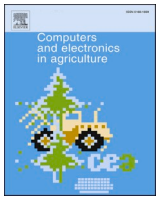
Computers and Electronics in Agriculture 212 (2023) 108073

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag



Case instance segmentation of small farmland based on Mask R-CNN of feature pyramid network with double attention mechanism in high resolution satellite images

Yangyang Cao [a], Zuoxi Zhao [a,b,*], Yuan Huang [a], Xu Lin [a], Shuyuan Luo [a], Borui Xiang [a], Houcheng Yang [a]

[a] College of Engineering, South China Agricultural University, Guangzhou 510642, China
[b] Key Laboratory of Key Technology on Agricultural Machine and Equipment, South China Agricultural University, Ministry of Education, Guangzhou 510642, China

ARTICLE INFO

Keywords:
Farmland segmentation
Mask R-CNN
Small farms
Attention mechanism
VHR images

ABSTRACT

Accurate spatial information of farmland in small farms is very important to provide operable information to farmers, managers and decision makers. However, small farms have small area, irregular shape, and use a variety of planting crops, which makes their boundaries blurred, and the standard edge detection algorithm cannot accurately segment the farmland boundary. Therefore, the automatic delimitation of fields in small farms is a challenging task. Aiming at the above problems, this paper proposes an example segmentation method of Mask R-CNN based on dual attention mechanism feature pyramid network (DAFPN) to describe small farms. DAFPN is composed of two attention modules: spatial attention module (SPA) and channel attention module (CHA) to enhance its feature extraction ability. Spatial attention module (SPA) generates spatial attention map by using the spatial relationship of features, and generates information to be emphasized or suppressed in spatial location; The channel attention module (CHA) learns an adaptive channel merging method based on the attention mechanism. Our proposed DAFPN can be easily inserted into the existing FPN model. We have conducted extensive experimental analysis on very high resolution (VHR) satellite images based on the Mask R-CNN deep learning framework of DAFPN. The standard COCO dataset evaluation index and F1-score evaluation strategy are used to compare the algorithm. AP_{50}, AP_{75} and F1-score reach 82.86%, 55.51% and 70.90% respectively, which is 8.7%, 8.31% and 6.87% higher than Mask R-CNN respectively. Our results highlight the ability of Mask R-CNN based on DAFPN to accurately depict small farms in VHR satellite images, which lays a foundation for the automatic segmentation of small farms.

1. Introduction

Accurately depicting and detecting the spatial distribution of agricultural resources is crucial to increasing agricultural production and ensuring food security in many parts of the world (Debats et al., 2016a). By 2050, the world's population is expected to increase to 10billion, and food production needs to be increased by 98%, in which small farms will play a key role. According to statistics, 80% of the food produced in Asia and sub-Saharan Africa comes from small farms, and about 90% of the farmers in the world are small farmers, with less than two hectares of land (Kienzle, 2013). However, the progress of various technologies (especially digital technology) and their decreasing costs are providing accurate farmland information for small farms in developing countries. In addition to activities such as pesticides and fertilizers, accurate farmland information can also be associated with detailed farmland information, such as boundaries, soil types and humidity, pest invasion, crops planted and yield applications achieved. By collecting and analyzing field data frequently, farmers can make more informed decisions (Marvaniya et al., 2021). In addition, field plot mapping indirectly provides information about agricultural practice, mechanization and production efficiency (Debats et al., 2016a; Shawon et al., 2020). An important prerequisite for precision agriculture detection is to understand the planting range, and the information of field space can also be introduced into crop type mapping and land use classification (Turker and Kok, 2013).

In this paper, a field is an area of land used for agricultural purposes

* Corresponding author..
E-mail address: zhao_zuoxi@scau.edu.cn (Z. Zhao).

https://doi.org/10.1016/j.compag.2023.108073
Received 4 August 2022; Received in revised form 9 June 2023; Accepted 12 July 2023
Available online 26 July 2023

to grow specific crops or crop mixtures (Nations, 2007). Field boundaries are defined as boundaries where crop types, crop mixes, or farm management practices change, or adjacent fields are separated by structures such as roads, fences, or very thin uncultivated areas. Our goal is to segment small farmland from VHR satellite images.

Contributions: This paper introduces an instance segmentation method of Mask R-CNN deep learning framework based on double attention feature pyramid network (DAFPN), which is used to segment fields of small farms from VHR satellite images. Mask R-CNN is trained to detect segmented fields. In the double attention feature pyramid network, spatial attention module (SPA) can be regarded as an adaptive spatial region selection mechanism. Only task related regions are the most important ones to be concerned about, and irrelevant regions in the image are discarded. The channel attention module (CHA) models the importance of each channel, and then enhances or suppresses different channels for different tasks. Our main contributions are:

(1) An instance segmentation method based on mask R-CNN deep learning framework based on double attention feature pyramid network (DAFPN) is introduced to segment the farmland of small farms from VHR satellite images;
(2) We designed two attention modules: spatial attention module and channel attention module, and proposed a new feature pyramid network structure for farmland segmentation;
(3) Our proposed double attention feature pyramid network (DAFPN) model can be easily inserted into the existing FPN structure.

2. Related work

Delimiting boundaries or detecting a single region is a special example of image segmentation. The developed methods can be roughly divided into non-learning methods using the salient features of low-level images (Zhu et al., 2016) and deep-learning methods (Badrinarayanan et al., 2017; Chen et al., 2018; Garcia-Garcia et al., 2017). Non-learning methods can be further divided into edge or boundary-based methods, region-based methods or hybrid methods. As described below, some of these methods have been reused and applied to farmland mining.

Edge based methods, such as Canny detector, focus on identifying discontinuities in the image to select candidate pixels to represent the field boundary (Canny, 1986; Martin et al., 2004), however, it cannot organize image features into consistent fields because they cannot guarantee closed polygons (Nevatia and Babu, 1980). Region based methods, such as multi-resolution segmentation, focus on grouping pixels into objects according to some homogeneity criteria (Kettig and Landgrebe, 1976; Pal and Mitra, 2002), but sometimes it is impossible to locate the boundary at the highest gradient or linear natural or visible edge (Chen et al., 2015). To overcome these limitations, researchers developed hybrid methods, these methods have been proved to improve the accuracy of farmland boundary detection (Crommelinck et al., 2017). Although these methods have shown promise, they usually recognize long boundaries, have straight object shapes, and have high brightness contrast compared with adjacent areas (Evans et al., 2002; Mueller et al., 2004). Other studies combine a set of extracted image features with classifiers (such as neural networks) to detect boundaries (Wagner and Oppelt, 2020).

Research shows that deep learning models have been used in many recent studies on automatic edge and contour detection algorithms. They have shown extraordinary ability in learning advanced data representation for object recognition, image classification and segmentation (Bertasius et al., 2015; Maninis et al., 2017; Yan and Roy, 2016).

LeNet model is the first convolution neural network model proposed, which lays the foundation for the development of deep convolution neural network (Lecun et al., 1998). The first popular modern CNN model is AlexNet proposed in 2012 (Krizhevsky et al., 2012), AlexNet has greatly improved the performance of image classification and other tasks, and won the championship of ImageNet challenge (Deng et al., 2009), making deep convolution neural network gradually become a hot spot in the field of computer vision. In recent years, many more advanced architectures have been introduced. For example, GoogLeNet adopts the Inception-v1 module, which uses sparse connections to reduce the number of model parameters, while ensuring the efficiency of computing resources, and improves the performance of the network when the depth reaches 22 layers (Szegedy et al., 2015). With the emergence of more and more models, CNN has more and more application scenarios, including scene classification (Cheng et al., 2018), land-cover and land-use classification (Bergado et al., 2016; Fu et al., 2017; Maggiori et al., 2016). In particular, deep convolution networks such as U-Net, ResU-Net, and SegNet have been used to map field boundaries (Masoud et al., 2019; Persello et al., 2019; Waldner and Diakogiannis, 2020).

