Recognition of field roads based on improved U-Net++ Network

Lili Yang[1,2], Yuanbo Li[1,2], Mengshuai Chang[1,2], Yuanyuan Xu[1,2], Bingbing Hu[3], Xinxin Wang[1,2], Caicong Wu[1,2*]

(1. *College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China*;
2. *Key Laboratory of Agricultural Machinery Monitoring and Big Data Application, Ministry of Agriculture and Rural Affairs, Beijing 100083, China*;
3. *Kunlun Beidou Intelligence Technologies Co. Ltd., Beijing 102200, China*)

Abstract: Unmanned driving of agricultural machinery has garnered significant attention in recent years, especially with the development of precision farming and sensor technologies. To achieve high performance and low cost, perception tasks are of great importance. In this study, a low-cost and high-safety method was proposed for field road recognition in unmanned agricultural machinery. The approach of this study utilized point clouds, with low-resolution Lidar point clouds as inputs, generating high-resolution point clouds and Bird's Eye View (BEV) images that were encoded with several basic statistics. Using a BEV representation, road detection was reduced to a single-scale problem that could be addressed with an improved U-Net++ neural network. Three enhancements were proposed for U-Net++: 1) replacing the convolutional kernel in the original U-Net++ with an Asymmetric Convolution Block (ACBlock); 2) adding a multi-branch Asymmetric Dilated Convolutional Block (MADC) in the highest semantic information layer; 3) adding an Attention Gate (AG) model to the long-skip-connection in the decoding stage. The results of experiments of this study showed that our algorithm achieved a Mean Intersection Over Union of 96.54% on the 16-channel point clouds, which was 7.35 percentage points higher than U-Net++. Furthermore, the average processing time of the model was about 70 ms, meeting the time requirements of unmanned driving in agricultural machinery. The proposed method of this study can be applied to enhance the perception ability of unmanned agricultural machinery thereby increasing the safety of field road driving.

Keywords: image segmentation, unmanned agricultural machinery, field roads, point cloud super-resolution, point cloud bird's eye view

DOI: 10.25165/j.ijabe.20231602.7941

Citation: Yang L L, Li Y B, Chang M S, Xu Y Y, Hu B B, Wang X X, et al. Recognition of field roads based on improved U-Net++ Network. Int J Agric & Biol Eng, 2023; 16(2): 171–178.

1 Introduction

The scarcity of labor in agricultural production currently hinders its expansion on a broad scale, and unmanned and intelligent agricultural technology is increasingly seen as a key solution to this issue[1,2]. During the busy farming season, the lighting conditions in agricultural production settings fluctuate drastically, and agricultural machinery must travel between hangars and farmland day and night. Unlike structured urban roads, field roads lack clear boundaries, and there are various obstacles such as tree shadows and weeds on both sides of the road, which make identifying road sections more challenging. Lidar has anti-interference characteristics and is not affected by lighting conditions[3,4], making it a popular choice for obstacle detection[5], tracking[6], and road recognition[7] in unmanned driving. Road recognition based on Lidar can be divided into traditional and deep learning methods. Compared with the complex process of traditional methods[8,9], 3D convolution of point clouds[10] and 2D convolution of point clouds projection views[11] have self-learning capability and significantly improve the accuracy of road semantic segmentation. The input data types for 3D convolution in road semantic segmentation are points and voxels. Although the original point clouds as input to a neural network preserve 3D properties, they are computationally demanding and unsuitable for real-time segmentation[12,13]. The representation of point clouds as standard 3D voxel grids can result in empty voxel grids[14], which leads to redundant calculations in voxel-based 3D convolutional road segmentation and a long average inference time, negatively impacting real-time performance. By contrast, 2D convolution methods of point cloud projection views transfer point clouds into Bird's Eye View (BEV)[11] or Front View (FV)[15], reducing computational costs.

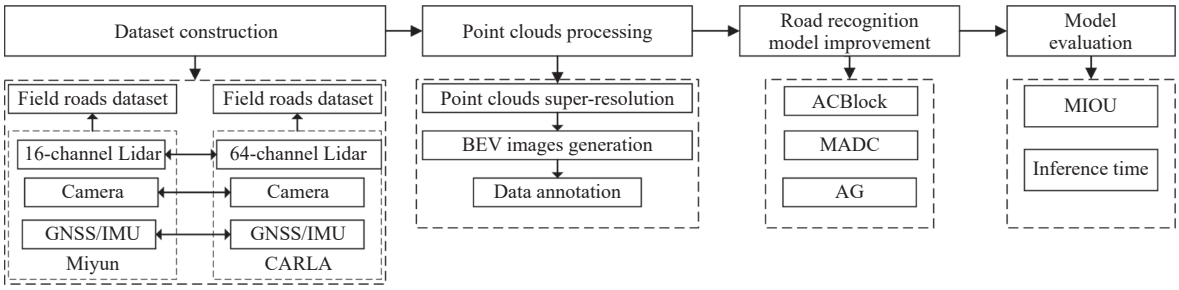
Other existing Lidar-based road recognition methods[16,17] usually require height differences or regular features at the boundaries when extracting road features. However, they are limited in their ability to extract field roads with slight height changes around the road boundary, and they typically require the use of high-channel Lidar or 16-channel Lidar for a certain type of road recognition (straight type). To address these limitations and build a low-cost unmanned system for agricultural machinery, we selected a 16-channel Lidar as the road sensing device. The

Received date: 2022-09-29 **Accepted date:** 2023-03-09

Biographies: Lili Yang, PhD, Associate Professor, research interest: unmanned agricultural machinery, Email: llyang@cau.edu.cn; **Yuanbo Li**, MS, research interest: unmanned agricultural machinery, Email: yeqianchen@cau.edu.cn; **Mengshuai Chang**, MS, research interest: unmanned agricultural machinery, Email: cmsfamily@126.com; **Yuanyuan Xu**, MS, research interest: unmanned agricultural machinery, Email: s20203081484@cau.edu.cn; **Bingbing Hu**, MS, research interest: unmanned agricultural machinery, Email: 1353232901@qq.com; **Xinxin Wang**, MS, research interest: big data mining of agricultural machinery, Email: 939994970@qq.com.

