agriculture

MDPI

Article

AM-UNet: Field Ridge Segmentation of Paddy Field Images Based on an Improved MultiResUNet Network

Xulong Wu [1,2], Peng Fang [1,2], Xing Liu [1,2], Muhua Liu [1,2], Peichen Huang [3], Xianhao Duan [1,2], Dakang Huang [1,2] and Zhaopeng Liu [1,2,*]

[1] College of Engineering, Jiangxi Agricultural University, Nanchang 330045, China; 18397804248@163.com (X.W.); fangpeng@jxau.edu.cn (P.F.); qqqq34782681@163.com (X.L.); liumuhua2024@163.com (M.L.); duan2022573466@gmail.com (X.D.); jxau613@163.com (D.H.)
[2] Jiangxi Key Laboratory of Modern Agricultural Equipment, Nanchang 330045, China
[3] College of Automation, Zhongkai University of Agriculture and Engineering, Guangzhou 510225, China; michaelpayson@163.com
* Correspondence: lzp841108@jxau.edu.cn; Tel.: +86-138-7483-8211

Abstract: In order to solve the problem of image boundary segmentation caused by the irregularity of paddy fields in southern China, a high-precision segmentation method based on the improved MultiResUNet model for paddy field mapping is proposed, combining the characteristics of paddy field scenes. We introduce the attention gate (AG) mechanism at the end of the encoder–decoder skip connections in the MultiResUNet model to generate the weights and highlight the response of the field ridge area, add an atrous spatial pyramid pooling (ASPP) module after the end of the encoder down-sampling, use an appropriate combination of expansion rates to improve the identification of small-scale edge details, use 1×1 convolution to improve the range of the sensory field after bilinear interpolation to increase the segmentation accuracy, and, thus, construct the AM-UNet paddy field ridge segmentation model. The experimental results show that the IoU, precision, and F1 value of the AM-UNet model are 88.74%, 93.45%, and 93.95%, respectively, and that inference time for a single image is 168ms, enabling accurate and real-time segmentation of field ridges in a complex paddy field environment. Thus, the AM-UNet model can provide technical support for the development of vision-based automatic navigation systems for agricultural machines.

Keywords: segmentation of ridges on fields; deep learning; attention gate mechanism; atrous spatial pyramid pooling

Citation: Wu, X.; Fang, P.; Liu, X.; Liu, M.; Huang, P.; Duan, X.; Huang, D.; Liu, Z. AM-UNet: Field Ridge Segmentation of Paddy Field Images Based on an Improved MultiResUNet Network. *Agriculture* **2024**, *14*, 637. https://doi.org/10.3390/agriculture14040637

Academic Editor: Maciej Zaborowicz

Received: 15 March 2024
Revised: 16 April 2024
Accepted: 19 April 2024
Published: 21 April 2024

1. Introduction

Unmanned driving in an agricultural setting is the technical core and construction foundation of intelligent farm machinery [1]. In recent years, as the construction of unmanned farms has gradually increased, agricultural unmanned machines have been successfully applied to large areas in standardized farms [2]. However, in the southern hilly areas of China, the construction of standardized farmland is relatively lagging behind, the farmland boundaries tend to form irregular curves or folded lines [3], and the farmland boundary positioning information on which the unmanned agricultural machinery relies is not easy to obtain. In particular, it is difficult and labor-intensive to manually obtain the boundary in paddy field environments; additionally, the boundary is mostly formed after cyclic cultivation, and the frequent modification of its edge ridges will also lead to poor timeliness of the point-picking work [4]. Therefore, the development of an environment-aware real-time detection technology for border ridge agricultural machinery is of great practical significance and scientific value in order to promote unmanned operations on farmlands with irregular borders in the hilly areas of south China.

Border ridge identification is the basis for obtaining boundary positioning information, and there are two main ways to achieve this using environment perception technology,

namely, the use of LiDAR or machine vision to obtain environmental information. Varghese et al. [5] used LiDAR point cloud tracking and modified the convex hull formation algorithm to decompose point cloud data, thus realizing the accurate extraction of building boundaries. Sun et al. [6] proposed a redundant feature point filtering algorithm, whereby LiDAR point cloud data were filtered and road boundaries were accurately segmented through the method of fusing boundary features with vehicle prediction trajectories, with recognition accuracy reaching 93%. Chen et al. [7] used the local maxima filtering algorithm and marking control watershed segmentation algorithm to detect the single-tree canopies of fruit trees, and extracted single-tree canopy information based on the canopy contour features. Chen et al. [8] pre-processed images using the Cb component in the YCr-Cb color space and carried out filtering, threshold segmentation, and morphological operations to extract the rice–wheat harvesting boundary line; the processing of a single image took 0.64 s. Hou et al. [9] used visual algorithms such as coarse extraction and dynamic segmentation of the lane line region to complete real-time detection of lane line boundaries. Cheng et al. [10] proposed a non-local mean denoising method based on integral graph acceleration and Tukey's bi-weight kernel function to achieve wear detection on tool wear images by means of morphological reconstruction of extreme points extracted from the wear region. These studies indicate that devices based on LiDAR and cameras are the main sensor devices used for boundary detection; however, there are significant differences between the boundary detection methods, and there are still problems such as the singular nature of scenes and low segmentation accuracy.

Traditional image segmentation mainly uses the Otsu method to obtain the threshold required for image segmentation. The Otsu method uses the image histogram to select the appropriate threshold to maximize the variance between the target and the background to achieve the purpose of image segmentation. When the difference in the gray value of the image is not significant, it is difficult to achieve accurate image segmentation [11]. In traditional image processing, to extract local features such as color, texture, and shape through expert knowledge and complex parameter adjustments. However, manual feature extraction methods do not fully represent image semantics, and extractors are generally application-specific and are poor in generalization and robustness [12]. Due to the dynamic characteristics of paddy fields, such as uneven terrain, changing weed conditions, varying weather conditions, and fluctuating lighting, traditional image processing methods lack the robustness to cope with the complex and ever-changing environment of paddy fields and advanced machine learning algorithms need to be considered [13]. Convolutional neural networks, with strong feature-learning ability, can be effectively used for the extraction of features using image spatial information [14,15], and good results have been achieved in terms of the segmentation of farmland areas [16,17]. In recent years, image segmentation methods based on deep neural networks have been widely used for farmland image segmentation and boundary extraction from UAV remote sensing images [18–21]. Fully convolutional networks can automatically extract deep semantic features, thus realizing end-to-end learning, and are more accurate and faster than the traditional support vector machine methods. These studies show that deep learning semantic segmentation can achieve pixel-level accurate segmentation with good generalization ability and robustness, which provides technical support for our subsequent research on navigation path extraction and the development of vision-based automatic navigation systems for agricultural machinery. The UNet model, due to its symmetric encoder–decoder architecture, has been applied for the semantic segmentation of medical images with great success [22], and has been increasingly used for agri-environmental image segmentation. Song et al. [23] added skip connections and depth-separable convolution on the basis of the SegNet segmentation model to achieve the effect of recognizing sunflower collapse in high-resolution remote sensing images. Ibtehaz et al. [24] introduced the MultiRes module with a residual structure, which achieved significant segmentation results on a more difficult small-sample medical dataset. Diao et al. [25] replaced the DoubleConv structure in the traditional UNet network by the ASPP structure to generate an ASPP-UNet network with a better row segmentation

performance for maize crop, which successfully meets the accuracy and real-time demands of agricultural robot vision navigation, and can function effectively under varying environmental pressures. Chen et al. [26] proposed a detection model for sesame and weeds based on YOLOv4. They integrated an attention mechanism into the SPP structure by using local importance pooling, and introduced an adaptive spatial feature fusion structure at the feature fusion level. This effectively improved the detection accuracy while maintaining a fast detection speed.

In agriculture, image processing tasks for paddy field environments present unique challenges due to the complex background and varying object scales. Objects in these fields exhibit significant scale differences within images, necessitating models with robust multi-scale information processing capabilities. The ASPP technique offers significant advantages in this context, effectively enhancing the model's ability to process multi-scale information and accurately capture semantic details across different object scales. Furthermore, research based on the UNet network combined with ASPP technology demonstrates the applicability of ASPP in the UNet framework. Integrating ASPP technology with the UNet network effectively enhances the model's ability to perceive complex backgrounds and object boundaries. Particularly, with the addition of an attention mechanism, the model's semantic segmentation ability is further improved, resulting in more accurate and refined semantic segmentation of paddy field environments while maintaining rapid detection speeds.