However, the influence of semantic segmentation model on farmland segmentation is limited, and it does not necessarily produce separated fields. Additional post-processing is required to define areas with closed contours and obtain individual instances (Waldner et al., 2021). In contrast, instance segmentation can be used to detect and segment each field separately. Taking the field boundary description as an instance segmentation problem can be an alternative method to directly generate a complete closed field polygon. (Zhang et al., 2018) use Mask R-CNN to complete the characterization process of automatic recognition of Arctic ice wedge polygons. Similarly, (Feng et al., 2019) developed a ship detection model based on Mask R-CNN, in which Mask R-CNN is used as the backbone, and another module called sequence local context module is used to obtain better remote sensing image results. (Quoc et al., 2020) use high-resolution satellite images from Google maps, the accuracy of Mask R-CNN and U-Net in segmenting farmland is compared. It is found that Mask R-CNN has achieved higher average accuracy in various fields in many countries/regions. And (Potlapally et al., 2019) successfully uses Mask R-CNN network to segment farmland instances in remote sensing images, however, it is not clear how effective Mask R-CNN is in mapping single field boundaries in areas with very small field sizes.

This paper presents a new agricultural method, which uses Mask R-CNN based on double attention mechanism feature pyramid network (DAFPN) to segment small farmland from VHR satellite images. This method accurately extracts the field boundary on the new image. Even if the model is trained on the image of a limited region, it will also detect the field boundaries of other regions. We compare the output of the improved Mask R-CNN, the original Mask R-CNN and the Mask R-CNN deep learning model after adding the attention mechanism of the convolutional block attention module (CMAM) (Woo et al., 2018), and analyze and discuss the results.

3. Study area and dataset

3.1. VHR satellite images data

Our research area is mainly located in Guangdong Province, the southernmost tip of the Chinese Mainland and bordering on the South China Sea. Guangdong province belongs to the seasonal climate region in East Asia, and is the region with the most abundant light, heat and water resources in China. The cultivated land in the whole region is “three crops a year”, with an average quality grade of 5.47, which is more than 4 levels higher than the average quality grade of cultivated land in China. Although the quality of cultivated land is not much, the number of cultivated lands in Guangdong, a populous province, is not much. By the end of 2020, the total cultivated land in the province was 1.899 million hectares, and the per capita cultivated land area was only 0.165 acres, less than 1/5 of the Chinese average level. The number of cultivated land spots below 0.2 ha accounted for 80%, but the area accounted for only 7%. Therefore, it is more in line with our research content and is used to divide small-scale farmland. Our VHR satellite image research area is located in Guangzhou and Foshan (Fig. 1). Its

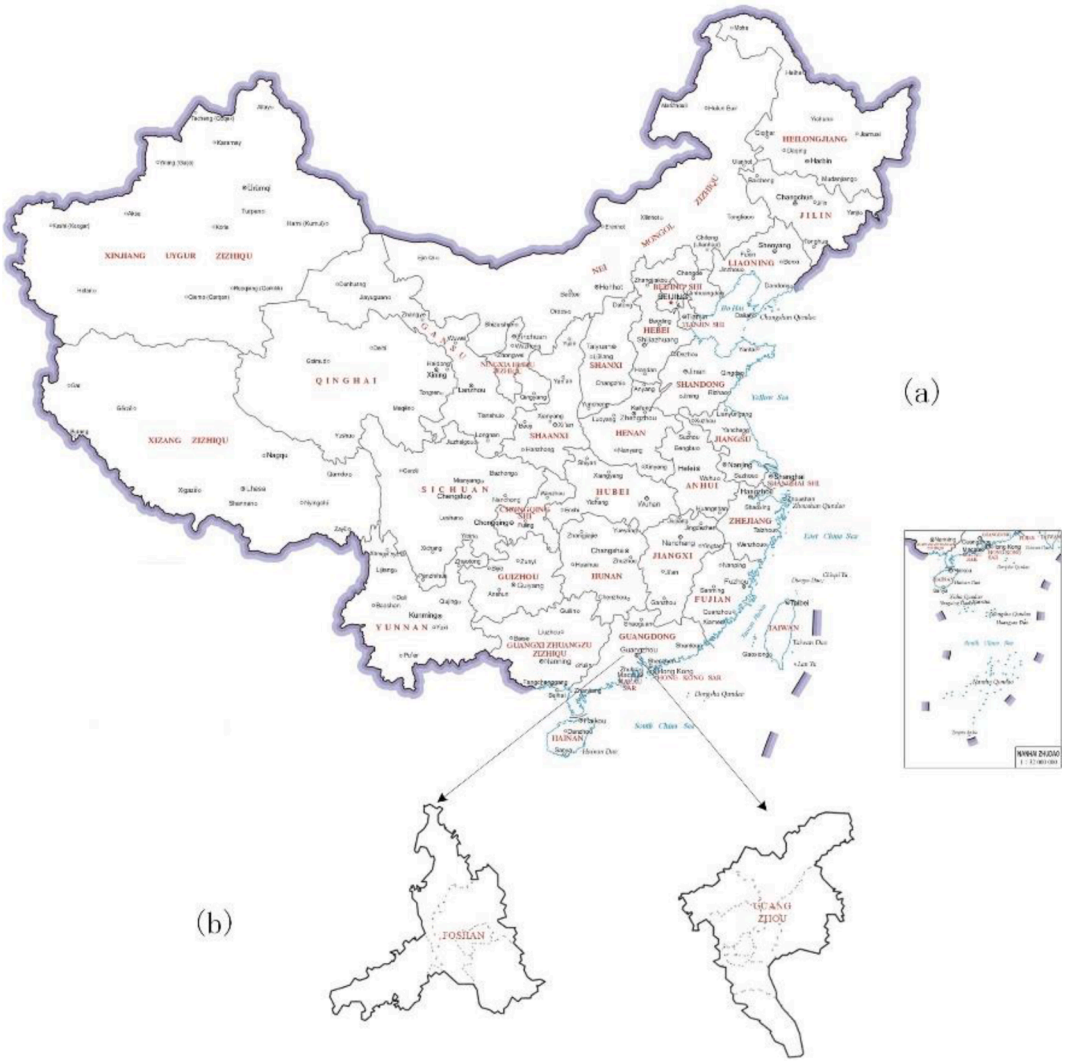


Fig. 1. Dataset distribution area. (a) Shows the location of the study area in China; (b) The geographical location of the specific study area is shown (Guangzhou City 23.13 N, 113.27E; Foshan City 23.02 N, 113.06E).

distribution is mainly affected by terrain and landform, showing a patchy distribution pattern, and all of them are mainly small farmland farms. The satellite data used in this paper is constructed from 0.3 m satellite images of Guangzhou and Foshan in December 2021, with three bands of red, green and blue.

3.2. Dataset preparation

First, we get images from Google Earth (https://www.google.com/intl/zh-CN/earth/), after geometric registration and preprocessing, in view of the high resolution and large amount of calculation of each region, each region is trimmed to 1024 × 1024-pixel small image, no overlap. Considering Tobler's first law, that is, everything is related to everything else, but things near are more relevant than things far away (Tobler, 1970), training, verification and test sets are split so that they are spatially independent, thereby enhancing the differences between split sets.