***Corresponding author: Caicong Wu**, PhD, Professor, research interest: big data mining of agricultural machinery, and driverless and cooperative operation. College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China. Tel: +86-13810521813, Email: wucc@cau.edu.cn.

proposed method of this study started from an unstructured low-resolution point cloud, which is then super-resolved and generates a Bird's Eye View (BEV) image of the agricultural machinery's surroundings[18,19]. An improved U-Net++ neural network was trained to carry out road detection in the BEV image, which achieves a balance between accuracy and computational cost. The road recognition process is depicted in Figure 1. The sensor parameter settings of the unmanned simulation platform CARLA were used, which correspond to those of the actual Miyun scene, to compile a dataset. This dataset included 64-channel point cloud simulated field road data obtained in CARLA and 16-channel point cloud field road data collected in the actual scene. The dataset was then super-resolved, and BEV images were generated for two resolutions, which were used to train and test the road recognition model.



Note: GNSS: Global Navigation Satellite System; IMU: Inertial Measurement Unit; CARLA is an open-source simulator for autonomous driving research; BEV: Bird's Eye View; ACBlock: Asymmetric Convolution Block; MADC: Multi-branch Asymmetric Dilated Convolutional Block; AG: Attention Gate; MIOU: Mean Intersection over Union.

Figure 1 Road recognition process

2 Materials and methods

2.1 Dataset construction

2.1.1 Miyun field roads dataset construction

In this study, field road data were collected in Henanzhai town, Miyun District, Beijing, China in June 2021 using a data acquisition platform mounted on a John Deere 1204 tractor. The platform consisted of three modules: point cloud data acquisition, image data acquisition, and vehicle position and posture acquisition. Figure 2 shows a schematic of the data acquisition platform.

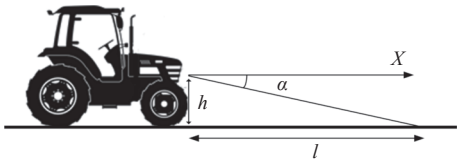
Figure 2 Agricultural scene field roads dataset collection platform

To ensure the safe operation of agricultural machinery, the minimum recognition distance for obstacles should exceed 15 m, taking into account the opposing driving of agricultural machines and a tractor's braking distance of around 7.5 m when the average speed was 25 km/h. The relationship between the Lidar height above the ground and discernible distance is presented in Figure 3 and Equation (1).

$$l = h/tan(\alpha) \quad (1)$$

where, l and h represent the height of the Lidar above the ground and the identifiable distance, respectively, m; α is the angle between the farthest point cloud on the ground and the horizontal, (°).

Based on the minimum recognition distance criterion, the point cloud data acquisition model included a 3D Lidar (Velodyne VLP-16 (VLP-16)) mounted at a height of 1.28 m above the ground, enabling the recognition of objects up to a maximum distance of

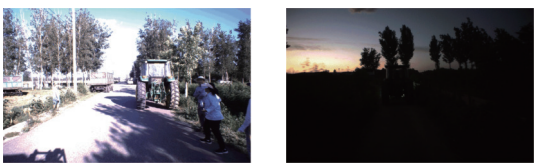


Note: X is the forward direction of the vehicle; h and l represent the height of the Lidar above the ground and the identifiable distance, respectively, m; α is the angle between the farthest point cloud on the ground and the horizontal, (°).

Figure 3 Schematic diagram of Lidar height above ground and identifiable distance

36.50 m when h=1.28 m. The 3D Lidar can rotate 360° in the horizontal direction with 16 channels in the vertical direction to collect point clouds. The image data acquisition module included an industrial camera mounted 1.78 m above the ground, with a resolution of 1920×1200 pixels and an acquisition frequency of 20 Hz. The vehicle position and posture acquisition module comprised a high-precision MEMS integrated navigation receiver (CHCNAV CGI-610) mounted 1.78 m above the ground. The receiver provided real-time high-precision carrier position, attitude, speed, and sensor information. Its antennas were tightly mounted on the roof of the platform to ensure data reliability.

