number of pixels in that category,

$$MPA = \frac{1}{k+1} \sum_{i=0}^k \frac{p_{ii}}{\sum_{j=0}^k p_{ij}}. \quad (5)$$

- (3) Mean Intersection over Union (MIoU):

$$MIoU = \frac{1}{k+1} \sum_{i=0}^k \frac{p_{ii}}{\sum_{j=0}^k p_{ij} + \sum_{j=0}^k p_{ji} - p_{ii}}. \quad (6)$$

Among the above criteria, the *MIoU* is the most representative and easy to implement. Many competitions and researchers use these criteria to evaluate their results [31]. In this paper, *PA* and *MIoU* were used as evaluation indicators for different categories of segmentation and overall network models.

4.3. Test Results and Analysis

When the deep residual U-type network model was trained, the model was saved every five iterations, and the model with the highest average intersection ratio among all models was selected as the test model. In order to verify the superiority of the proposed

network model, it was compared with the fully convolutional neural network model, the U-type network model, and the Front-end+Large network. Figures 5–8 show the loss values and *MIoU* of each iteration of the four network models. Table 2 reports the highest Pixel Accuracy (*PA*) and the highest *MIoU* for category segmentation.

Figure 5. The iteration loss and mean intersection over union (*MIoU*) of the fully convolution neural network.

Figure 6. The iteration loss and *MIoU* of the U-type network.

Figure 7. The iteration loss and *MIoU* of the Front-end+Large network.

From its training effects, the fully convolutional neural network has a larger floating amplitude and higher frequency. After about 100 iterations, the loss value and the *MIoU* tended to stabilize, but the floating range was large. The fully convolutional neural network discards the fully connected layer and uses deconvolution to realize semantic segmentation. When recognizing the complex environment of the orchard, there are more network parameters in the fully convolutional neural network, which leads to problems such as the segmentation of the details in the image not being clear enough, the segmentation boundary between different categories being blurred, and the training time being long. The training effect of the U-type network in Figure 6 was significantly better than that

of the fully convolutional neural network. The network tended to be stable after about 60 iterations. Up-sampling was adopted to map the feature images, which greatly reduced the network parameters. However, there were still fluctuations in the preliminary training process, which was due to insufficient network depth and poor ability to distinguish the boundaries of different categories in the image during training. The effect of Front-end-Large network training in Figure 7 was significantly improved compared with that for the first two networks. The fluctuation in the previous training was small and tended to be stable around 54 iterations. The training effect of the deep residual U-type network in Figure 8 was significantly better than that of the first three networks. There was no obvious fluctuation in the early training period, the change in the loss value and *MIoU* was small after about 40 iterations, and the overall training time was short and stable. The deep residual U-type network adopts the U-type network structure but adds a residual block in the coding layer and fuses the image feature information, which better processes the image boundary information. In addition, a jump connection and constraints were added to the decoding layer, which effectively reduced network parameters. Among the four networks, the deep residual U-type network had the lowest loss value, the highest *MIoU*, and the best training effect.

Figure 8. The iteration loss and *MIoU* of the deep residual U-type network.

Table 2. Semantic segmentation results of different network models.

Category	Pixel Accuracy (%)			
	Fully Convolutional Neural Network	U-Type Network	Front-End+Large Network	Deep Residual U-Type Network
Background	77.9	86.2	85.4	91.4
Road	75.8	80.1	83.6	88.5
Peach	66.7	73.1	79.9	84.3
Sundries	60.1	70.4	69.8	79.6
<i>MIoU</i> /%	70.13	77.45	78.6	85.95

As can be seen from Table 2, in the category segmentation of the orchard environment, the four semantic segmentation network models had high accuracy in pixel recognition for backgrounds and roads, but low accuracy in recognition of fruit trees and debris. The average intersection ratio of the deep residual U-type network proposed in this paper was 85.95%, which is higher than those of the first three network models, and achieved better results in orchard environment recognition.