Secondly, the image dataset uses the open-source script LabelMe on GitHub (https://github.com/wkentaro/labelme) Comment. After running the labelMe script, label the small farmland in each satellite image for object instance segmentation; After that, the training data is saved in the format of JavaScript object notation (.Json), which is also the data format used in some other machine learning training data sets for training on the data sets. The detailed data operation steps are shown in Fig. 2.

4. Method

4.1. Mask R-CNN network

As one of the most advanced instance segmentation models, Mask R-CNN (He et al., 2017) is improved by adding segmentation mask to generate branches on the basis of faster R-CNN (classification + regression branches) (Girshick, 2015). The farmland segmentation framework based on Mask R-CNN is shown in Fig. 3. It consists of four parts: (1) the backbone structure and feature pyramid are responsible for feature extraction of the whole image to generate features of different scales; (2) Regional suggestion network (RPN) is used to generate regions of interest; (3) ROI classifier for category prediction of each ROI and bounding box regressor for refining ROI; (4) FCN with RoIAlign and bilinear interpolation for predicting pixel accurate masks (He et al., 2017). In our research, the feature pyramid network (DAFPN) based on dual attention mechanism and ResNet50 is used as the backbone to improve accuracy and speed (Hariharan et al., 2014; Lin et al., 2017). For each image, the CNN feature is extracted by the backbone, and then the RPN uses the sliding window method to calculate the bounding box

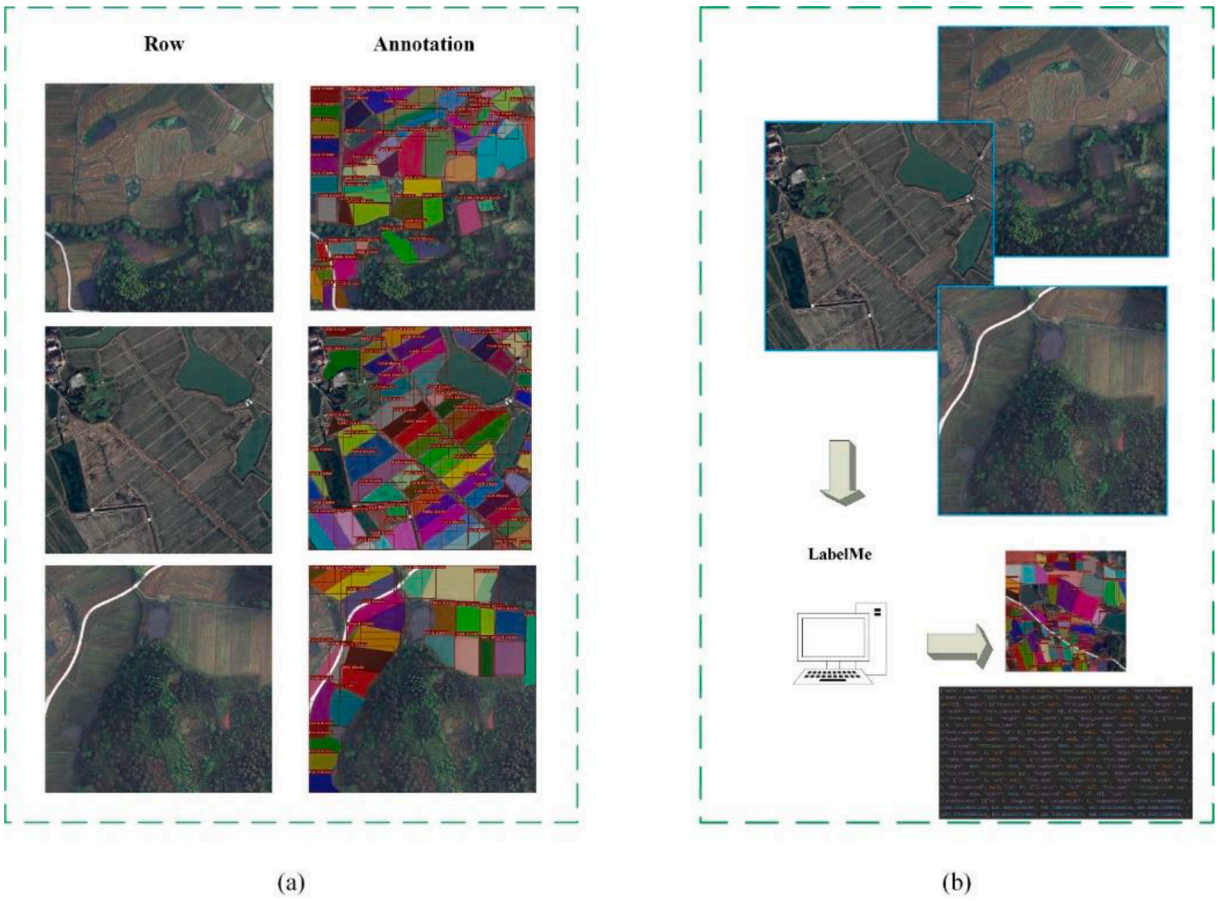


Fig. 2. Some annotation data and data processing. (a) Farmland annotation image. (b) Data processing.

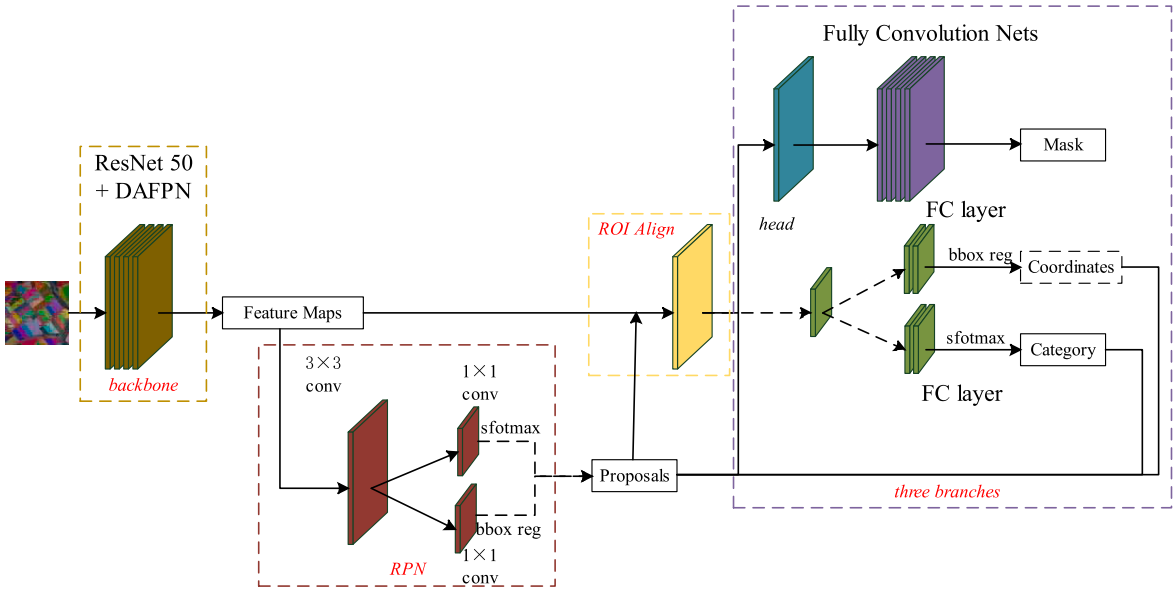


Fig. 3. Mask R-CNN farmland segmentation framework based on DAFPN.

proposal on the feature map (Girshick, 2015). Before the next step, RoIAlign is used to map the spatial region of interest of any size in the feature to a fixed spatial resolution by using bilinear interpolation. Finally, Mask R-CNN header predicts the object category, refines the bounding box location, and generates the segmentation mask at the same time.