During data collection, the driving speed of the agricultural machinery was around 10 km/h. Figure 4 illustrates the road data captured under different lighting conditions, including semi-structured and unstructured roads. The Miyun agricultural scene field roads dataset contained a total of one thousand frames.



a. Normal light conditions b. Poor light conditions

Figure 4 Camera images of Miyun field roads under different lighting conditions

The field roads data collection platform employed in this study allowed for the collection of high-quality and comprehensive data

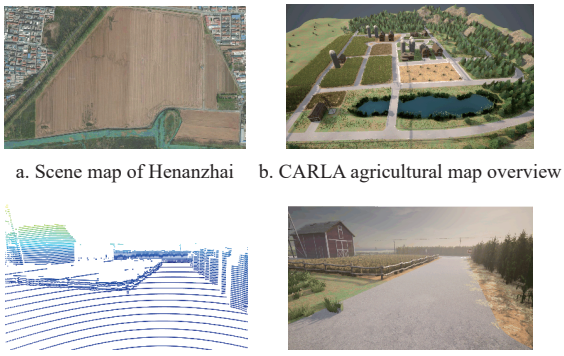
for agricultural machinery research, thereby advancing the development of autonomous agricultural machinery.

2.1.2 CARLA unmanned driving simulation dataset construction

Various open-source software packages, such as CARLA, Autoware, Gazebo, and Unity, are available for simulating different types of Lidar systems on ground vehicles. For this study, the CARLA simulator was utilized since the point cloud super-resolution model required 64-channel point clouds for training, and CARLA simulation maps contain an agricultural production landscape that closely resembles the real-world Henanzhai agricultural scene, including farmland, field roads, pedestrians, and vehicles.

The simulated data acquisition platform consisted of an RGB camera, a "VLP-64" Lidar, and an integrated navigation system. The horizontal and vertical views of the "VLP-64" Lidar with 64 channels in the vertical direction were identical to those of the VLP-16. The parameter settings of other sensors were consistent with those of their real-world counterparts.

To construct datasets, a vehicle was manually driven on the CARLA agricultural production simulation map and gathered 2000 frames of data, which was recorded in the KITTI dataset format. Figures 5a and 5b show the scene map of Henanzhai and the entire CARLA agricultural production landscape, respectively. Figures 5c and 5d depict the point cloud schematic and RGB camera image in the same frame.



a. Scene map of Henanzhai b. CARLA agricultural map overview

c. CARLA field roads point clouds d. CARLA field roads image

Figure 5 Henanzhai and CARLA agricultural scenes

2.2 Point cloud super-resolution

In this study, an point cloud super-resolution model was employed that was trained using the CARLA simulated field roads dataset. The model treats the point cloud super-resolution problem as an image super-resolution problem, as shown in Figure 6. First, the 16-channel point cloud is projected onto a low-resolution range image with a resolution of 16-by-1024, which can be processed by an image neural network. The encoder consists of a sequence of convolutional blocks and average pooling layers, while the decoder has a reversed structure with transposed convolutions for upsampling the feature spatial resolutions. The output layer produces the final high-resolution range image, which has a resolution of 64-by-1024. The high-resolution range image in 2D space is then converted to a 64-channel high-resolution point cloud in 3D space.

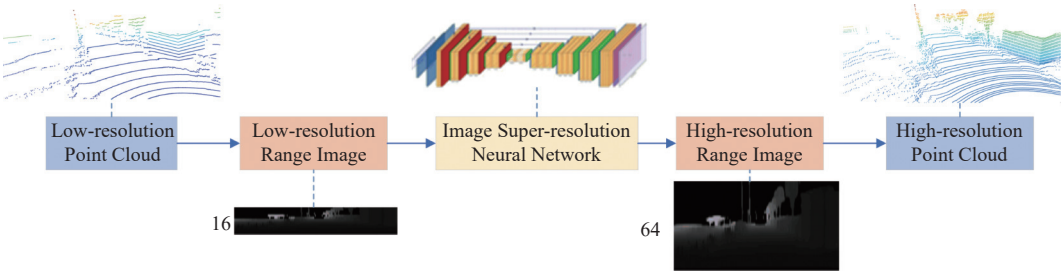


Figure 6 Point cloud super-resolution process

For the experiments of this study, the trained model was used to generate one thousand frames of high-resolution 64-channel point clouds from one thousand frames of sparse 16-channel point clouds from the real Miyun scene.

2.3 Point cloud BEV images generation

Before feeding the unstructured, unordered point clouds to the U-Net++ network, they need to be converted into an appropriate image form[11]. In this work, a region of interest (ROI) was utilized 30 m wide ($y \in [-15\text{ m}, +15\text{ m}]$) and 40 m long ($x \in [0\text{ m}, +40\text{ m}]$), as shown in Figure 7, to account for the existence of many forks in the field roads and the recognition range of Lidar. Firstly, a grid is created in the x-y plane of the Lidar's ROI, and each point cloud element was assigned to one of its cells. Due to the sparsity of point clouds and the necessary image resolution for deep learning, the cell size was set at 0.10 m×0.10 m. Next, for each grid cell, the number of points, maximum height (the difference between the maximum height and the minimum height), and mean reflectivity were calculated. These three attributes were used as the R-channel value, G-channel value, and B-channel value, respectively, in each cell during the conversion of the 3D Lidar data into a bird's eye view (BEV) image. All cells were then aggregated to generate a BEV image of the point cloud with a resolution of 300×400 pixels. Given that the U-Net++ network required the input image to be a multiple of 32 pixels, the BEV image's resolution was adjusted to 512×512 pixels. The resulting point cloud BEV images include a total of 2000 frames, with 1000 frames each of 16-channel point cloud BEV images from Miyun and 64-channel point cloud BEV images generated by the proposed point cloud super-resolution model. It should be noted that although the BEV transformation of the 3D point cloud results in a loss of spatial information, the proposed point cloud super-resolution model indirectly compensates for this loss by upsampling the 16-channel Lidar point cloud to 64-channel.