For the semantic segmentation of field ridges in paddy field environments, this study proposes a network structure combining an attention gate (AG) module and an atrous spatial pyramid pooling (ASPP) module based on the MultiResUNet model, and the feasibility and high accuracy of the proposed model are verified through an ablation experiment and model comparison experiment. The main points of improvement are as follows:

(1) Adding attention gate structure skip connections at the end of encoder–decoder skip connections in MultiResUNet, generating attention coefficients through sigmoid activation and passing them to the input coding layer using trilinear interpolation to highlight semantic regions related to the field ridges during up-sampling, and suppressing target-independent feature responses.
(2) Adding the ASPP module after the end of coding down-sampling, using a smaller combination of expansion rates to improve the identification of field ridge edge details, adding one-way parallel average pooling to integrate global information, and using 1×1 convolution to achieve channel dimensionality reduction after bilinear interpolation, in order to enhance the range of sensory fields for feature semantics and further improve the segmentation accuracy of the field ridge region.

2. Materials and Methods

2.1. Data Acquisition

The agricultural field information acquisition system based on the Yanmar rice transplanter used in the experiment mainly included a ZED camera (StereoLabs, San Francisco, CA, USA), an NVIDIA TX2 (Nvidia, Santa Clara, CA, USA), a display, and so on, as shown in Figure 1. The ZED camera was fixed directly above (50 cm) the front of the vehicle using a flip bracket, and the camera turning angle and height could be adjusted. The camera's properties include a frame rate of 100 FPS, a field of view of 90°60°110°, an aperture of f/2.0, and a depth range of 0.5–20 m. The embedded processor of choice—the NVIDIA TX2—is based on the Pascal™ (Bellevue, WA, USA) architecture and has 8 GB of memory, as well as 59.7 GB/s memory bandwidth.

The paddy field image acquisition time period was divided into four phases: 18 January 2020, 5 March 2020, 18 April 2021, and 5 June 2021. The acquisition times were between 7:00 and 11:00 and 14:00 and 17:00, including three time periods in both the morning and afternoon, and the weather included cloudy and sunny days. The image acquisition locations were Nanchang City, Jiangxi Province (28°64′80″ E, 115°68′75″ N), and Yichun City,

Jiangxi Province (27°94′73″ E, 114°26′21″ N), as shown in Figure 2. For each experiment, the rice transplanter was driven along the edge of the field, the distance between the front wheels and the field was 40 cm, the driving speed was 0.75 m/s, the camera was set to shoot diagonally at a 45° pitch angle, the height of the camera from the ground was 40 cm, and the image acquisition interval was 0.1 s. In total, 840 samples of paddy field images with a complete set of scenes and clear features of the field were screened.

Figure 1. Data acquisition platform.

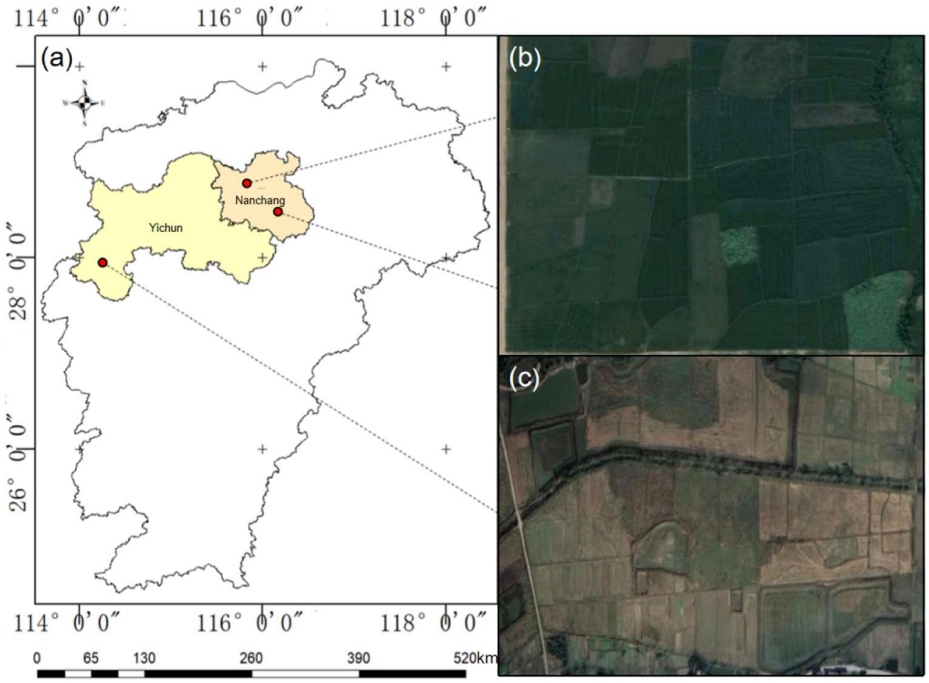

Figure 2. Image acquisition locations: (**a**) map of Jiangxi Province; (**b**) Nanchang City; and (**c**) Yichun City.

2.2. Dataset Construction

Due to the complex environment of the paddy fields, data acquisition in the ploughed paddy fields was carried out in the early, middle, and late rice cultivation periods. The paddy field ridge images acquired in the early and middle rice cultivation periods were mainly composed of six types of image, namely: scenes of green vegetation cover, no vegetation cover, water surface reflections, water surface shadows, uneven distributions of

the water surface, and rutted marks. The green vegetation field images were collected in early June, with a large amount of green vegetation on the surface of the field, typical of a paddy field after plowing and before rice transplanting. Images of fields without vegetation were collected in early January (i.e., in winter), with withered vegetation covering the surface of the fields and a small amount of water frozen on the field surface, accompanied by part of the rice stubble stalks; this is typical of a fallow paddy field after harvesting in winter. Water surface reflections were associated with the outdoor light intensity and caused a loss in imaging performance due to localized whitening of the reflective region of the paddy field content; this phenomenon mostly occurred in summer at around midday. Shadow blocking, due to oblique illumination from the Sun causing shadows on the field, was mostly distributed in the morning and afternoon hours. The uneven distribution of the water surface was due to excessive water release after plowing, and the surface of the paddy field had localized bare mud surfaces. Rutted marks were generated after the operation of agricultural machinery, and were mostly distributed in the scenes where the surface water level was low and the field was not fully plowed.

In order to improve the data distribution balance and enhance the model's generalization ability, we applied various transformations to each original image, as shown in Figure 3. We employed data enhancement such as random rotation, shifting, cropping, and scaling, while color transformations include adding noise and adjusting the hue, saturation, and brightness values. To implement these transformations, we used Python (3.6) scripts to call OpenCV vision library functions to expand the collected 840 images of paddy field ridges to 3400 images.

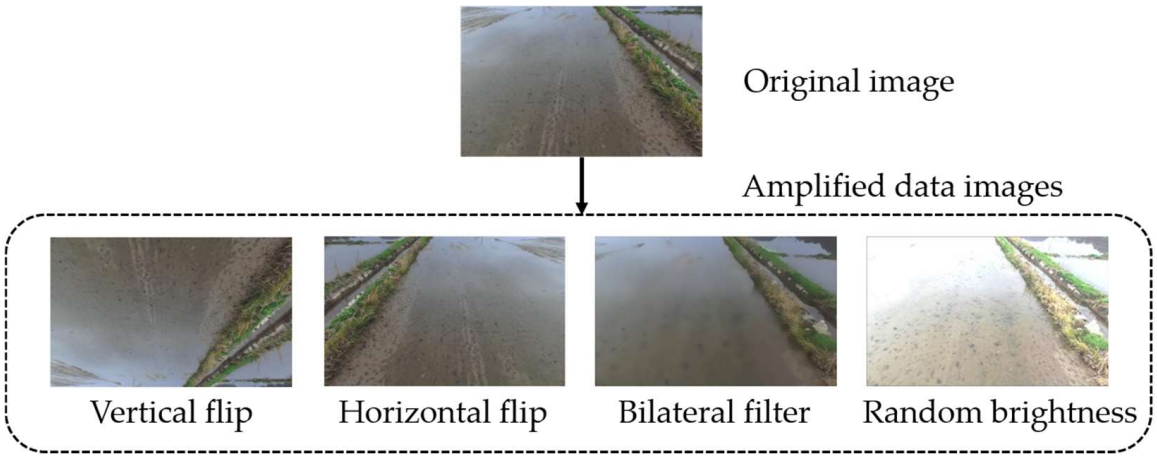

Figure 3. Schematic diagram of dataset supplementation.