4.2. Dual attention feature pyramid network

The overall framework of DAFPN is shown in Fig. 4. We designed two

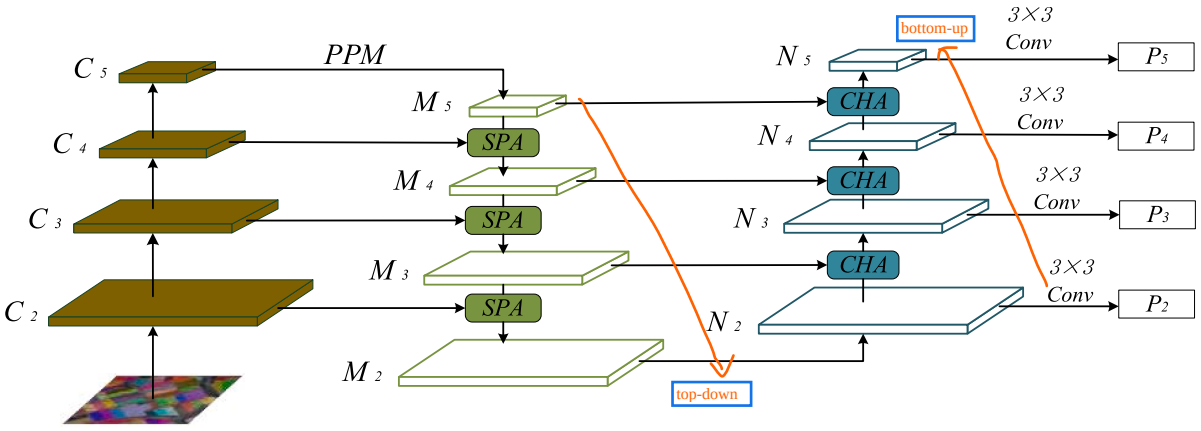


Fig. 4. Overview of double attention mechanism feature pyramid network. SPA: spatial attention module CHA: channel attention module.

attention modules: spatial attention module (SAP) and channel attention module (CHA). We designed SPA and CHA to optimize different problems. In order to reduce the coupling between the two modules in the joint optimization process, we alternately optimize spatial attention and channel fusion. Inspired by PANet (Liu et al., 2018), we extended a bottom-up path after the lowest level of FPN. Then we embed spa into the top-down path and cha into the bottom-up path. Like DRFPN (Ma and Chen, 2020), we also use the pyramid pooling module because it has a powerful ability to capture context information (Zhao et al., 2017), we connect PPM with the highest-level feature extracted by the backbone as the input of the top-down path. Resnet50, as the first part of the backbone network, consists of five stages, corresponding to five characteristic maps with different scales {C2, C3, C4, C_5}, and its corresponding step size is {4, 8, 16, 32} pixels. These feature maps are used to establish the feature pyramid of DAFPN network and obtain new features respectively {P_2, P_3, P_4, P_5}.

4.2.1. Spatial Attention Module

In FPN, it is inaccurate to up sample the feature map with large resolution only based on the position information. Inspired by CBAM and DRFPN (Ma and Chen, 2020; Woo et al., 2018), this paper first combines the context information of adjacent layers, and then in order to calculate spatial attention, we use the operations of average pooling and maximum pooling to act on the channel direction. In order to aggregate spatial information, average pooling has been widely used. For example, it is used to effectively learn the range of target objects and calculate spatial statistics (Hu et al., 2018; Zhou et al., 2016).

As shown in Fig. 5, given two adjacent feature graphs C_{low} and M_{high}, we first use the convolution layer to compress the channel of M_{high} to reduce the computational cost, and then up sample the compressed m to the same size as C_{low}, and then connect them, and perform average pooling and maximum pooling operations on the connected feature graphs respectively, to obtain two 2D graphs representing the average pooling feature and maximum pooling feature in the channel: $F^s_{avg} \in \mathbb{R}^{1\times H\times W}$ and $F^s_{max} \in \mathbb{R}^{1\times H\times W}$, mathematically expressed as:

$$F^s_{avg} = AvgPool(cat(C_{low}, upsample(conv_0(M_{high})))) \quad (1)$$

$$F^s_{\max} = MaxPool(cat(C_{low}, upsample(conv_0(M_{high})))) \quad (2)$$

Where $cat(\bullet)$ represents connection operation, $conv_0(\bullet)$ represents 1×1 convolution layer of channel compression, and *upsample* represents up sampling operation. After getting the two feature maps, connect them and get the spatial attention map through the standard convolution layer and activation function. The spatial attention map generates the position information that needs to be highlighted or suppressed. Finally, multiply the spatial attention map by C_{low} to get the feature map M_i:

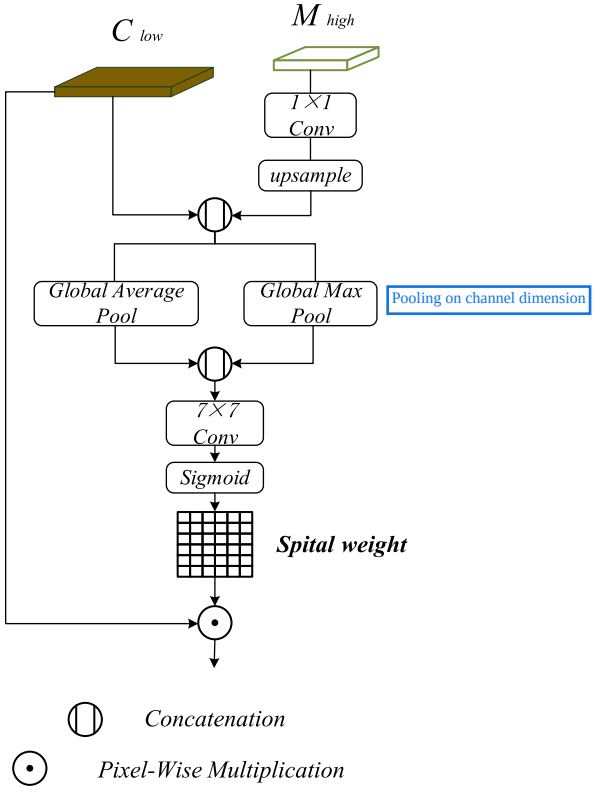


Fig. 5. Spatial Attention Module Network.

$$Attention^s = \sigma(conv_1(cat(F^s_{avg}, F^s_{\max}))) \quad (3)$$

$$M_i = Attention^s \odot C_{low} \quad (4)$$

Where $conv_1(\bullet)$ represents 7×7 convolution layer and σ represents sigmoid function.

4.2.2. Channel Attention Module

Inspired by CBAM (Woo et al., 2018), we propose channel attention module (CHA), which aims to optimize the fusion method between channels according to the context. As shown in Fig. 6, the module works in a bottom-up manner, using the channel attention mechanism to guide the network to learn the weight of channels when merging adjacent layers.

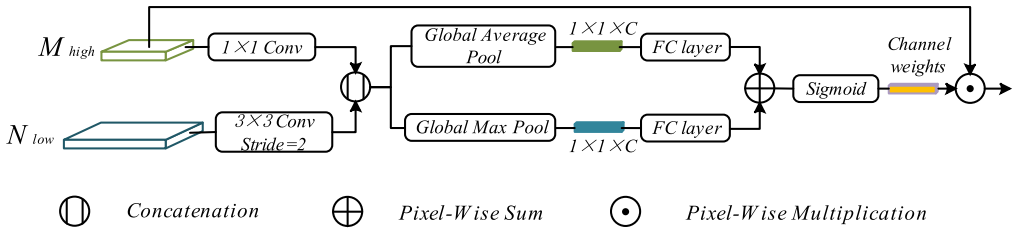


Fig. 6. Channel Attention Module Network.