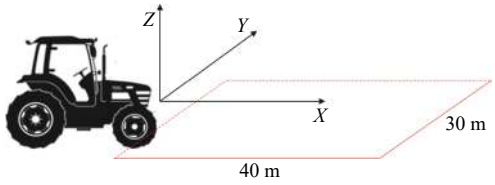


Figure 7 Schematic diagram of grid range

2.4 Data annotation

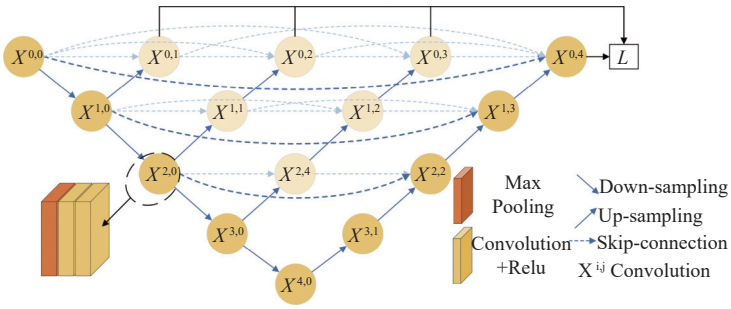
The aforementioned 2000 frames of point cloud BEV images were split into a training set, validation set, and test set in an 8:1:1 ratio. The BEV images were manually labeled and their pixels into two categories were classified, namely road and non-road, using the

Labelme image labeling tool. Next, the labeled training and validation sets were augmented using the Albumentations tool, which included horizontal and vertical flips, as well as 45° counterclockwise rotation.

2.5 Improved U-Net++ model

2.5.1 Model overview

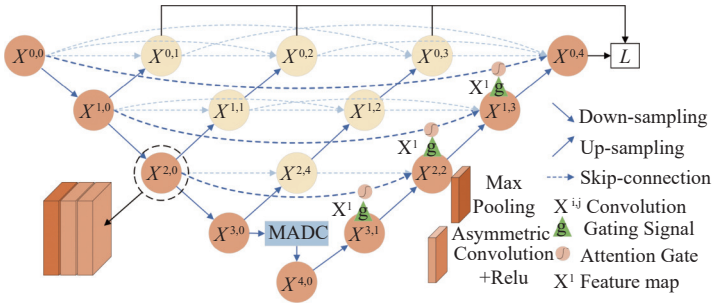
In 2015, Fully Convolutional Network (FCN) was proposed to apply deep learning to road recognition for the first time, enabling road area extraction in urban road scenarios. The U-Net network is a variant of FCN and has been widely used. The U-Net++[20] semantic segmentation network can be viewed as composed of multiple U-Net with different depths. The long-skip-connected features of U-Net are passed without additional operations, which lack the fusion of dense feature graphs. In order to solve this problem, U-Net++ defines short-skip-connection to densely connect the encoder stage and decoder stage feature maps, which were combined into dense short-skip-connection to further fuse the feature map information. However, even though the short-skip-connection of U-Net++ indirectly fused the characteristics of different receptive fields, it only fused the information of the next layer, and the information of the upper layer was not fused, causing the fine granularity of the encoder stage and the decoder stage to still not be fine enough. Additionally, the features of the short-skip connections undergo too much intermediary convolution during the transfer process, leading to the extracted information being combined with irrelevant information, thereby making the decoding path of the image longer and more challenging to train in backpropagation. To address these issues, U-Net++ combines long and short skip connections for feature fusion. Long-skip connections solve the problem of difficult training during backpropagation, allowing original features to be better trained. Short-skip connections can be utilized for dense feature map fusion, which improves feature extraction capability. The network structure is shown in Figure 8. However, U-Net++'s combination of long and short skip connections does not address the issue of the original features containing a lot of redundant information, nor does it resolve the issue of the encoder and decoder stages lacking sufficient granularity.



Note: Relu: Rectified linear unit.

Figure 8 U-Net++ network structure

In this study, several modifications were made to the U-Net++ model in order to address the issues of insufficient granularity in the encoder and decoder stages, as well as the loss of edge information in the road segmentation results due to the lack of direct fusion of information in the upper layers. To address these issues, the original ordinary convolution kernels were replaced in the outermost U-shaped structure with two Asymmetric Convolution Blocks (ACBlock)[21] of size 3. Additionally, a Multi-branch Asymmetric Dilated Convolutional Block (MADC) was added in the highest semantic information layer to fuse information from different receptive fields and directly fuse information from the upper layers. To address the problem of redundant information in the original features resulting from the presence of long-skip-connection in the original network model, an Attention Gate model (AG)[22] was added to the long-skip-connection in the decoder stage, which reduces the extraction of non-road redundant information from the original features via the attention mechanism. The improved U-Net++ structure, incorporating these modifications, is shown in Figure 9.

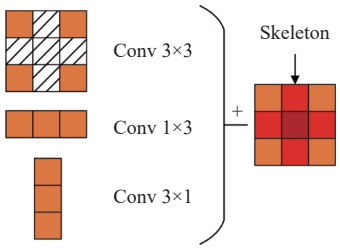


Note: Relu: Rectified linear unit.