Figure 9 shows the respective semantic segmentation prediction images from the four network models. Among them, the segmentation image generated by the fully convolutional neural network model has the following shortcomings: One is that some areas' categories were lost in the segmentation, and small objects such as small branches could not be identified. The other is that for the segmentation of a large region, the boundary detail information processing capability is insufficient; as shown in the figure,

the segmentation of the road boundary is not clear enough. This result is due to the fact that the fully convolutional neural network model uses deconvolution in the decoding process. Although this method is simple and feasible, it causes problems such as violent pooling, blurring of segmented images, and lack of spatial consistency. Compared with the fully convolutional neural network model, the segmentation image generated by the U-type network model has higher *MIoU* and better segmentation effect. However, for the overlapping parts of the categories, the difference in the features of the separated categories is not obvious, the boundary of the overlapping part is rough, and the overlapping parts are easily lost, such as details of the intersection of fruit trees and sundries. For the complex environment of the orchard, due to insufficient depth of the U-type network, the detailed information of each category cannot be fully utilized, and overfitting occurs in the training process. The segmentation image generated by the Front-end+Large network has improved overall effect compared to those by the fully convolutional neural network and the U-type network, but some details, such as branches and debris, were still lost. It can be seen from the segmentation result image that the orchard environment recognition model based on the deep residual U-type network can well reflect the boundary information in the large area and small region category segmentation. The recognition accuracy was higher than that of the previous three segmentation network models.

Figure 9. Semantic segmentation prediction images from different network models.

In summary, the U-type network model based on deep residuals proposed in this paper can effectively improve the recognition accuracy of orchard environments, and the segmentation model can also show better robustness for complex orchard environments and light changes.

5. Conclusions and Future Work

- (1) Orchard environment recognition based on the deep residual U-type network was realized by collecting orchard environment information and constructing a deep residual U-type network. Compared with the fully convolutional neural network, the U-type network, and the Front-end+Large network, the deep residual U-type network showed strengthened ability to extract contextual information and process details of an orchard environment image.
- (2) The fully convolutional neural network, the U-type network, the Front-end+Large network, and the deep residual U-type network were tested and compared. The test results showed that the segmentation accuracy of the deep residual U-type network was better than that of the other three networks.
- (3) The semantic segmentation model based on the deep residual U-type network presented high recognition accuracy and strong robustness in actual orchard environment recognition, showing potential to provide environmental perception for the autonomous operation of horticultural tractors in orchards. At the same time, this method also has shortcomings, such as a large number of annotations required to make a data set, which is time-consuming and labor-intensive; a large amount of graphics card memory consumed during the training process; and insufficient post-processing optimization of the prediction segmentation image. The next research work will focus on these shortcomings to further improve the accuracy of model recognition.

Author Contributions: G.L., P.Z. and K.J. carried out experiments; G.L. wrote the manuscript with assistance of G.S., J.H. and C.X.; All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by Jiangsu Province Key R&D Plan (Modern Agriculture)—R&D of Horticultural Electric Tractor (BE2017333) and the Key R&D Program of Jiangsu Province (BE2018343).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: The authors appreciate Lin Zhong, Sichuan Chuanlong Tractors Co. Ltd. (Chengdu, Sichuan, China), for kindly providing valuable suggestions and help.

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

References

1. Liu, M. A Brief Discussion on the Intelligence of Agricultural Machinery and Analysis of the Current Situation of Chinese Agriculture. *Theor. Res. Urban Constr.* **2017**, *8*, 224.
2. Fan, Z.; Zhang, L.; Luo, F.; Wang, X. Applying Wide-Angle Camera in Robot Vision Localisation. *Comput. Appl. Softw.* **2014**, *31*, 191–194.
3. Radcliffe, J.; Cox, J.; Bulanon, D.M. Machine vision for orchard navigation. *Comput. Ind.* **2018**, *98*, 165–171. [[CrossRef](#)]
4. Lyu, H.-K.; Park, C.-H.; Han, D.-H.; Kwak, S.W.; Choi, B. Orchard Free Space and Center Line Estimation Using Naive Bayesian Classifier for Unmanned Ground Self-Driving Vehicle. *Symmetry* **2018**, *10*, 355. [[CrossRef](#)]
5. An, Q. *Research on Illumination Issue and Vision Navigation System of Agriculture Robot*; Nanjing Agriculture University: Nanjing, China, 2008.
6. Zhao, B.; Zhang, X.; Zhu, Z. A Vision-Based Guidance System for an Agricultural Vehicle. In Proceedings of the World Automation Congress, Kobe, Japan, 19–23 September 2010; pp. 117–121.