In our network, given two adjacent characteristic graphs N_{low} and M_{high}, we first use a 1×1 convolution layer to compress the channel of M_{high} to reduce the computational cost, then use a 3×3 convolution layer to down sample N_{low} to the same size as M_{high}, and then connect them, and use global average pooling and global maximum pooling to aggregate two different spatial context information for the connected characteristic graphs respectively. Two different spatial context descriptors are generated: F^c_{avg} and F^c_{max}. next, the two descriptors enter the full connection layer respectively, and the channel weight is obtained by summing the sigmoid function element by element, that is, the channel attention graph. Finally, the channel attention graph is multiplied by M_{high} to obtain the characteristic graph N_i. this process can be expressed as:

$$F^c_{avg} = AvgPool(cat(conv_2(M_{high}), conv_3(N_{low}))) \tag{5}$$

$$F^c_{\max} = AvgPool(cat(conv_2(M_{high}), conv_3(N_{low}))) \tag{6}$$

$$Attention^c = \sigma(fc(F^c_{avg}) + fc(F^c_{\max})) \tag{7}$$

$$N_i = Attention^c M_{high} \tag{8}$$

Where $cat(\bullet)$ represents the connection operation, $conv_2(\bullet)$ represents the 1×1 convolution layer of the compression channel, and $conv_3(\bullet)$ represents 3×3 convolution layer, and the stride is 2. $fc(\bullet)$ represents the full connection layer. In order to reduce parameters, the full connection layer is shared, and σ represents sigmoid function.

Unlike CBAM (Woo et al., 2018), Cha focuses on how adjacent layers guide each other through context when merging, and learn weights by fusing low-level feature maps and high-level feature maps. Cha can more accurately capture the semantic relationship between channels and improve the final detection ability.

4.3. RPN and RoIAlign operation

RPN receives the characteristic image $\{P_2, P_3, P_4, P_5, P_6\}$ (P_6 is obtained by down sampling P_5), performs sliding convolution operation on it, generates candidate regions without categories, distinguishes the candidate regions with or without farmland targets, calculates the central coordinates, length and width of the input farmland data image corresponding to the region of interest containing farmland, and determines the coordinate position of the candidate frame. After RPN network processing and prediction, a large number of overlapping candidate boxes will be generated. It is necessary to use the non-maximum suppression (NMS) algorithm to screen out some accurate candidate boxes with high foreground confidence, and input them into RoIAlign (He et al., 2017) together with the output feature map. RoIAlign adjusts the size of the anchor box to a fixed size (He et al., 2017). The Mask R-CNN model selects the bilinear interpolation method in the RoIAlign layer to calculate the position coordinates, so that the originally discrete pooling process is continuous, and the pixel values with coordinates corresponding to floating-point numbers can be mapped without any quantization of the candidate areas. In the back-propagation of RoIAlign layer, $i^*(r,j)$ is the coordinate position of a floating-point number (the sampling point calculated during forward propagation), so in the characteristic map before pooling, each point with the horizontal and vertical coordinates of $i^*(r,j)$ less than 1 should accept the gradient of the corresponding point y_{rj} back-propagation, so the back-propagation formula of RoIAlign is:

$$\frac{\partial L}{\partial x_i} = \sum_r \sum_j [d(i, i^*(r,j)) < 1](1 - \Delta h)(1 - \Delta w) \frac{\partial L}{\partial y_{rj}} \tag{9}$$

Where $d(\cdot)$ represents the distance between two points, Δh and Δw represent the difference between the horizontal and vertical coordinates of i and $i^*(r,j)$. the extracted features are correctly aligned with the original region recommendations through RoIAlign, which helps to produce better pixel segmentation results.

4.4. Farmland segmentation and loss function

The features obtained by RoIAlign are fed to the full connection (FC) layer for classification and bounding box regression, and also to the convolution layer for segmentation. Classification is accomplished by passing the output of the FC layer using all features to the softmax layer.

For network training, the loss function represents the estimation of the gap between the prediction result and the target. In our farmland segmentation network of Mask R-CNN based on the feature pyramid of double attention mechanism, the total loss function is mainly composed of three parts: the classification loss of candidate boxes, the location regression loss and the target mask loss. The loss function used is as follows:

$$L = L_{cls} + L_{bbox} + L_{mask} \tag{10}$$

L_{cls} refers to the loss of category, and the calculation formula is as follows:

$$L_{cls} = \frac{1}{N_{cls}} \sum_i -\log[p^*_i p_i + (1 - p^*_i)(1 - p_i)] \tag{11}$$

Where p_i is the probability that the ROI of sequence number i is predicted to be a positive sample, N_{cls} is the normalization parameter, $p^*_i = 0$ is the negative sample in the proposed area, and $p^*_i = 1$ is the positive sample in the proposed area.

L_{bbox} represents the regression loss of the bounding box, and the calculation formula is as follows:

$$L_{bbox} = \frac{1}{N_{reg}} \sum_i p^*_i R(t_i, t^v_i) \tag{12}$$

$$Smooth_{L1} = \begin{cases} 0.5X^2 if|X| < 1 \\ |X| - 0.5 otherwise \end{cases} \tag{13}$$

Where N_{reg} is the normalized parameter, t_i is the predicted migration parameter, t^v_i is the actual migration parameter, $p^*_i = 1$ and $p^*_i = 0$ represent the positive and negative samples of the proposed area respectively, and R is the $Smooth_{L1}$ loss.

L_{mask} represents the mask loss, and the calculation formula is as follows:

$$L_{cls} = \frac{1}{N_{cls}} \sum_{i} - \log[p_i^* p_i + (1 - p_i^*)(1 - p_i)] \quad (14)$$

Where y_v is the target true tag value, and y_v^k is the predicted value in Mask R-CNN.

5. Evaluation index

AP (average precision) is a commonly used evaluation index in target detection and instance segmentation tasks, which represents the average accuracy of a single category. The calculation of AP needs to involve accuracy (P) and recall (R). F1-score is the harmonic average of accuracy and recall. F1-score gives a good overview of the area difference between the reference section and the test section. 1 represents the best output of the model, 0 represents the worst output of the model, and the mathematical expression is as follows:

$$P = \frac{T_P}{T_P + F_P} \times 100\% \quad (15)$$

$$R = \frac{T_P}{T_P + F_N} \times 100\% \quad (16)$$

$$F1 - score = \frac{2P \cdot R}{2P + R} \times 100\% \quad (17)$$

Where T_P means that the farmland positive sample is predicted to be a positive sample, F_P means that the farmland negative sample is predicted to be a positive sample, F_N means that the negative sample type is predicted to be a negative sample, and AP is equal to the area under a certain type of precision recall curve. According to the different setting of IOU threshold, it can be specifically divided into AP, AP_{50}, AP_{75}. AP refers to taking the value of IOU threshold from 0.5 to 0.95 in steps of 0.05, and then taking the average value of AP under each threshold. Ap_{50} is the AP value when the IOU threshold is 0.5, and AP_{75} is the AP value when the IOU threshold is 0.75. In this paper, AP, AP_{50}, AP75 and F1-score will be used to evaluate the experimental results.