Figure 9 Improved U-Net++ structure

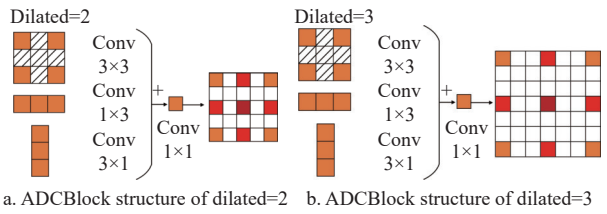
2.5.2 ACBlock

The ACBlock was designed to have higher weights on the "skeleton" structure near the center point, which creates an asymmetric structure with "high skeleton weights and low boundary weights". This uneven weight distribution during the convolution process improves the effective feature extraction ability of the road, especially in curved road recognition scenarios[21], as shown in Figure 10. In the improved U-Net++ model, the original convolutional kernel was replaced in the outermost U-shaped structure with ACBlock, which improved the model's ability to adapt to curved roads. Because the ACBlock was the same size as the original convolutional kernel, the inference time of the



Note: Conv: Convolution; '+' means the outputs are summed up.

Figure 10 Schematic diagram of the ACBlock structure



a. ADCBlock structure of dilated=2 b. ADCBlock structure of dilated=3

Note: ADCBlock: Asymmetric Dilated Convolution. '+' means the outputs are summed up.

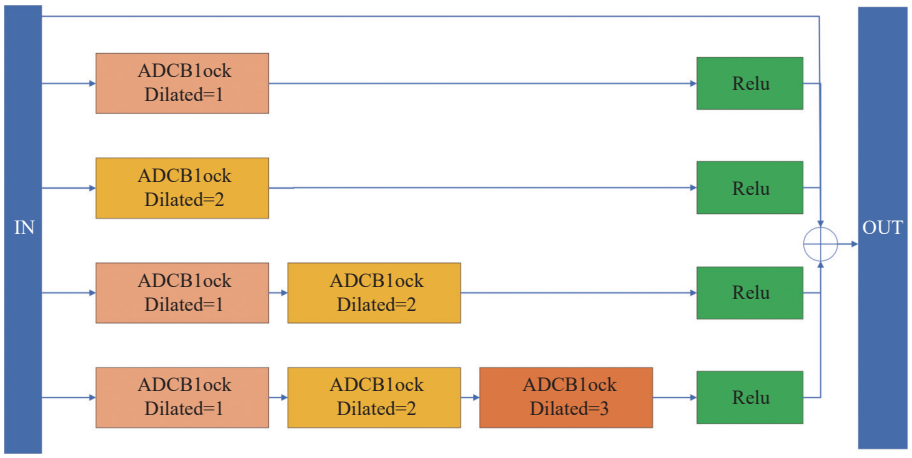
Figure 11 ADCBlock structure with different expansion rates

improved U-Net++ model did not increase.

2.5.3 MADC

Building upon the concepts of dilated convolution[23] and asymmetric convolution, the MADC module introduces Asymmetric Dilated Convolution (ADCBlock), which expands the receptive field without reducing the image's resolution or introducing new hyperparameters. On the basis of ACBlock, it is expanded with an expansion rate of 2 or 3 and smoothed with 1×1 convolution to obtain two kinds of ADCBlock, as shown in Figure 11.

To extract road information from various receptive fields, the structural concept of Inception was integrated[24] with the multi-branch ACBlock combination illustrated in Figure 12 to construct the MADC module. The MADC module consists of five branches that allow for the merging of different receptive fields. The first branch uses the upper layer feature map as input, and the other four branches use ADCBlocks with different expansion rates. The second branch uses an ADCBlock with an expansion rate of 1, the third branch uses an ADCBlock with an expansion rate of 2, the fourth branch uses an ADCBlock with an expansion rate of 1 followed by an ADCBlock with an expansion rate of 2, and the fifth branch uses an ADCBlock with an expansion rate of 1, 2, and 3 in succession. The MADC module applies the Relu function to preserve nonlinearity after each ADCBlock convolution. Compared to the ADCBlock structure, the MADC module increases the network depth and widens the receptive field. The broader perceptual field is better suited for capturing information about large areas of roads and producing more abstract features, while the higher network depth can handle inputs for more complicated road features.



Note: IN: Input; OUT: Output.

Figure 12 Schematic diagram of the MADC structure

2.5.4 Attention gate

Attention Mechanism is a kind of well-performing approach in computer image segmentation and target detection tasks that focuses on target information in the image and ignores irrelevant information. For road recognition, the attention mechanism was introduced to focus on the road features to be learned and ignore non-road regions in the image.

The attention gate model, as shown in Figure 13, involves the encoder's downsampled features X^1 and the decoder's upsampled features g. Firstly, feature maps X^1 and g are subjected to 1×1×1 convolution operation to obtain feature maps A and B, respectively, and feature maps A and B are summed to obtain feature map C, and the Relu function is performed to keep nonlinearity. Then 1×1×1 convolution, sigmoid activation function, and resampling are performed to obtain the attention score α. Finally, α and X^1 are multiplied to assign attention weights to the original features, and the feature maps are fused in the upsampling. This feature map, after multiplying with the attention score, reduces the values of regions in the image that are not related to road features and increases the values of road regions relatively, improving the road segmentation accuracy.