2.3. Dataset Preparation

In this study, the paddy field ridge dataset was labeled at the pixel-level using the image annotation tool LabelMe (5.3.1) with manual visual annotation. As shown in Figure 4, the pixel value for a labeled field ridge area was 1, while the pixel value for a non-field ridge area was 255. The data were stored according to the format of publicly available dataset PASCAL VOC 2012, and the dataset was divided into a training set of 2380 images, a validation set of 680 images, and a test set of 340 images in a ratio of 7:2:1. Table 1 demonstrates the division of the dataset.

Table 1. Dataset segmentation.

Dataset	Proportions	Images
Training dataset	70%	2380
Validation dataset	20%	680
Test dataset	10%	340
Full data	100%	3400

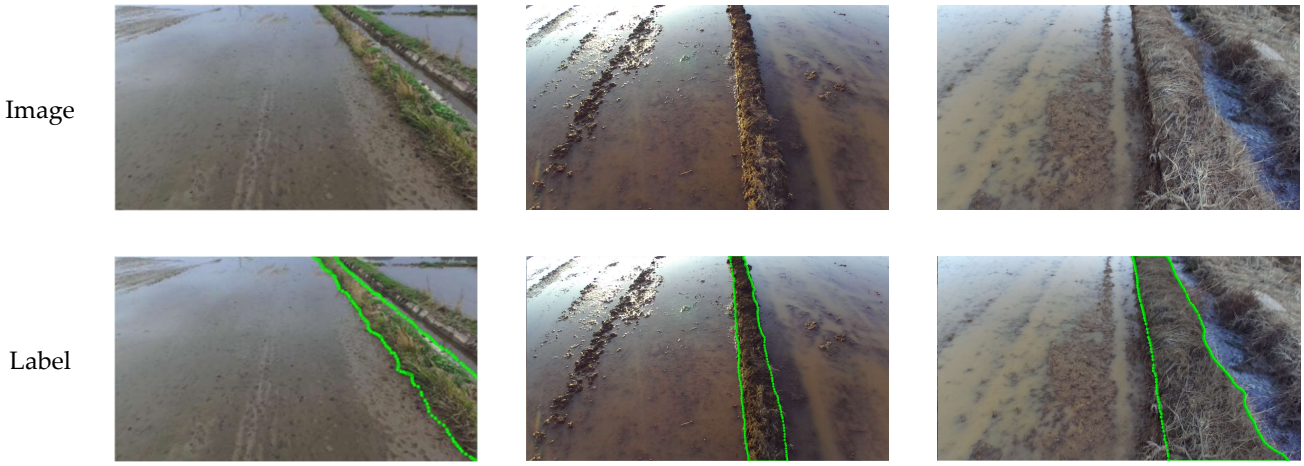

Figure 4. Tagging of the ridge area.

2.4. MultiResUNet Model

In this study, the image segmentation and extraction method for paddy field ridges was developed on the basis of the MultiResUNet network. The MultiResUNet model is mainly divided into an encoder–decoder architecture, as shown in Figure 5. In the encoding stage, the input image is first subjected to four instances of MultiResBlock convolution, which contains three 3 × 3 serial convolutions and 1 × 1 residual connection for image feature extraction, in order to enhance its multi-scale semantic information representation ability. A maximum pooling operation with a size of 2 × 2 and a step size of 2 is performed after each convolution. Subsequently, in the decoding stage, the extracted feature matrix is mapped and up-sampled using a 2 × 2 transpose matrix, after which the features are convolved through four MultiResBlock convolutions. The ResPath residual join is also introduced in the encoder–decoder stage, in order to ensure depth consistency before and after the join, thus improving the prediction accuracy.

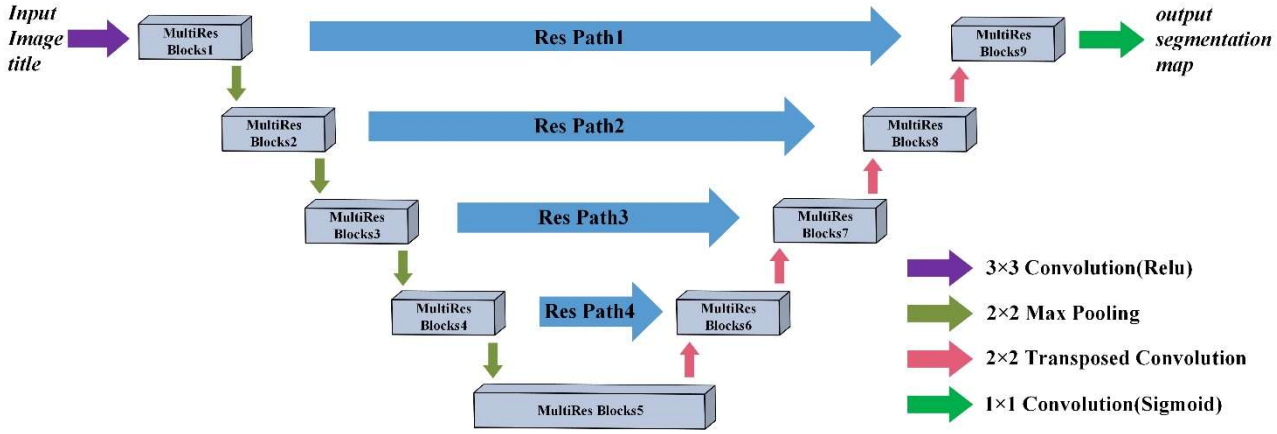

Figure 5. MultiResUNet network structure.

In this study, the paddy field scenes have strong environmental complexity, the field area is small compared to the background area, and the number of field data is small. MultiResUNet has excellent segmentation advantages on datasets of complex scenes, and possesses strong multi-scale semantic extraction and classification ability, allowing it to quickly extract small field areas. It also has good applicability for the segmentation of field images that show strong characteristics of perspective (i.e., where near objects appear big and distant objects appear small).

2.4.1. MultiResBlock Module

Figure 6 shows the schematic structure of the MultiResBlock. In order to prevent too large a number of parameters from affecting the model training effect, the MultiResBlock convolutional structure uses three 3 × 3 convolutional kernels, in order to simulate the effect of 5 × 5 and 7 × 7 convolutional kernels and to enhance the non-linearity of the model. The memory requirement of the convolutional layers is increased through incrementally increasing the memory requirement of the convolutional layers from 1 to 3, instead of maintaining the same memory requirement of the convolutional layers for all three layers. Furthermore, to obtain additional spatial information and to maintain the image size, a residual join is added and a 1 × 1 convolutional layer is introduced, which is finally fused using an add operation to form the MultiResBlock module. The model adds several convolutional layers to strengthen the depth of the network and avoid the gradient disappearance or explosion phenomena that tend to occur in deep neural networks during the training process.

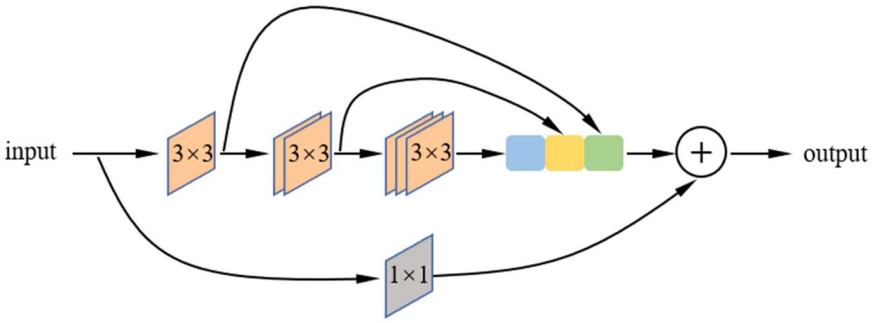

Figure 6. MultiResBlock structure.