6. Experimental

6.1. Network training

During the training, the software platform in this paper adopts Pytortch1.10.0 + cu102 and torchvision0.11.0 + cu102 with GPU under Windows system to build the farmland segmentation model, and the whole algorithm is implemented by python. The details of hardware information in this experiment are shown in Table 1. The data set used in this paper includes 317 satellite images, a total of 3805 labels. We divide these tags into training set, verification set and test set in the ratio of 7:2:1. The learning rate is 0.02 and the number of iterations is 200epoch. In addition, the pre-training ResNet50 weight based on the COCO dataset is also used to accelerate the training process (Lin et al., 2014).

In order to verify the accuracy of the improved mask R-CNN in farmland segmentation, ablation experiments were carried out to verify the unchanged mask R-CNN, the CBAM attention mechanism and the DAFPN based mask R-CNN network, respectively, to achieve the segmentation performance comparison with the improved mask R-CNN.

Table 1
The experimental hardware.

Configuration	Parameter
CPU	Intel Core i9-11900 K
GPU	Nvidia GeForce RTX 3060Ti
Development environment	Windows 10
Memory	16 G
Hard disk	1 TB

6.2. Smallholder farmland segmentation results

As shown in Fig. 7, for regular farmland (Fig. 7a and I), the Mask R-CNN segmentation result cannot distinguish the fuzzy boundary, and the Mask R-CNN segmentation result after adding CBAM attention mechanism is significantly improved compared with the Mask R-CNN method, but it also fails to segment accurately, and our method is more accurate than the above two methods. For the small and irregular shaped farmland shown in Fig. 7e, the Mask R-CNN detection and segmentation results will miss recognition for some large farmland (Fig. 8f). After adding the attention mechanism, the detection rate is further improved. The recognition and segmentation results of our method (Fig. 8h) are more satisfactory than CBAM attention mechanism. In general, our method compares Mask R-CNN and the network with CBAM attention mechanism to achieve the best segmentation performance, as shown in Fig. 8 and Table 2.

However, there are several common segmentation problems when our model is applied to farmland segmentation, as shown in Fig. 8. The first common depiction error is segmentation overlap (Fig. 8b, f and j). The same field is identified as two or more fields. When the field boundary changes not significantly, there is also excessive segmentation or irregular polygons are generated on multiple fields (Fig. 8b and j). When the brightness of the field satellite image changes greatly, there is also insufficient segmentation (Fig. 8j and d). When the field boundary is very small, the model can never detect any polygons (Fig. 8l and h).

6.3. Comparison with other instance segmentation methods

In order to further analyze the performance of the improved Mask R-CNN farmland instance segmentation method based on Dual Attention Pyramid Network (DAPPN), standard COCO evaluation indicators such as AP, AP_{50}, AP_{75}, and F1 scores were used to evaluate the method. The performance of the method was compared with Mask R-CNN, SOLO, Mask2Former, YOLACT, BlendMask, and Mask R-CNN instance segmentation method with CBAM attention mechanism. We use the same training, validation, and testing set to train and test these networks. The target detection and segmentation results of these methods on the test set are shown in Tables 2 and 3.

In farmland target detection, the F1 values of SOLO, Mask2Former, BlemdMask, Mask R-CNN with CBAM, and Mask R-CNN with DAFPN are all greater than 70%, and our method has the highest F1 value, reaching 75.9%, while Mask R-CNN and YOLACT have the lowest F1 value. And by comparing AP, AP_{50}, and AP_{75}, it can be found that the data of AP_{50} is at the highest, and choosing an IOU threshold greater than or equal to 0.5 is more suitable for this study. And our method performs excellently in both AP_{50} and AP_{75}, reaching 86.66% and 62.29% respectively. In terms of farmland target segmentation, both Mask2Former and our method have F1 values of 70%, and our method is 0.59% higher than Mask2Former. And on the evaluation indicators of AP, AP_{50}, and AP_{75}, our method achieved 49.95%, 82.86%, and 55.51% respectively, achieving good performance, proving the excellent accuracy of our method in farmland segmentation. From the comparison results, it can be concluded that the Mask R-CNN farmland instance segmentation method based on DAFPN proposed in this study can more effectively and accurately segment farmland.

Farmland in remote sensing images is a visual object with complex knot shapes and rich texture features. Even within the same species, there are significant differences in form and color. The main reason why our method achieved good recognition is that DAFPN combines the ideas of top-down and bottom-up methods, integrating instance level rich spatial information and accurate pixel channel features, making it very suitable for the agricultural remote sensing images in this study. When Mask2Form was proposed, it was the best instance segmentation network on the COCO dataset. When Swin Transformer was selected as its backbone network, its performance was the best. At the same time, SOLO used instance classes to achieve direct instance segmentation,

Fig. 7. Five example segmentation of 1024 × 1024-pixel satellite images. The column of satellite image is RGB image with ground reality. Mask R-CNN is the prediction result of the original Mask R-CNN. Mask R-CNN with CBAM is the prediction result after adding CBAM attention mechanism. Mask R-CNN with DAFPN lists the prediction results after adding the pyramid feature network of double attention mechanism.

which is not affected by target detection. It can achieve good segmentation results by segmenting instances of different objects with pixel level feature alignment. The main reason for the poor recognition performance of YOLACT is that it is a single stage instance segmentation network that uses a global image-based method to process images. This method can better preserve the position information of objects. However, in cases where the texture of the farmland boundary is not clear, it may not be able to accurately locate the farmland boundary, leading to errors.

7. Discussion

Our work shows the potential of Mask R-CNN model based on DAFPN in segmenting small farmland and DAFPN in applying to other models, even in images with very small and variable fields. In general, our model correctly segmented most of the farmland in the image (Fig. 7), but underestimated the fuzzy degree of the farmland boundary in the image (Fig. 8), which may be caused by the complete convolution network ignoring the object boundary that is difficult to classify (Cheng et al., 2020). Although our overall accuracy is moderate, there are several consistent problems that have reduced the prediction accuracy of our model. Specifically, the Mask R-CNN based on DAFPN is challenged when the field boundary, field shadow change or field shape is not obvious in the image (Fig. 8). Considering the above problems, and considering that the human eye sometimes cannot see the boundary, it is not surprising that the model performs poorly when the farmland boundary is fuzzy. Shadows cause color changes in the image, which leads to undetectable farmland or insufficient segmentation. When the farmland boundary in the image is blurred, our model will also lead to over segmented farmland or segmentation overlap, although this