7. Liu, B.; Wu, H.; Wang, Y.; Liu, W. Main Road Extraction from ZY-3 Grayscale Imagery Based on Directional Mathematical Morphology and VGI Prior Knowledge in Urban Areas. *PLoS ONE* **2015**, *10*, e0138071. [[CrossRef](#)]
8. Sujatha, C.; Selvathi, D. Connected component-based technique for automatic extraction of road centerline in high resolution satellite images. *EURASIP J. Image Video Process.* **2015**, *2015*, 4144. [[CrossRef](#)]
9. Oliveira, G.L.; Burgard, W.; Brox, T. Efficient Deep Models for Monocular Road Segmentation. In Proceedings of the IEEE/RISJ International Conference on Intelligent Robots and Systems, Daejeon, Korea, 9–14 October 2016; pp. 4885–4891.
10. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [[CrossRef](#)]
11. He, H.; Yang, D.; Shicheng, W.; Wang, S.; Li, Y. Road Extraction by Using Atrous Spatial Pyramid Pooling Integrated Encoder-Decoder Network and Structural Similarity Loss. *Remote. Sens.* **2019**, *11*, 1015. [[CrossRef](#)]
12. Li, Y.; Xu, J.; Liu, D.; Yu, Y. Field Road Scene Recognition in Hilly Regions Based on Improved Dilated Convolutional Networks. *Trans. Chin. Soc. Agric. Eng.* **2019**, *35*, 150–159.
13. Wang, Y.; Liu, B.; Xiong, L.; Wang, Z.; Yang, C. Research on Generating Algorithm of Orchard Road Navigation Line Based on Deep Learning. *J. Hunan Agric. Univ.* **2019**, *45*, 674–678.
14. Zhang, Z.; Liu, Q.; Wang, Y. Road extraction by deep residual u-net. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 749–753. [[CrossRef](#)]
15. Anita, K.; Narendra, D.; Gupta, S.; Ashish, S. A deep Residual U-Net convolutional neural network for automated lung segmentation in computed tomography images. *Biocybern. Biomed. Eng.* **2020**, *40*, 1314–1327.
16. Chen, L.; Yan, N.; Yang, H.; Zhu, L.; Zheng, Z.; Yang, X.; Zhang, X. A Data Augmentation Method for Deep Learning Based on Multi-Degree of Freedom (DOF) Automatic Image Acquisition. *Appl. Sci.* **2020**, *10*, 7755. [[CrossRef](#)]
17. Zhou, J.; Li, B.; Chen, S. A Real Time Semantic Segmentation Method Based on Multi-Level Feature Fusion. *Bull. Surv. Mapp.* **2020**, *1*, 10–15.
18. Russell, B.C.; Torralba, A.; Murphy, K.P.; Freeman, W.T. LabelMe: A Database and Web-Based Tool for Image Annotation. *Int. J. Comput. Vis.* **2008**, *77*, 157–173. [[CrossRef](#)]
19. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *39*, 640–651.
20. Wu, X.; Du, M.; Chen, W. Exploiting Deep Convolutional Network and Patch-Level CRFs for Indoor Semantic Segmentation. *Ind. Electron. Appl.* **2016**, *11*, 2158–2297.
21. 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.* **2017**, *40*, 834–848. [[CrossRef](#)]
22. Hao, F.; Florent, L. Pyramid Scene Parsing Network in 3D: Improving Semantic Segmentation of Point Clouds with Multi-Scale Contextual Information. *ISPRS J. Photogramm. Remote Sens.* **2019**, *154*, 246–258.
23. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the International Conference on Medical Image Computing & Computer-assisted Intervention, Munich, Germany, 5–9 October 2015; Volume 351, pp. 234–241.
24. He, K.; Zhang, X.; Ren, S. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
25. Lofte, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv* **2015**, arXiv:1502.03167.
26. Betere, I.J.; Kinjo, H.; Nakazono, K.; Oshiro, N. Investigation of multi-layer neural network performance evolved by genetic algorithms. *Artif. Life Robot.* **2019**, *24*, 183–188. [[CrossRef](#)]
27. Yan, H.; Wang, P.; Dong, Y.; Luo, C.; Li, H. Image Classification and Identification Method Based on Improved Convolutional Neural Network. *Comput. Appl. Softw.* **2018**, *35*, 193–198.
28. Lukmandono, M.; Basuki, M.; Hidayat, J. Application of Saving Matrix Methods and Cross Entropy for Capacitated Vehicle Routing Problem (CVRP) Resolving. In Proceedings of the IOP Conference Series: Materials Science and Engineering, Surabaya, Indonesia, 29 September 2018.
29. Bengio, S.; Bengio, Y.; Courville, A. Why Does Unsupervised Pre-training Help Deep Learning? *J. Mach. Learn. Res.* **2010**, *11*, 625–660.
30. Tieleman, T.; Hinton, G. RMSProp: Divide The Gradient by A Running Average of Its Recent Magnitude. COURSE: Neural Networks for Machine Learning. *Tech. Rep.* **2012**, *4*, 26–31.
31. Garciagarcia, A.; Ortsescolano, S.; Oprea, S. A Review on Deep Learning Techniques Applied to Semantic Segmentation. *arXiv* **2017**, arXiv:1704.06857.