2.4.2. Res Path Structure

The Res Path structure is shown in Figure 7. The use of skip connections in the MultiResUNet structure enables semantic information to be passed from the encoder to the decoder and, as there are large semantic differences between the shallow features in the encoder and the corresponding deep features in the decoder, directly splicing the two will affect the accuracy of the prediction results. Therefore, in order to reduce the difference between the encoder and decoder features, the encoder features are functionally mapped using a convolution operation before connecting them with the corresponding features in the decoder. Using 4, 3, 2, and 1 convolutional blocks in the Res Path, respectively, the encoder–decoder structure maintains a consistent depth before connecting through the non-linear operation of 3 × 3 convolutional layers with a 1 × 1 residual structure.

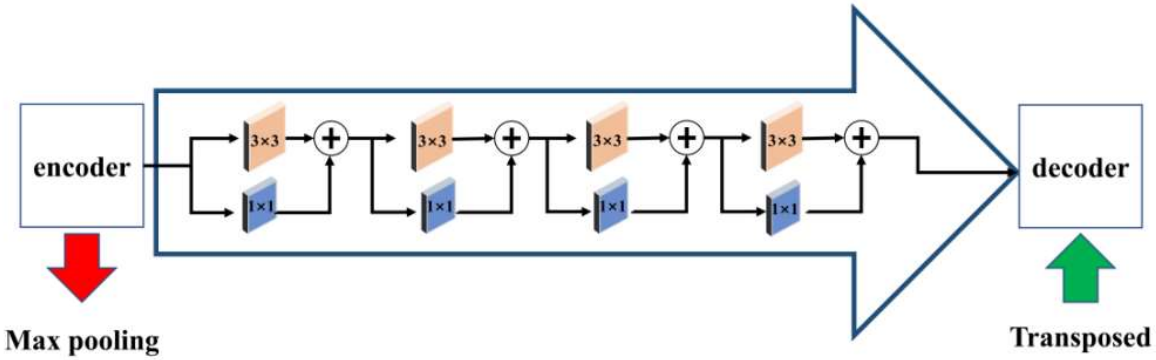

Figure 7. Res Path structure.

2.5. AM-UNet Semantic Segmentation Model for Paddy Field Ridges

The research object of this study is paddy field ridges, which are similar in appearance to the background color and texture of the paddy field. The background (paddy field surface) also accounts for a much larger proportion of the image than the target area (field ridges), and its specific differentiation is reflected in the edges of the field ridge area, as well as the degree of severance between the local area of the field ridges and the overall paddy field. Therefore, the segmentation task is more sensitive to the model's context awareness and its ability to correctly transfer multi-scale semantic information. In the original model, the Res Path structure has the problem of insufficient fusion of multi-scale feature information in the process of semantic information transfer, resulting in inaccurate segmentation of paddy fields. Therefore, in this study, an attention gate (AG) mechanism is added to the front end of the Res Path, which highlights the significant target region and suppresses the background region response that is not related to the segmentation target, in order to improve the accuracy of semantic information transfer. Meanwhile, the MultiResBlock module in the fifth down-sampling of the encoder and the up-sampling of the decoder is replaced with an ASPP null-space module, which improves the model's ability to extract multi-scale features. Based on the above improvements, an AM-UNet semantic segmentation model for paddy fields is proposed in this study. The improved AM-UNet semantic segmentation network model structure is shown in Figure 8.

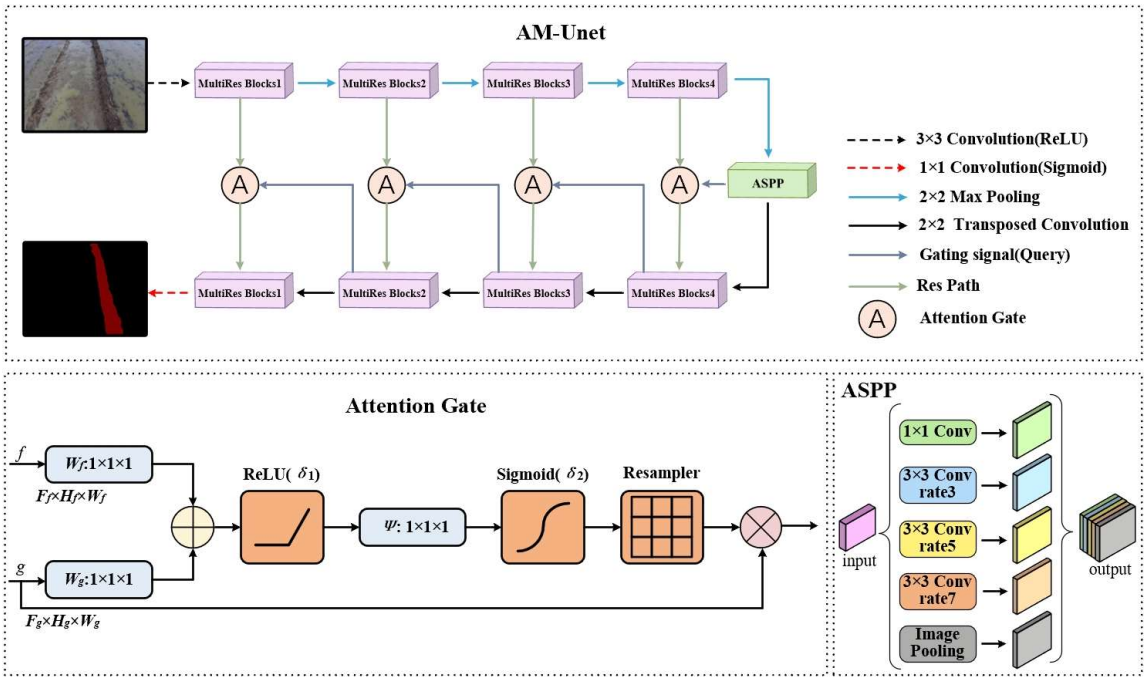

Figure 8. AM-UNet semantic segmentation model.

2.5.1. Attention Gate (AG) mechanism

The main focus of this study is a binary classification problem based on paddy fields and ridges, wherein the background paddy field region is large and the target ridge features account for a relatively small proportion. Therefore, in order to inhibit the feature response of the background region of paddy fields and to highlight the feature response of the ridge region, an attention gate (AG) [27] is introduced as a spatial attention module, which can be expressed using the following mathematical formulae:

$$T_{att}^{j} = \xi^{k}\left(\varepsilon_1\left(W_X^k X_i^j + W_g^k g_i + b_g\right)\right) + b_\xi, \tag{1}$$

$$a_i^j = \varepsilon_2\left(T_{att}^j\left(x_i^j, g_i; \Theta_{att}\right)\right), \tag{2}$$

where ε_1 is the ReLU function; ε_2 is the Sigmoid function; W_g^k, W_x^k, and ξ_k are convolution operations; and b_g and b_ξ are bias terms. An additive attention map T_{att}^j can be obtained after the computation of Equation (1), following which a 1×1 convolution operation is performed on it. Based on this, the Sigmoid computation of Equation (2) is introduced to compress the features, output the weight values within the range [0, 1], and obtain the soft attention map.

The structure of the AG attention mechanism is shown in Figure 9, where g is the feature matrix of the decoding region and x is the feature matrix of the encoding region. After an additive attention map is obtained through summing the g and x features, a soft attention map is obtained by applying the ReLU activation function, 1×1 convolution, and Sigmoid activation. To ensure that the image size is constant, a Resampler resampling step is added after the activation function and the feature representation of the background region is suppressed. Introducing the AG module into the skip connection of the UNet network structure can significantly enhance the feature extraction ability of the network for the paddy field ridge region, effectively suppressing the influence of the background region on target segmentation.

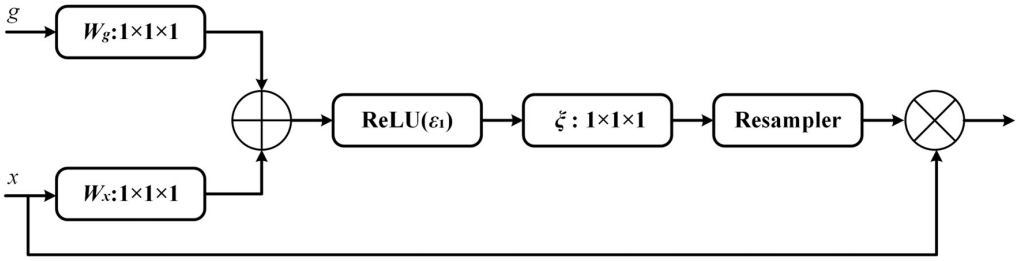

Figure 9. Attention gate mechanism structure.