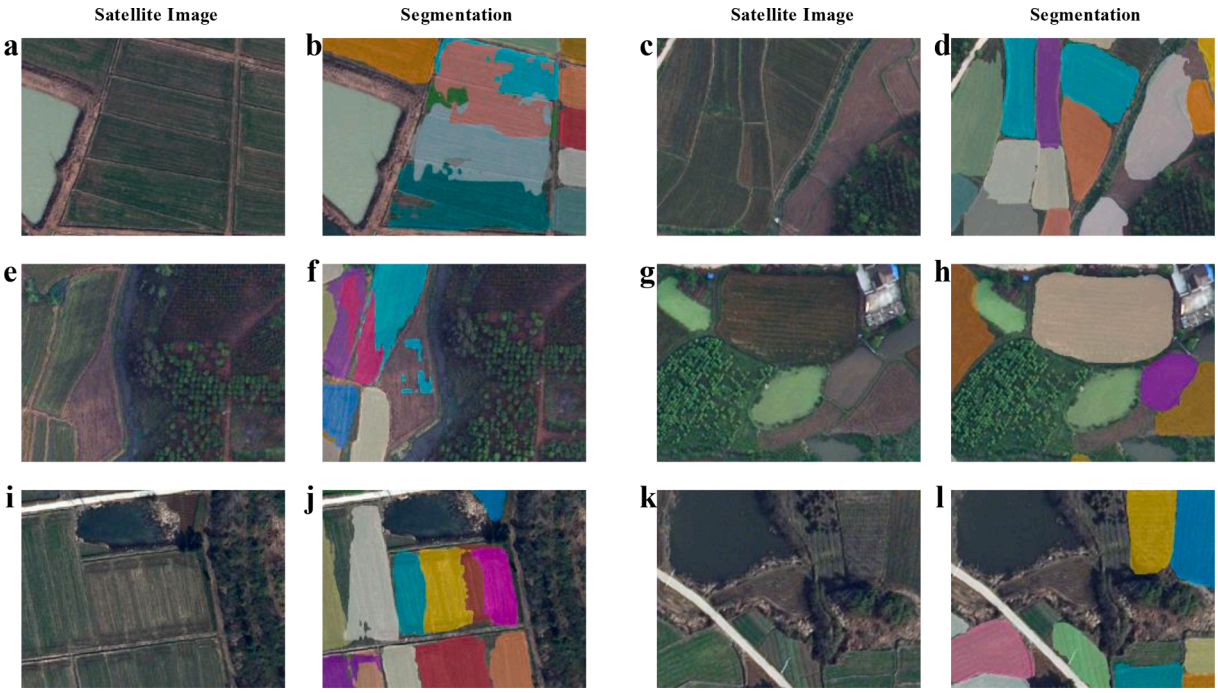


Fig. 8. Examples of common problems in farmland segmentation. The column satellite image refers to RGB images with ground reality, and the column segmentation refers to the predicted segmentation results.

Table 2
Standard COCO metrics results of IoU bounding box.

Methods	AP (%)	AP_{50} (%)	AP_{75} (%)	F1-score (%)
Mask R-CNN	45.45	77.88	48.56	68.99
SOLO	48.42	80.32	53.29	70.03
Mask2Former	53.12	83.46	57.38	70.82
YOLACT	44.56	75.18	47.84	66.83
BlendMask	49.88	81.32	55.58	70.35
Mask R-CNN with CBAM	**55.00**	85.29	61.28	70.91
Mask R-CNN with DAFPN	54.96	**86.66**	**62.29**	**75.90**

Table 3
Standard COCO metrics results of IoU segmentation box.

Methods	AP (%)	AP_{50} (%)	AP_{75} (%)	F1-score (%)
Mask R-CNN	43.19	74.16	47.20	64.03
SOLO	45.90	77.83	52.56	65.76
Mask2Former	49.28	81.23	55.39	70.31
YOLACT	42.31	72.38	46.10	61.28
BlendMask	48.59	78.48	53.43	68.65
Mask R-CNN with CBAM	49.72	80.67	55.45	69.68
Mask R-CNN with DAFPN	**49.95**	**82.86**	**55.51**	**70.90**

phenomenon is rare. Considering the long, narrow and small farmland, we think our model performs poorly, because these areas are also relatively rare in the landscape, resulting in a small proportion of training data for relevant features. Although these problems are not very common in the whole environment, the existence of these problems still leads to medium segmentation and detection accuracy.

Although our results show that Mask R-CNN based on DAFPN can successfully map small farmland, there are several important ways for future work. First, future work should examine the extent to which the model can be extended to more different smallholder systems. Our model has higher accuracy than Mask R-CNN in both large regional farmland and small irregular farmland. Secondly, we focus on training and testing only using images in the same time period, but images of crop fields with multiple phenological differences may improve the performance of farmland segmentation, as found in previous work (Aung et al., 2020; Debats et al., 2016a). Third, the future work should use the field farmland spatial information collected on the ground as the standard to evaluate the accuracy of the model. Although the farmland boundary is visible to a large extent when we process the data and mark it, in a few cases, because the boundary is not obvious, we can only determine the boundary based on our own judgment, and we cannot accurately digitize the boundary.

8. Conclusion

In this study, we developed a Mask R-CNN based on double attention mechanism feature pyramid network(DAFPN) to accurately segment small farmland in high-resolution satellite images. In order to improve the feature extraction ability of backbone networks, an improved feature pyramid network model integrating double attention mechanism is designed based on Mask R-CNN network. The model consists of two attention modules: spatial attention module (SPA) and channel attention module (CHA), which optimize the ability of spatial feature extraction and channel feature extraction respectively. Compared with the original Mask R-CNN, the improved network model shows superior segmentation performance. The example segmentation method of Mask R-CNN based on DAFPN can effectively and accurately segment small farmland. The AP, AP_{50}, AP_{75} and F1 score of our method are 49.95%, 82.86%, 55.51% and 70.90% respectively. The proposed method effectively divides small-scale farmland. Finally, we develop a feature pyramid network (DAFPN) with dual attention mechanism. Our model can be easily inserted into the existing FPN structure and used in combination with other models. However, the network model is slightly larger, and the segmentation accuracy needs to be further improved. In the future, we will further simplify the method of network structure.

Funding

The authors would like to acknowledge the support of this study by the following funding sources: Equipment and Practices for the Fully Mechanized Production of Staple Forage and Fodder, the State Key

Research Program of China, a project sponsored by the Ministry of Science and Technology of China (Grant No. 2022YDF2001901-01), the Guangdong Provincial Department of Agriculture's Modern Agricultural Innovation Team Program for Animal Husbandry Robotics (Grant No. 2019KJ129), the Vehicle Soil Parameter Collection and Testing Project (Grant No. 4500-F21445), and the Special Project of Guangdong Provincial Rural Revitalization Strategy in 2020 (YCN[2020] No. 39) (Fund No. 200–2018-XMZC-0001–107–0130).

CRediT authorship contribution statement

Yangyang Cao: Conceptualization, Methodology, Software, Investigation, Formal analysis, Writing – original draft. **Zuoxi Zhao:** Conceptualization, Funding acquisition, Resources, Supervision, Writing – review & editing. **Yuan Huang:** Data curation, Writing – original draft. **Xu Lin:** Visualization, Investigation. **Shuyuan Luo:** Resources, Supervision. **Borui Xiang:** Visualization, Writing – review & editing. **Houcheng Yang:** Software, Validation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

Aung, H.L., Uzkent, B., Burke, M., Lobell, D., Ermon, S., 2020. Farm parcel delineation using spatio-temporal convolutional networks. In: In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 76–77.

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. Segnet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE T. Pattern Anal. 39 (12), 2481–2495.

Bergado, J.R., Persello, C., Gevaert, C., 2016. A deep learning approach to the classification of sub-decimetre resolution aerial images. In: 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS). IEEE, IEEE, pp. 1516–1519.

Bertasius, G., Shi, J., Torresani, L., 2015. Deepedge: A multi-scale bifurcated deep network for top-down contour detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4380-4389.

Canny, J., 1986. A computational approach to edge detection. IEEE T. Pattern Anal. 6, 679–698.

Chen, B., Qiu, F., Wu, B., Du, H., 2015. Image segmentation based on constrained spectral variance difference and edge penalty. Remote Sens.-Basel. 7 (5), 5980–6004.

Chen, L., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 801–818.

Cheng, T., Wang, X., Huang, L., Liu, W., 2020. Boundary-preserving mask r-cnn. In: European Conference on Computer Vision. Springer, Springer, pp. 660–676.

Cheng, G., Yang, C., Yao, X., Guo, L., Han, J., 2018. When deep learning meets metric learning: remote sensing image scene classification via learning discriminative CNNs. IEEE T. Geosci. Remote. 56 (5), 2811–2821.