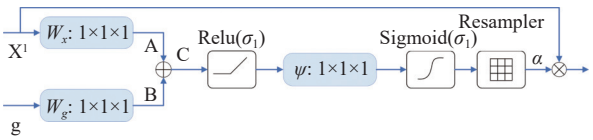


Figure 13 Attention Gate structure

3 Results and analysis

3.1 Network Training

The network of this work was implemented in Pytorch and trained on a single NVIDIA Tesla V100 GPU with 32 GB of memory. The Adam optimization algorithm was used with an initial learning rate of 1e-4, and the learning rate was dynamically adjusted

using exponential decay. To avoid overfitting, the early stop method was employed during training. Specifically, if the loss function value of the validation set did not decrease in 5 consecutive rounds of training, the training was stopped immediately. A hybrid loss function was used, which combines binary cross-entropy with dice coefficients, as the objective function. The hybrid loss L is defined as,

$$L(Y,P) = -\frac{1}{n}\sum_{i=1}^{n}\left(Y_i \cdot \log P_i + \frac{2 \cdot Y_i \cdot P_i}{Y_i^2 + P_i^2}\right) \quad (2)$$

$$L = \sum_{c=1}^{d} L_c(Y,P) \quad (3)$$

where, $Y_i \in Y$ and $P_i \in P$ denote the target labels and predicted probabilities for the class road at the ith pixel in the batch, and n indicates the number of pixels within one batch. The overall loss function for improved U-Net++ is then defined as the weighted summation of the hybrid loss from each individual decoder, where d indexes the decoder.

3.2 Improved model evaluation

To better evaluate the effectiveness of each improved structure, the network was progressively enhanced to construct six different groups of networks based on the U-Net++ architecture. The experimental results are listed in Table 1. Experiments 1-5 used 16-channel point cloud BEV images, while experiment 6 utilized 64-channel point cloud BEV images completed by super-resolution. Experiment 1 represents the original U-Net++ network. Experiment 2 replaced convolution kernels with ACBlock on the basis of Experiment 1. Experiments 3 and 4 added the MADC module and the AG module, respectively, based on Experiment 2. Finally, Experiments 5 and 6 integrated the MADC module into Experiment 4. The performance improvements of different enhancement structures were compared and the point cloud super-resolution method was proposed in this paper by testing the different network structures on their corresponding test sets. The evaluation metrics were Mean Intersection over Union (MIOU) and average inference time (TIME) per image on the test set.

Table 1 Comparison of effects of different improved network structures

Test No	Dataset		Coding structure		MADC	Decoding structure	MIOU/%	Inference time/ms
	16-channel point clouds	64-channel point clouds	CONV	ACBlock		AG		
1	√	--	√	--	--	--	89.19	15.45
2	√	--	--	√	--	--	91.01	15.45
3	√	--	--	√	√	--	92.46	16.04
4	√	--	--	√	--	√	93.85	16.23
5	√	--	--	√	√	√	94.27	16.55
6	--	√	--	√	√	√	96.54	16.55

Note: CONV: Convolution; ACBlock: Asymmetric Convolution Blocks; MADC: Asymmetric Dilated Convolutional Block; AG: Attention Gate; MIOU: Mean Intersection over Union; '√' means the partial content or structure is used; '--' means the partial content or structure is not used.

As demonstrated by the experimental findings listed in Table 1, the MIOU for segmentation of 16-channel point cloud BEV images using the original U-Net++ network was only 89.19%. The experimental results reveal that replacing the standard convolutional kernel with ACBlock improved the convolutional kernel's capacity to extract road characteristics, as evidenced by the 1.82 percentage points increase in MIOU in Experiments 1 and 2, without increasing the number of model parameters and inference time. Experiment 3's addition of the MADC module expanded and fused road information from different respective fields, resulting in a 1.45 percentage points increase in MIOU compared to Experiment 2. Experiment 4 introduced the AG module to focus on the target road features, reducing erroneous extraction of non-road regions and increasing MIOU by 2.84 percentage points compared to Experiment 2. The improved U-Net++ network in Experiment 5 integrated all the improved modules, resulting in a 5.08 percentage points increase in MIOU compared to Experiment 1, to 94.27%. Experiment 6 utilized the improved U-Net++ network to segment 64-channel point cloud BEV images, improving the MIOU by 7.35 percentage points compared to Experiment 1, to 96.54%. Due to the inclusion of MADC and AG modules into the network structure, the inference time slightly increased as the model precision improved. Experiments 1, 2, 3, 5, and 6 were selected as comparison experiments, used the point cloud BEV images of different road sections as the input image, and employed the image obtained by Labelme software annotation as the true values. Figure 14 shows a comparison of the prediction results of the network.

In more detail, the first row shows a semi-structured road without obstacles, and the improved U-Net++ network can perform better road recognition. In the second row, a semi-structured straight road with distant agricultural machinery and pedestrians is shown. Although the improved U-Net++ network using 16-channel point cloud BEV images as input can segment the road near the obstacles, there may be some mis-extraction due to the limited Lidar beams hitting the distant objects and fewer extractable features. In contrast, the improved U-Net++ network using 64-channel point cloud BEV images as input can segment the roads near distant obstacles more accurately, thanks to the increased number of Lidar beams and richer features. In the third row, a semi-structured intersection is shown. The improved U-Net++ network using 64-channel point cloud BEV images as input mistakenly extracts the water channel next to the distant car as a road area, because the point cloud data at the canal was mistakenly expanded during super-resolution, making the features here similar to the road surface. The fourth row shows an unstructured road with farmlands beside it, which have more similar characteristics to the road surface. Our method can eliminate the mis-extraction of farmland beside the road, thereby improving the accuracy of road segmentation. Overall, the proposed method in this paper can segment point cloud BEV images of field roads more accurately than the original U-Net++ network in most cases, thereby ensuring the safety of agricultural machinery.