2.5.2. Atrous Spatial Pyramid Pooling (ASPP) Module

ASPP was first proposed for the DeepLabv2 [28] model, and has been widely used in the field of image semantic segmentation due to its superior multi-scale feature extraction performance. In this study, we use an ASPP module composed of 3×3 null convolutions with expansion rates of 1, 3, 5, and 7, and a global average pooling layer in parallel. We add a Batch Normalization layer and ReLU activation function after each parallel null convolution operation, and use a cascade operation to fuse the features of each channel, allowing for better extraction of the fine-scale land features in the images of the ridges and furrows. The structure of ASPP is shown in Figure 10.

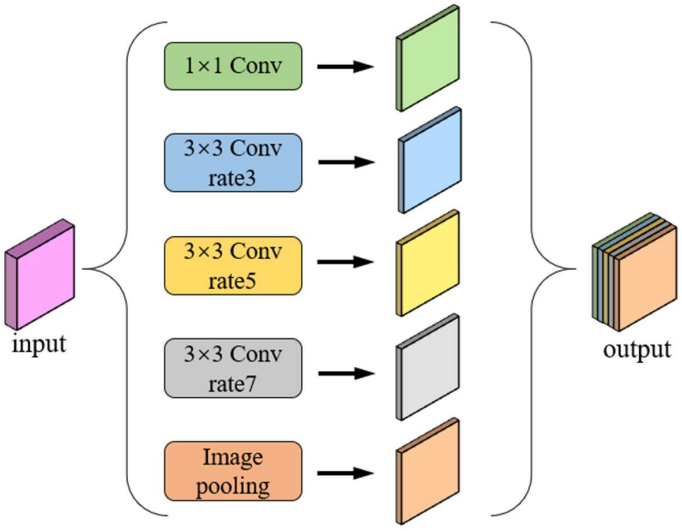

Figure 10. ASPP module.

3. Results and Analysis

3.1. Platform Parameters

The experimental hardware environment was a Lenovo P720 workstation with 256 G of running memory, an NVIDIA model RTX5000 GPU, and an Intel Xeon Gold 6142 CPU (2.6 GHz). The running environment was based on the Ubuntu 16.04 operating system, CUDA version 10.0, and CUDNN version 7.4.0, and was implemented in the pytorch deep learning framework using the Python programming language.

3.2. Evaluation Indices

In order to quantify the segmentation performance of the AM-UNet model on the paddy field ridge dataset, we selected the intersection over union (*IoU*), precision (*P*), and *F*1 score as metrics to evaluate the performance of the network structure [29]. The formulae for these evaluation indices are as follows:

$$IoU = \frac{TP}{TP + FN + FP}, \tag{3}$$

$$Precision = \frac{TP}{TP + FP}, \tag{4}$$

$$Recall = \frac{TP}{TP + FN}, \tag{5}$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}, \tag{6}$$

where *TP* indicates that the model prediction and label values are both true, *FP* indicates that the prediction value is true and the label value is false, and *FN* indicates that the prediction value is false and the label value is true. *IoU* is a common evaluation metric, which is used here to indicate the degree of overlap between the mask predicted by the model and the label mask. The *F*1 value is the harmonic mean of the precision and recall, making it a combined evaluation metric of the precision and recall.

3.3. Evaluation Index

The stochastic gradient descent (SGD) [30] used in the MultiResUNet model is an improved algorithm based on batch gradient descent (BGD). It approximates the loss function with a first-order Taylor expansion and solves for the minimum value of the approximated function, which is then used as the initial value for the next iteration. Although the convergence speed is fast on large data samples and an optimal solution can be obtained without training on all data, it is highly dependent on the current batch and the update is unstable, presenting periodic oscillations. Therefore, the improved AM-UNet network training procedure was optimized for convergence using Adam's first-order algorithm [31]. The loss function used in the training process was Dice Loss [32], with the following formulae:

$$\text{dice} = 1 - \frac{2|X \cap Y|}{|X| + |Y|}, \tag{7}$$

where X is the sum of predicted field pixels and Y is the sum of real field pixels.

In this study, we used the Adam first-order optimization algorithm in comparison with the SGD algorithm, combined with momentum and the RMSProp algorithm, and synchronously optimized and updated the network parameters as follows:

$$w_t = w_{t-1} - \Phi_t \frac{\widehat{m_{dw}}}{\sqrt{\widehat{n_{dw}}} + \gamma}, \tag{8}$$

$$y_t = y_{t-1} - \Phi_t \frac{\widehat{m_{dy}}}{\sqrt{\widehat{n_{dy}} + \gamma}}, \tag{9}$$

where w_t, y_t, and Φ_t are the network weights, bias, and learning rate in the tth iteration, respectively; $\widehat{m_{dw}}$ and $\widehat{m_{dy}}$ are the bias-corrected momentums of w_t and y_t, respectively; $\sqrt{\sqrt{\widehat{n_{dw}}}}$ and $\sqrt{\widehat{n_{dy}}}$ are the bias-corrected values of w_t and y_t, respectively; and γ is a hyperparameter, the value of which is taken as 10^{-8} in this study. To accelerate the convergence of the network, exponential decay was used to regulate the learning rate of the iterations, calculated as follows:

$$\Phi = \Phi_0 g^{f(t/10)}, \tag{10}$$

where Φ is the learning rate in the current iteration, $f()$ denotes downward rounding, g is the rate decay factor, t is the number of current iterations, and Φ_0 is the initially defined learning rate.

The AM-UNet model was trained using the Adam first-order optimization algorithm and the SGD algorithm separately, in order to compare the performance of both algorithms with an increase in the number of iterations. The initial learning rate was set to 0.0001 for training, the exponential decay rate for first-order moment estimation was set to 0.9, the exponential decay rate for second-order moment estimation was set to 0.999, the batch size was set to 4, and the number of running iterations was 20,000. The training loss curve is shown in Figure 11. The blue curve presents the performance change when using the Adam first-order optimization algorithm. In the first 10,000 iterations of training, the loss value decreased greatly, and there were local oscillations; when the number of iterations reached 16,000, the loss value tended to flatten out, and the model was more stable; and, when the number of iterations reached 20,000, the loss value converged to 0.056. The red curve presents the performance change for the SGD stochastic gradient descent algorithm. As the number of iterations increased, the loss value gradually decreased. In the first 12,000 training iterations, the loss value decreased greatly and the oscillation amplitude was larger; when the number of iterations reaches 18,000, the loss value tended to be flat; and, at 20,000 iterations, the loss value converged to 0.092. In the SGD algorithm, each iteration's weights are changed in the training process, and therefore the model is more stable; however, as the weights of the SGD algorithm are adjusted in each iteration during the training process, the model can easily skip the optimal solution, and the oscillation causes the loss value to fluctuate up and down. Therefore, it can be seen that the use of the Adam first-order optimization algorithm resulted in a more stable model performance.

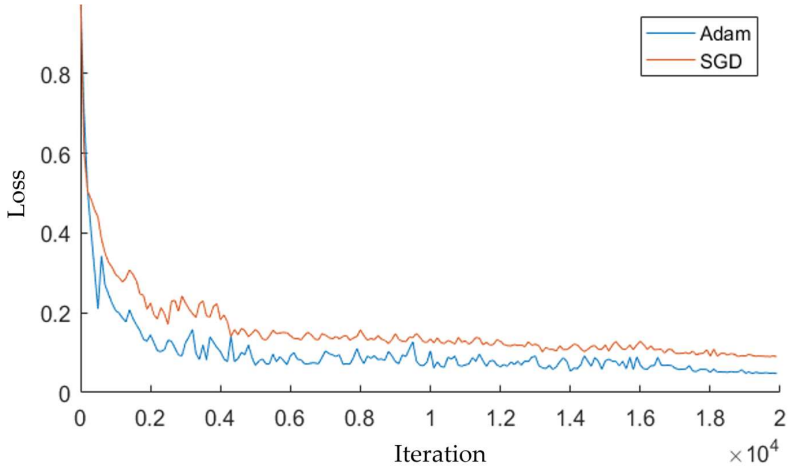

Figure 11. AM-UNet model training loss value change curve.