Crommelinck, S., Bennett, R., Gerke, M., Yang, M.Y., Vosselman, G., 2017. Contour detection for UAV-based cadastral mapping. Remote Sens.-Basel. 9 (2), 171.

Debats, S.R., Luo, D., Estes, L.D., Fuchs, T.J., Caylor, K.K., 2016a. A generalized computer vision approach to mapping crop fields in heterogeneous agricultural landscapes. Remote Sens. Environ. 179, 210–221.

Deng, J., Dong, W., Socher, R., Li, L., Li, K., Fei-Fei, L., 2009. Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. Ieee, Ieee, pp. 248–255.

Evans, C., Jones, R., Svalbe, I., Berman, M., 2002. Segmenting multispectral Landsat TM images into field units. IEEE T. Geosci. Remote. 40 (5), 1054–1064.

Feng, Y., Diao, W., Zhang, Y., Li, H., Chang, Z., Yan, M., Sun, X., Gao, X., 2019. Ship instance segmentation from remote sensing images using sequence local context module. In: IGARSS 2019–2019 IEEE International Geoscience and Remote Sensing Symposium. IEEE, IEEE, pp. 1025–1028.

Fu, G., Liu, C., Zhou, R., Sun, T., Zhang, Q., 2017. Classification for high resolution remote sensing imagery using a fully convolutional network. Remote Sens.-Basel. 9 (5), 498.

Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena-Martinez, V., Garcia-Rodriguez, J., 2017. A review on deep learning techniques applied to semantic segmentation. arXiv preprint arXiv:1704.06857.

Girshick, R., 2015. Fast r-cnn. In: In: Proceedings of the IEEE international conference on computer vision, pp. 1440–1448.

Hariharan, B., Arbeláez, P., Girshick, R., Malik, J., 2014. Simultaneous detection and segmentation. In: European Conference on Computer Vision. Springer, Springer, pp. 297–312.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn. In: In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969.

Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks. In: In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141.

Kettig, R.L., Landgrebe, D.A., 1976. Classification of multispectral image data by extraction and classification of homogeneous objects. IEEE Trans. Geosci. Electron. 14 (1), 19–26.

Kienzle, J., 2013. Precision agriculture for smallholder farmers. Agric. Development. 19, 12–15.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, p. 25.

Lecun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. P. Ieee. 86 (11), 2278–2324.

Lin, T., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: common objects in context. In: European Conference on Computer Vision. Springer, Springer, pp. 740–755.

Lin, T., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017. Feature pyramid networks for object detection. In: In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125.

Liu, S., Qi, L., Qin, H., Shi, J., Jia, J., 2018. Path aggregation network for instance segmentation. In: In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8759–8768.

Ma, J., Chen, B., 2020. Dual refinement feature pyramid networks for object detection. arXiv preprint arXiv:2012.01733.

Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P., 2016. Convolutional neural networks for large-scale remote-sensing image classification. IEEE T. Geosci. Remote. 55 (2), 645–657.

Maninis, K., Pont-Tuset, J., Arbeláez, P., Van Gool, L., 2017. Convolutional oriented boundaries: from image segmentation to high-level tasks. IEEE T. Pattern Anal. 40 (4), 819–833.

Martin, D.R., Fowlkes, C.C., Malik, J., 2004. Learning to detect natural image boundaries using local brightness, color, and texture cues. IEEE T. Pattern Anal. 26 (5), 530–549.

Marvaniya, S., Devi, U., Hazra, J., Mujumdar, S., Gupta, N., 2021. Small, sparse, but substantial: techniques for segmenting small agricultural fields using sparse ground data. Int. J. Remote Sens. 42 (4), 1512–1534.

Masoud, K.M., Persello, C., Tolpekin, V.A., 2019. Delineation of agricultural field boundaries from Sentinel-2 images using a novel super-resolution contour detector based on fully convolutional networks. Remote Sens.-Basel. 12 (1), 59.

Mueller, M., Segl, K., Kaufmann, H., 2004. Edge-and region-based segmentation technique for the extraction of large, man-made objects in high-resolution satellite imagery. Pattern Recogn. 37 (8), 1619–1628.

Nations, U., 2007. A system of integrated agricultural censuses and surveys: world programme for the census of agriculture 2010. Food and Agriculture Organization of the United Nations Rome, Italy.

Nevatia, R., Babu, K.R., 1980. Linear feature extraction and description. Comput. Graphics Image Process 13 (3), 257–269.

Pal, S.K., Mitra, P., 2002. Multispectral image segmentation using the rough-set-initialized EM algorithm. IEEE T. Geosci. Remote. 40 (11), 2495–2501.

Persello, C., Tolpekin, V.A., Bergado, J.R., De By, R.A., 2019. Delineation of agricultural fields in smallholder farms from satellite images using fully convolutional networks and combinatorial grouping. Remote Sens. Environ. 231, 111253.

Potlapally, A., Chowdary, P.S.R., Shekhar, S.R., Mishra, N., Madhuri, C.S.V.D., Prasad, A., 2019. Instance segmentation in remote sensing imagery using deep convolutional neural networks. In: 2019 International Conference on Contemporary Computing and Informatics (IC3I). IEEE, IEEE, pp. 117–120.

Quoc, T.T.P., Linh, T.T., Minh, T.N.T., 2020. Comparing U-Net convolutional network with mask R-CNN in agricultural area segmentation on satellite images. In: 2020 7th NAFOSTED Conference on Information and Computer Science (NICS). IEEE, IEEE, pp. 124–129.

Shawon, A.R., Ko, J., Ha, B., Jeong, S., Kim, D.K., Kim, H., 2020. Assessment of a proximal sensing-integrated crop model for simulation of soybean growth and yield. Remote Sens.-Basel. 12 (3), 410.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. In: In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9.

Tobler, W.R., 1970. A computer movie simulating urban growth in the Detroit region. Econ. Geogr. 46 (sup1), 234–240.

Turker, M., Kok, E.H., 2013. Field-based sub-boundary extraction from remote sensing imagery using perceptual grouping. ISPRS J. Photogramm. 79, 106–121.

Wagner, M.P., Oppelt, N., 2020. Extracting agricultural fields from remote sensing imagery using graph-based growing contours. Remote Sens.-Basel. 12 (7), 1205.

Waldner, F., Diakogiannis, F.I., 2020. Deep learning on edge: extracting field boundaries from satellite images with a convolutional neural network. Remote Sens. Environ. 245, 111741.

Waldner, F., Diakogiannis, F.I., Batchelor, K., Ciccotosto-Camp, M., Cooper-Williams, E., Herrmann, C., Mata, G., Toovey, A., 2021. Detect, consolidate, delineate: scalable mapping of field boundaries using satellite images. Remote Sens.-Basel. 13 (11), 2197.

Woo, S., Park, J., Lee, J., Kweon, I.S., 2018. Cbam: convolutional block attention module. In: In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19.
Yan, L., Roy, D.P., 2016. Conterminous United States crop field size quantification from multi-temporal Landsat data. Remote Sens. Environ. 172, 67–86.
Zhang, W., Witharana, C., Liljedahl, A.K., Kanevskiy, M., 2018. Deep convolutional neural networks for automated characterization of arctic ice-wedge polygons in very high spatial resolution aerial imagery. Remote Sens.-Basel. 10 (9), 1487.
Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid scene parsing network. In: In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2881–2890.
Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A., 2016. Learning deep features for discriminative localization. In: In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2921–2929.
Zhu, H., Meng, F., Cai, J., Lu, S., 2016. Beyond pixels: a comprehensive survey from bottom-up to semantic image segmentation and cosegmentation. J. Vis. Commun. Image R. 34, 12–27.