3.3 Real-time Assessment

Table 2 lists the time required for each stage of our method. As shown, the proposed method of this study achieves an average processing time of 69.56 ms/frame, which is well within the range of real-time performance. Considering that the typical speed of tractors on field roads is 20 km/h, the proposed method can accurately identify field roads in real time.

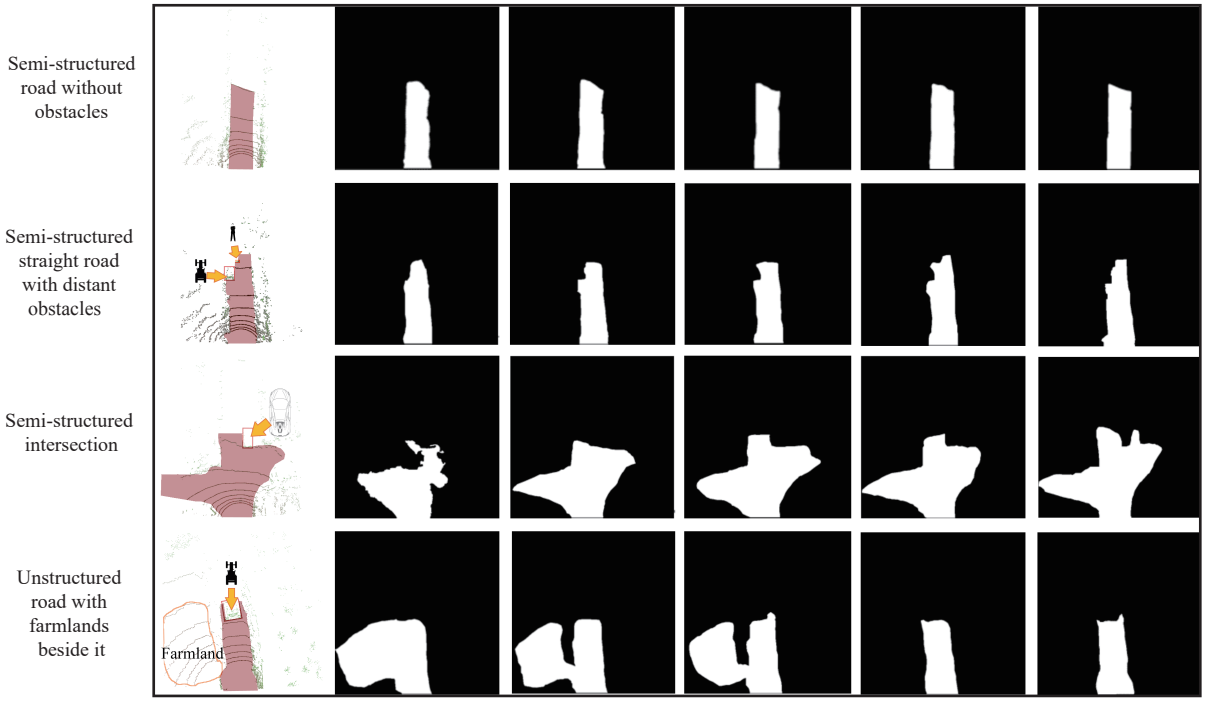


Figure 14 Comparison of network prediction effects under different scenarios

Table 2 Average time required to process a frame in each stage

Data process	16-channel point clouds	64-channel point clouds
Super resolution/ms	×	8
BEV image/ms	42.17	45.01
Road prediction/ms	16.55	16.55
Total/ms	58.72	69.56

Note: '×' means that it has not gone through this operation.

4 Conclusions

In this study, a technical solution was proposed for the recognition of field roads in BEV images using 16-channel Lidar. By performing super-resolution processing on the 16-channel point cloud data to generate 64-channel point cloud data, converting 3D point clouds into the 2D point cloud BEV images, and improving the U-Net++ road segmentation network, the accuracy of road segmentation had been significantly improved. The proposed method of this study employed the U-Net++ as the basic semantic segmentation network structure, and incorporated three key improvements:

1) The ACBlock was used to replace the original U-Net++ convolutional kernel in the outermost U-shaped structure. The asymmetric structure of the ACBlock enhanced the feature extraction capabilities of convolutional kernels without introducing additional hyperparameters, resulting in superior generalization performance;

2) The MADC was added to the highest semantic information layer. By combining asymmetric dilated convolutions with various receptive fields, the MADC improved the robustness of the model;

3) The AG model was added to the long-skip-connection in the decoding stage, which effectively filters out invalid non-road information in the image and mitigates the false extraction of farmland with a similar structure to unstructured roads;

The proposed method of this study achieved an MIOU of 96.54% and an average processing time of 69.56 ms on 16-channel point clouds, which was 7.35 percentage points higher than U-Net++, meeting the requirements for unmanned driving in agricultural machinery. Moreover, this study's approach was also applicable to lower channel Lidars, such as 8-channel Lidar, by super-resolving 8-channel Lidar point clouds by a factor of eight. Multi-sensor fusion is the main research direction of future unmanned driving, and future work will be extended to multi-sensor fusion for environment perception.

Acknowledgements

The authors acknowledge that this work was financially supported by the National Key R&D Program of China and Shandong Province, China (Grant No. 2021YFB3901300).