3.4. Ablation Experiment

In order to verify the contributions obtained through introducing the AG mechanism and ASPP structure into the model, the model using MultiResUNet as the backbone feature extraction network was used as the baseline model, and an ablation test was conducted on the basis of this model. The results of the experiment are shown in Table 2. The introduction of the AG module improved the IoU, precision, and F1 value on the paddy field ridge dataset by 0.6, 0.55, and 0.74 percentage points, respectively, with respect to the baseline model; the introduction of the ASPP structure improved the IoU, precision, and F1 value on the paddy field ridge dataset by 1.33, 0.92, and 1.38 percentage points, respectively, with respect to the baseline model. Referring to the AG module and the ASPP structure alone, each evaluation index was improved to different degrees, and the introduction of the ASPP structure resulted in a greater improvement than the AG module.

Table 2. Results of the paddy field ridge dataset ablation experiment.

Method	IoU	P	F1
Baseline	85.71	90.84	91.77
+AG	86.31	91.39	92.51
+ASPP	87.04	91.76	93.15
AG + ASPP	88.78	93.56	94.01

The introduction of the AG module and ASPP structure at the same time strengthened the model's ability to segment targets; in particular, the IoU, precision, and F1 value were improved by 3.07, 2.72, and 2.24 percentage points, respectively, on the paddy field ridge dataset. The AG module significantly enhanced the feature extraction ability of the network on the paddy field ridge region and effectively suppressed the influence of the background region on target segmentation. The ASPP structure plays an important role in the identification of small targets; it combines the semantics of the different sensory fields without losing information, better extracts the target image information, and better extracts the fine-scale features of landmarks in field images. Therefore, the AG module and ASPP structure proposed in this study improved the accuracy of the model and allowed it to better recognize the paddy field ridges.

3.5. Comparison of the Performance of Different Models

In order to further validate the classification accuracy of paddy field ridges under different features, 100 images of paddy field ridges categorized as type a (green vegetated ridges), type b (unvegetated ridges), type c (uneven distribution of water surfaces), type d (reflections on the water surface), type e (shaded occlusion), and type f (rutted marks) were selected as a test set from the 680 validation images and 340 test images in the paddy field ridge dataset. The related training parameter settings were kept the same. Based on these data, the experimental results obtained with the proposed method were compared with the results of the UNet, MultiResUNet, and PSPNet models, as these models were widely used in semantic segmentation due to their advantages of simple structure, high accuracy, and generalizability. The segmentation results are shown in Figure 12, and the performance metrics are given in Table 3.

Figure 11 shows the test results for the original and labeled images obtained with the UNet, MultiResUNet, PSPNet [33], and AM-UNet models for the six different types of southern paddy field images.

For segmentation of the most common paddy field scene—fields with green vegetation—the UNet model showed local background misclassification, while the PSP and MultiResUNet models were not fine enough in their edge recognition. The edge lines on both sides of the field were too smooth, which made it difficult to express the complexity of the edges of the field. In comparison, the improved AM-UNet model achieved the best results in terms of segmentation completeness and detail retention. For the segmentation of scenes with no

vegetation, a reduction in the number of weeds in the field led to an improvement in the distinction between field ridges and paddy fields, and the four models were able to carry out accurate segmentation of field ridges. For the segmentation of scenes with an uneven water distribution, due to the presence of bare mud surfaces in the paddy field area, the UNet and MultiResUNet models presented erroneous segmentation, while the segmentation results of the PSPNet model were incomplete and the recognition effect was not good. Again, the AM-UNet model had the best results in terms of segmentation completeness and detail preservation, and therefore the AM-UNet model performed better in this scenario.

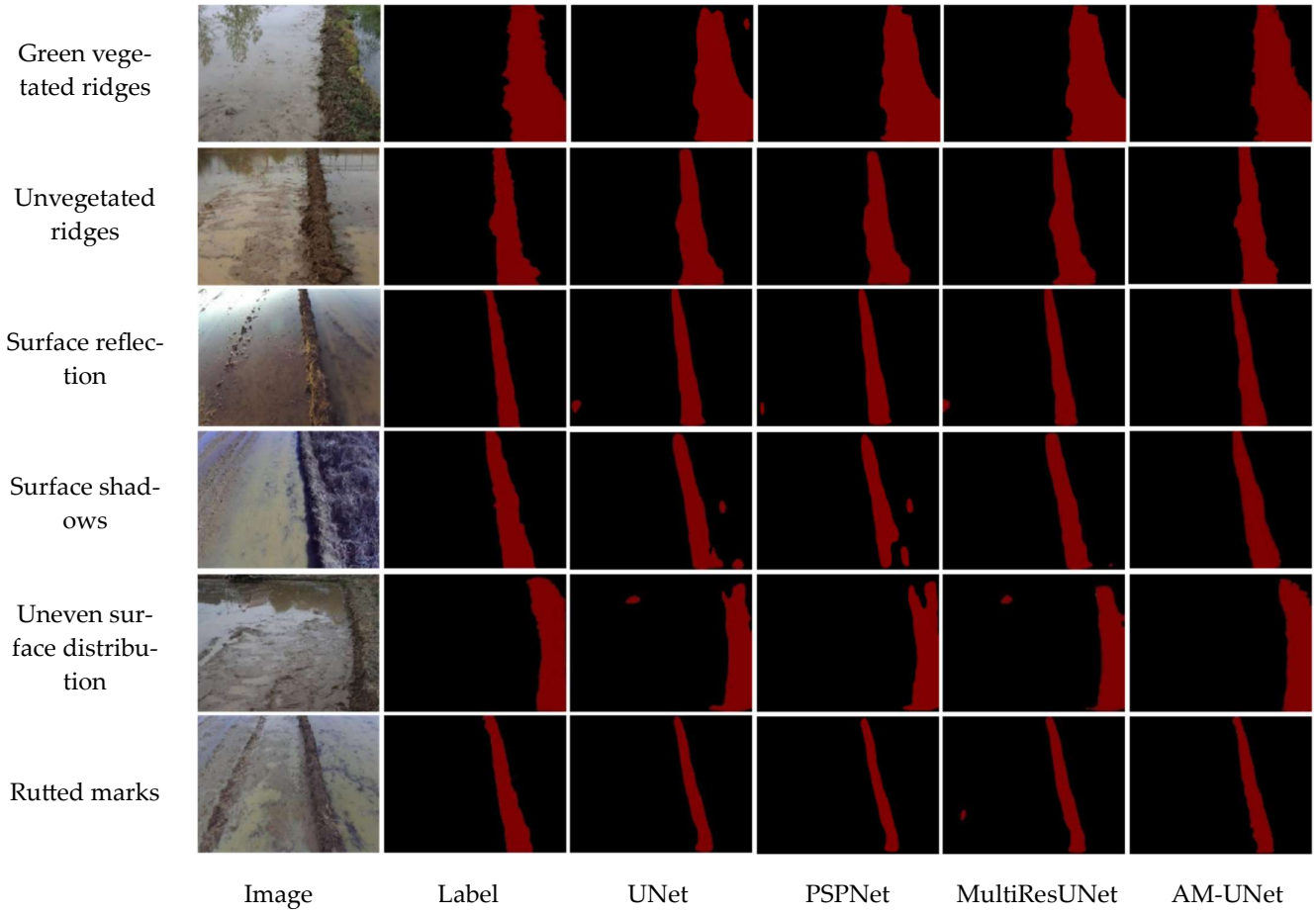

Figure 12. Comparison of different segmentation algorithms.

Table 3. Performance evaluation on different types of paddy field.

Type	UNet			PSPNet			MultiResUNet			AM-UNet		
	IoU	P	F1	IoU	P	F1	IoU	P	F1	IoU	P	F1
Green vegetated ridges	85.82	90.95	92.27	83.72	90.28	91.60	87.46	91.23	92.80	89.51	95.49	94.51
Unvegetated ridges	86.34	91.36	92.78	84.61	91.08	92.27	88.14	92.69	93.26	92.25	96.84	95.03
Surface reflection	82.89	89.17	91.06	81.24	88.46	90.20	84.83	90.23	91.69	87.13	92.18	93.68
Surface shadows	84.63	89.51	90.13	81.83	89.28	88.77	86.87	91.82	90.91	87.89	92.36	93.08
Uneven surface distribution	83.04	89.28	89.69	82.05	89.14	91.25	83.26	89.31	91.13	88.79	93.71	93.27
Rutted marks	82.72	88.86	91.52	80.27	88.13	89.71	85.36	91.09	92.55	86.87	90.14	94.14
Average	84.24	89.86	91.24	82.29	89.39	90.63	85.98	91.06	92.06	88.74	93.45	93.95

For the segmentation of fields with water reflection, the UNet, PSPNet, and MultiResUNet models all misidentified the soil clods in the paddy field, and did not achieve good results in recognizing the difference between local soil clods and the whole field. However,

the AM-UNet model accurately captured the difference between the two and realized accurate segmentation. From the comparison of scene segmentation under shadow masking, the segmentation results obtained with the UNet and PSPNet models showed different degrees of segmentation breaks in the field boundary region, mistakenly segmenting part of the field with shadow into the background paddy field. Both the MultiResUNet and AM-UNet models were able to accurately recognize the shadowed area. In the segmentation of scenes with rutted marks, the rutted marks formed a block similar to the field in the background area of the paddy field. The MultiResUNet model produced some incorrect segmentation in the rutted marks area. UNet and PSPNet did not present a high degree of completion of the segmentation of the field, given that the segmented area was smaller than the actual field area. In contrast, AM-UNet provided the most accurate segmentation in the rutted marks scenario.

In general, the AM-UNet model performed better than the other three models for field segmentation in scenes with green vegetated fields, unvegetated fields, water surface reflections, water surface shadows, an uneven distribution of the water surface, and rutted marks; it also had the highest field edge fit and greater robustness for field segmentation under multiple interference elements.

Table 3 provides the segmentation performance parameters of each model in the six types of paddy field scenarios, from which it can be seen that the mean precision of the AM-UNet model provided an improvement of 3.59, 4.06, and 2.39 percentage points, when compared with the UNet, PSPNet, and MultiResUNet models, respectively. Furthermore, the mean IoU of the improved AM-UNet model was better than that of the UNet, PSPNet, and MultiResUNet models by 4.5, 4.34, and 2.76 percentage points, respectively; and the F1 value of the AM-UNet model was improved by 2.71, 3.32, and 1.89 percentage points when compared with the UNet, PSPNet, and MultiResUNet models, respectively. Table 4 presents the training and inference times for different models. The AM-UNet model exhibited a noticeable increase in the number of network parameters compared to the UNet, PSPNet, and MultiResUNet models. However, considering the improvement in its segmentation capability, we deemed this increase acceptable.

Table 4. Performance evaluation on different models.

Model	IoU/%	F1/%	Training Time/h	Inference Time/ms
PSPNet	82.29	90.63	19.6	150
UNet	84.24	91.24	21.2	147
MultiResUNet	85.98	92.06	28.8	155
AM-UNet	88.74	93.95	31.2	168

Our comparative tests demonstrated that the improved AM-UNet model performed best, in terms of all evaluation indices, and achieved significant segmentation effects in the test set of paddy field ridges under six different classification scenarios, with good robustness and precision, IoU, and F1 values of 93.45%, 88.74%, and 93.95%, respectively. The model's inference time for a single image is 168ms, indicating that it can perform approximately 5.95 inferences per second. Given that the camera's field of view remains consistent, with the semantically segmented paddy field ridges measuring about 5m in length in each image, it can accurately, and in real time, segment the ridges in the complex paddy field environment at the normal operating speed of the agricultural machine. This capability provides crucial technical support for developing a vision-based automated navigation system for agricultural machinery.

4. Discussion

In this study, we analyzed the environmental characteristics of paddy field ridges and proposed the AM-UNet model. This model incorporates an attention gate mechanism at the front-end of the encoder–decoder skip connections within the MultiResUNet architecture. Additionally, it integrates an atrous spatial pyramid pooling (ASPP) module following

the downsampling stage. These enhancements are designed to refine the segmentation accuracy of paddy field ridges and minimize the misidentification between the background and target regions. The results showed that the model had better segmentation accuracy and applicability along the edges of the paddy field ridges and the surrounding background area. When compared with the improved Floodfill algorithm [34] for paddy field ridge segmentation, our method did not require the input of preset points. Unlike traditional image processing, which relied on manual feature design and selection and was susceptible to interference from factors such as field lighting and road conditions which resulted in poor generalization ability, we adopted an end-to-end deep learning framework. This framework automatically learned image features through a multilevel network structure, thereby eliminating the need for manual feature design and extraction inherent in traditional methods. Our approach exhibited strong adaptability and generalization capabilities.

Although the model proposed in this study accomplishes accuracy and speed in segmentation, it still has some limitations. (1) The incorporation of the attention gate (AG) and atrous spatial pyramid pooling (ASPP) module, despite enhancing segmentation accuracy, introduces a reduction in the model's inference speed due to the computational demands of the kernel. At present, our model achieves an inference speed of 168 ms per single image. While this meets the real-time segmentation demand under the normal operating speed of agricultural machines, considering the time consumption of other algorithms in the subsequent visual navigation system, there is still room for further improvement in the model's inference speed. (2) The paddy field ridge dataset is deficient in paddy field ridge images captured within well-lit environments. As a consequence, the segmentation performance in well-lit environments remains unvalidated. In upcoming research endeavors, we will strive to acquire additional paddy field ridge images captured in well-illuminated natural settings. Subsequently, these images will be integrated into the paddy field ridge dataset to validate the segmentation efficacy of the AM-UNet model across various environmental contexts. Additionally, we intend to optimize the structure of the enhanced network deployment by leveraging TensorRT acceleration technology to reduce the number of model parameters. This optimization aims to enhance segmentation efficiency and achieve real-time segmentation. Furthermore, we plan to employ the inverse perspective mapping method on the segmented images to obtain the coordinate set of the paddy field ridge boundary. The navigation route is formulated according to its boundary coordinates, preparing the technology for the next step of guiding unmanned agricultural machines in a paddy field environment.

5. Conclusions

(1) In this study, a segmentation model based on the MultiResUNet model was constructed to address the difficult problem of accurate segmentation of paddy field images representing the paddy field environment in southern China. The improved model introduces an attention gate (AG) at the end of the encoder–decoder skip connection in the MultiResUNet model, highlights the feature response of the field ridge region, and introduces an atrous spatial pyramid pooling (ASPP) module after down-sampling the encoder to improve the recognition accuracy regarding the small-scale edge details of field ridges. These improvements allow the model to realize the accurate recognition and segmentation of field ridge images in a complex environment, providing technical support for the development of vision-based automatic navigation systems for agricultural machines.

(2) The experimental results show that the segmentation accuracy, average intersection over union, and average F1 value of the optimized model in the validation set were 93.45%, 88.74%, and 93.95%, respectively, better than those obtained with the UNet, MultiResUNet, and PSPNet methods. When compared with the existing MultiResUNet model, the proposed model's segmentation accuracy, intersection over union, and average F1 value were improved by 2.39, 2.76, and 1.89 percentage points, respectively, and the inference time for a single image was 168ms, enabling accurate

and real-time segmentation of field ridges in a complex paddy field environment. Therefore, the introduction of the attention mechanism and spatial pyramid pooling significantly improved the segmentation effect of the model for paddy field ridge images.

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

Funding: This research was funded by the Key Project of Jiangxi Province, grant number JXNK2023080502; Jiangxi Province Unveiling and Commanding Project, grant number 20222-05125-03; Basic and Applied Basic Research of Guangzhou Basic Research Program in 2022, grant number 202201011691.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: Please contact the corresponding author for the model code and experimental data.

Acknowledgments: We are thankful to Xing Liu, Dakang Huang, Xianhao Duan, and Shan Li, who have contributed to our data collection.