[References]

[1] Cui X Z, Feng Q, Wang S Z, Zhang J H. Monocular depth estimation with self-supervised learning for vineyard unmanned agricultural vehicle. Sensors, 2022; 22(3): 721.

[2] Zhu N Y, Liu X, Liu Z Q, Hu K, Wang Y K, Tan J L, et al. Deep learning for smart agriculture: Concepts, tools, applications, and opportunities. Int J Agric & Biol Eng, 2018; 11(4): 32–44.

[3] He Y, Jiang H, Fang H, Wang Y, Liu Y F. Research progress of intelligent obstacle detection methods of vehicles and their application on agriculture. Transactions of the CSAE, 2018; 34(9): 21–32. (in Chinese)

[4] Yao L J, Hu D, Yang Z D, Li H B, Qian M B. Depth recovery for unstructured farmland road image using an improved SIFT algorithm. Int J Agric & Biol Eng, 2019; 12(4): 141–147.

[5] Pang S, Morris D, Radha H. CLOCs: Camera-LiDAR object candidates fusion for 3D object detection. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Las Vegas: IEEE, 2020; pp. 10386–10393. doi: 10.1109/IROS45743.2020.9341791.

[6] Cui Y P, Xu H, Wu J Q, Sun Y, Zhao J X. Automatic vehicle tracking with roadside LiDAR data for the connected-vehicles system. IEEE Intelligent Systems, 2019; 34(3): 44–51.

[7] Lyu Y C, Bai L, Huang X M. Real-time road segmentation using lidar data processing on an FPGA. In: 2018 IEEE International Symposium on Circuits and Systems (ISCAS), IEEE, 2018; pp.1–5.

[8] Kisner H, Thomas U. Segmentation of 3D point clouds using a new spectral clustering algorithm without a-priori knowledge. In: VISIGRAPP 2018, 2018; pp. 315–322.

[9] Zhang W. Lidar-based road and road-edge detection. In: 2010 IEEE Intelligent Vehicles Symposium. La Jolla: IEEE, 2010; pp.845–848. doi: 10.1109/IVS.2010.5548134.

[10] Charles R Q, Su H, Mo K C, Guibas L J. Pointnet: Deep learning on point sets for 3D classification and segmentation. In:Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2017; pp.652–660.

[11] Beltrán J, Guindel C, Moreno F M, Cruzado D, Garcia F, De La Escalera A. Birdnet: A 3D object detection framework from lidar information. In: 2018 21st International Conference on Intelligent Transportation Systems (ITSC), IEEE, 2018; pp. 3517–3523.

[12] Hua B S, Tran M K, Yeung S K. Pointwise convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2018; pp. 984–993.

[13] Zhang Y H, Wang J, Wang X N, Dolan J M. Road-segmentation-based curb detection method for self-driving via a 3D-LiDAR sensor. IEEE Transactions on Intelligent Transportation Systems, 2018; 19(12): 3981–3991.

[14] Zhou Y, Tuzel O. Voxelnet: End-to-end learning for point cloud based 3D object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, IEEE, 2018; pp. 4490–4499.

[15] Chang Y C, Xue F, Sheng F, Liang W T, Ming A L. Fast road segmentation via uncertainty-aware symmetric network. In: 2022 International Conference on Robotics and Automation (ICRA), Philadelphia: IEEE, 2022; pp. 1124–11130. doi: 1109/ICRA46639.2022.9812452.

[16] Zhu Z, Liu J L. Graph-based ground segmentation of 3D LIDAR in rough area. In: 2014 IEEE International Conference on Technologies for Practical Robot Applications (TePRA), IEEE, 2014; pp. 1–5. doi: 10.1109/TePRA.2014.6869157.

[17] Cheng Z Y, Ren G Q, Zhang Y. Ground segmentation from 3D point cloud using features of scanning line segments. Opto-Electronic Engineering, 2019; 46(7): 180268.

[18] Triess L T, Peter D, Rist C B, Enzweiler M, Zollner J M. CNN-based synthesis of realistic high-resolution LiDAR data. In: 2019 IEEE Intelligent Vehicles Symposium (IV), Paris: IEEE, 2019; pp. 1512–1519.

[19] Shan T X, Wang J K, Chen F F, Szenher P, Englot B Simulation-based lidar super-resolution for ground vehicles. Robotics and Autonomous Systems, 2020 134: 103647. doi: 10.1016/J.ROBOT.2020.103647.

[20] Zhou Z, Siddiquee M M R, Tajbakhsh N, Liang J M. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. IEEE Transactions on Medical Imaging, 2019; 39(6): 1856–1867.

[21] Ding G G, Han J G, Ding X H, Guo Y C. ACNet: Strengthening the kernel skeletons for powerful CNN via asymmetric convolution blocks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, IEEE, 2019; pp. 1911–1920.

[22] Oktay O, Schlemper J, Folgoc L L, Lee M, Heinrich M, Misawa K, et al. Attention U-Net: Learning where to look for the pancreas. arXiv preprint, 2018; arXiv: 1804.03999, 2018.

[23] Yang J D, Zhu J T, Wang H L, Yang X. Dilated MultiResUNet: Dilated multiresidual blocks network based on U-Net for biomedical image segmentation. Biomedical Signal Processing and Control, 2021; 68: 102643.

[24] Bala S A, Kant S. Dense dilated inception network for medical image segmentation. International Journal of Advanced Computer Science and Applications (IJACSA), 2020; 11(11): 0111195.