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

References

1. Liu, C.; Lin, H.; Li, Y.; Gong, L.; Miao, Z. Analysis on Status and Development Trend of Intelligent Control Technology for Agricultural Equipment. *Trans. Chin. Soc. Agric. Mach.* **2020**, *51*, 1–18.
2. Li, D.; Li, Z. System Analysis and Development Prospect of Unmanned Farming. *Trans. Chin. Soc. Agric. Mach.* **2020**, *51*, 1–12.
3. Lu, G.; Jin, T.; Shen, W. Research on the Development of Rice Moderate Scale and Mechanization in Small Paddy Area in the South. *Agric. Mach.* **2016**, *12*, 69–71. [CrossRef]
4. Wang, J.; Weng, W.; Liu, J.; Wang, J.; Na, M. Design and Experiment of Bi-Directional Ridger for Paddy Field. *Trans. Chin. Soc. Agric. Mach.* **2019**, *50*, 40–48. [CrossRef]
5. Varghese, V.; Shajahan, D.A.; Nath, A.G. Building Boundary Tracing and Regularization from LiDAR Point Cloud. In Proceedings of the 2016 International Conference on Emerging Technological Trends (ICETT), Kollam, India, 21–22 October 2016; IEEE: Kollam, India, 2016; pp. 1–6.
6. Sun, P.; Zhao, X.; Xu, Z.; Wang, R.; Min, H. A 3D LiDAR Data-Based Dedicated Road Boundary Detection Algorithm for Autonomous Vehicles. *IEEE Access* **2019**, *7*, 29623–29638. [CrossRef]
7. Chen, R.; Li, C.; Yang, G.; Yang, H.; Xu, B.; Yang, X.; Zhu, Y.; Lei, L.; Zhang, C.; Dong, Z. Extraction of Crown Information from Individual Fruit Tree by UAV LiDAR. *Trans. Chin. Soc. Agric. Eng.* **2020**, *36*, 50–59. [CrossRef]
8. Chen, J.; Sun, J.; Chen, H.; Song, J. A Study on Real-Time Extraction of Rice and Wheat Harvest Boundary Line in Shadow Environment. *J. Agric. Mech. Res.* **2022**, *44*, 26–31.
9. Hou, C. Research on Vision-Based Lane Line Detection Technology. Ph.D. Dissertation, Southwest Jiaotong University, Chengdu, China, 2017.
10. Chen, X.; Yu, J. Monitoring Method for Machining Tool Wear Based on Machine Vision. *J. ZheJiang Univ. (Eng. Sci.)* **2021**, *55*, 896–904. [CrossRef]
11. Pandey, R.; Lalchhanhima, R. Segmentation Techniques for Complex Image: Review. In Proceedings of the 2020 International Conference on Computational Performance Evaluation (ComPE), Shillong, India, 2–4 July 2020; IEEE: Shillong, India, 2020; pp. 804–808.
12. Qiao, Y.; Liu, H.; Meng, Z.; Chen, J.; Ma, L. Method for the Automatic Recognition of Cropland Headland Images Based on Deep Learning. *Int. J. Agric. Biol. Eng.* **2023**, *16*, 216–224. [CrossRef]
13. Yu, Y.; Bao, Y.; Wang, J.; Chu, H.; Zhao, N.; He, Y.; Liu, Y. Crop Row Segmentation and Detection in Paddy Fields Based on Treble-Classification Otsu and Double-Dimensional Clustering Method. *Remote Sens.* **2021**, *13*, 901. [CrossRef]
14. Peng, B.; Guo, Z.; Zhu, X.; Ikeda, S.; Tsunoda, S. Semantic Segmentation of Femur Bone from MRI Images of Patients with Hematologic Malignancies. In Proceedings of the 2020 IEEE Region 10 Conference (TENCON), Osaka, Japan, 16–19 November 2020; pp. 1090–1094.
15. Trebing, K.; Stanczyk, T.; Mehrkanoon, S. SmaAt-UNet: Precipitation Nowcasting Using a Small Attention-UNet Architecture. *Pattern Recognit. Lett.* **2021**, *145*, 178–186. [CrossRef]
16. He, Y.; Zhang, X.; Zhang, Z.; Fang, H. Automated Detection of Boundary Line in Paddy Field Using MobileV2-UNet and RANSAC. *Comput. Electron. Agric.* **2022**, *194*, 106697. [CrossRef]

17. Wang, S.; Su, D.; Jiang, Y.; Tan, Y.; Qiao, Y.; Yang, S.; Feng, Y.; Hu, N. Fusing Vegetation Index and Ridge Segmentation for Robust Vision Based Autonomous Navigation of Agricultural Robots in Vegetable Farms. *Comput. Electron. Agric.* **2023**, *213*, 108235. [CrossRef]
18. Marshall, M.; Crommelinck, S.; Kohli, D.; Perger, C.; Yang, M.Y.; Ghosh, A.; Fritz, S.; de Bie, K.; Nelson, A. Crowd-Driven and Automated Mapping of Field Boundaries in Highly Fragmented Agricultural Landscapes of Ethiopia with Very High Spatial Resolution Imagery. *Remote Sens.* **2019**, *11*, 2082. [CrossRef]
19. Xu, L.; Ming, D.; Zhou, W.; Bao, H.; Chen, Y.; Ling, X. Farmland Extraction from High Spatial Resolution Remote Sensing Images Based on Stratified Scale Pre-Estimation. *Remote Sens.* **2019**, *11*, 108. [CrossRef]
20. Waldner, F.; Diakogiannis, F.I. Deep Learning on Edge: Extracting Field Boundaries from Satellite Images with a Convolutional Neural Network. *Remote Sensing of Environment* **2020**, *245*, 111741. [CrossRef]
21. Hong, R.; Park, J.; Jang, S.; Shin, H.; Kim, H.; Song, I. Development of a Parcel-Level Land Boundary Extraction Algorithm for Aerial Imagery of Regularly Arranged Agricultural Areas. *Remote Sens.* **2021**, *13*, 1167. [CrossRef]
22. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015, Proceedings of the 18th International Conference, Munich, Germany, 5–9 October 2015*; Springer: Cham, Switzerland, 2015.
23. Song, Z.; Zhang, Z.; Yang, S.; Ding, D.; Ning, J. Identifying Sunflower Lodging Based on Image Fusion and Deep Semantic Segmentation with UAV Remote Sensing Imaging. *Comput. Electron. Agric.* **2020**, *179*, 105812. [CrossRef]
24. Ibtehaz, N.; Rahman, M.S. MultiResUNet: Rethinking the U-Net Architecture for Multimodal Biomedical Image Segmentation. *Neural Netw.* **2020**, *121*, 74–87. [CrossRef]
25. Diao, Z.; Guo, P.; Zhang, B.; Zhang, D.; Yan, J.; He, Z.; Zhao, S.; Zhao, C. Maize Crop Row Recognition Algorithm Based on Improved UNet Network. *Comput. Electron. Agric.* **2023**, *210*, 107940. [CrossRef]
26. Chen, J.; Wang, H.; Zhang, H.; Luo, T.; Wei, D.; Long, T.; Wang, Z. Weed Detection in Sesame Fields Using a YOLO Model with an Enhanced Attention Mechanism and Feature Fusion. *Comput. Electron. Agric.* **2022**, *202*, 107412. [CrossRef]
27. Oktay, O.; Schlemper, J.; Folgoc, L.L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N.Y.; Kainz, B.; et al. Attention U-Net: Learning Where to Look for the Pancreas. *arXiv* **2018**. [CrossRef]
28. 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] [PubMed]
29. Garcia-Garcia, A.; Orts, S.; Oprea, S.; Villena-Martinez, V.; Rodríguez, J.G. A Review on Deep Learning Techniques Applied to Semantic Segmentation. *arXiv* **2017**, arXiv:1704.06857.
30. Ruder, S. An Overview of Gradient Descent Optimization Algorithms. *arXiv* **2016**. [CrossRef]
31. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980.
32. Milletari, F.; Navab, N.; Ahmadi, S.-A. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; pp. 565–571.
33. 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]
34. Xie, S.; Huang, W.; Zhu, L.; Yang, C.; Zhang, S.; Fu, G. Vision Navigation System of Farm Based on Improved Floodfill Method. *J. Chin. Agric. Mech.* **2021**, *42*, 182. [CrossRef]