Investigation into Few-Shot Instance Segmentation of Agricultural Landscapes Utilizing SAM-Driven Model

Wenbo Chen [1, a], Hongbo Zhu [2, b], Chengwei Zhao [1, c], Xuqing Li [1, 3, 4, d, *], Tingxuan Wang [1, e], Zekun Zhang [1, f], Zhihe Hu [1, g]

[1] College of Remote Sensing and Information Engineering, North China Institute of Aerospace Engineering, Langfang Hebei, 065000, China

[2] Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China

[3] Agricultural Resources and Environmental Research Institute, Guangxi Academy of Agricultural Sciences/Guangxi Key Laboratory of Arable Land Conservation, Guangxi, China

[4] State Key Laboratory of Remote Sensing Science, Beijing Normal University. Beijing 100875, China

[a] 13290668106@163.com, [b] zhuhb@aircas.ac.cn, [c] zhaochengwei3414@163.com, [d] meililixuqing@163.com, [e] 13941817419@163.com, [f] 2536155481@qq.com, [g] 22774611711@qq.com

* Corresponding Author: Xuqing Li

Abstract. In complex agricultural landscapes, especially with limited sample sizes, current instance segmentation techniques often grapple with defining clear boundaries, thereby compromising segmentation accuracy. To mitigate this challenge, our research proposes enhancements through quantitative comparisons and the incorporation of the SAM model to bolster instance segmentation networks. Specifically, we have formulated SAM+Mask2Former and SAM+Mask R-CNN by adopting SAM's feature encoder as the foundational backbone. Further refinement is accomplished by integrating the Efficient Channel Attention (ECA) module and optimizing prompts, culminating in the SAM+ECA+Prompter model. Impressively, these models exhibit robust generalization capabilities when applied to high-resolution remote sensing imagery, even when based on smaller datasets. Empirical evidence demonstrates that, even with a mere 50 samples, the models achieve an Average Precision (AP) and Average Recall (AR) of 0.715. When scaled to 200 samples, AP surges to 0.875, while AR surpasses 0.903—figures that stand toe-to-toe with traditional field segmentation paradigms. Consequently, our research significantly elevates the efficacy of segmentation models, emphasizing a reduction in both labeling and computational expenses, thereby presenting a resource-efficient, high-precision solution for precision agriculture.

Keywords: Few-shot Learning; Quantification; Semantic Segmentation in Farmland; SAM Model; Cost.

1. Introduction

In precision agriculture, the effective monitoring and management of farmland crops are intrinsically linked to the precise analysis of remote sensing images, particularly in the context of crop planting and growth surveillance. Image segmentation technology is pivotal in the processing of these remote sensing images. This technology enables the accurate extraction of the target area from the intricate farmland environment, thereby providing robust data support for agricultural decision-making processes[1].

The advancement of deep learning technology has led to the instance segmentation method based on deep learning becoming increasingly prevalent. In the context of farmland instance segmentation tasks, Mask R-CNN effectively segments different crop instances and accurately extracts boundaries by integrating object detection and instance segmentation. However, it carries a high computational cost and performs suboptimally with small objects and indistinct boundaries[2]. Conversely,

Mask2Former merges the benefits of Transformer and CNN, enabling it to better manage complex backgrounds and multi-scale farmland images, demonstrating robust generalization capabilities. Nevertheless, its training process is intricate, and its performance may not meet expectations when dealing with limited sample data[3]. The Segment Anything Model(SAM)[4] exhibits considerable strengths in image segmentation tasks, particularly in versatility and unsupervised learning, as it can autonomously execute image segmentation and minimize manual labeling. However, SAM's adaptability may decrease in specific fields or low-resource environments, such as farmland scene segmentation tasks. Furthermore, due to the absence of category information and reliance on manual prompt provision, it cannot fully satisfy the requirements of existing farmland instance segmentation tasks[5].

In summary, this study introduces three enhanced models derived from the SAM framework and validates their performance in the task of farmland instance segmentation using a limited set of diverse test data. The findings indicate that the refined model, based on the SAM framework, can effectively segment farmlands even with a minimal sample size, achieving superior accuracy. This underscores the viability of farmland segmentation with limited samples. The experimental procedure is delineated in Fig. 1

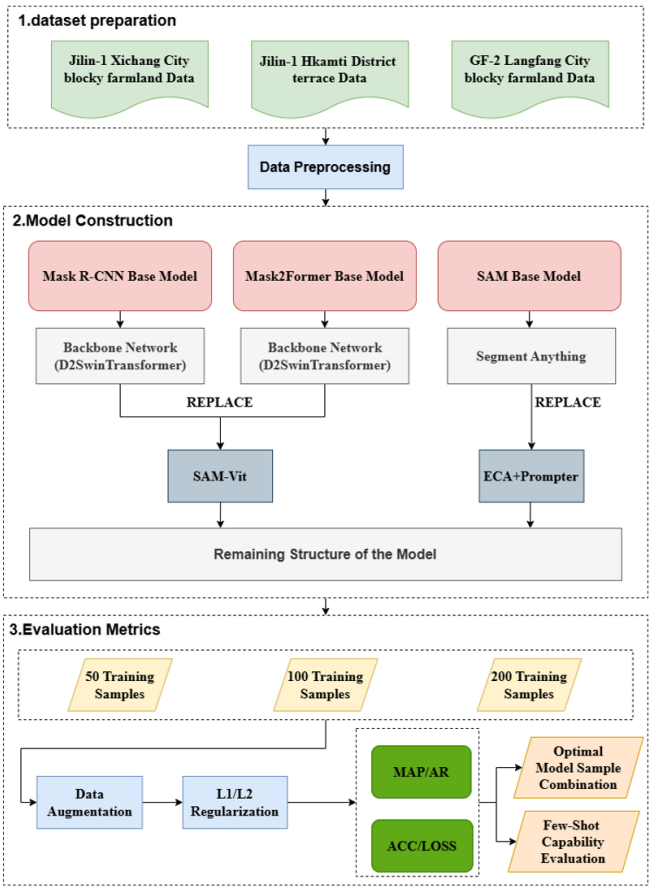


Fig 1. Flow chart

2. Study Area and Materials

2.1 Study Area

Farmland can be categorized based on its shape and spatial location into two types: blocky farmland and terraces. In areas with a relatively gentle terrain, a horizontal field construction project is implemented. This project typically features rectangular or near-rectangular geometry, aligned with the direction of irrigation and drainage. Paddy fields fall under this category. Conversely, terraced fields are found in regions with steeper ground slopes. The construction of these terraced fields is

contingent upon the specific terrain and contour lines, resulting in different sections. Terraced fields can further be divided into two subcategories: horizontal terraces and steep terraces[6].

To validate the stability of the proposed model, this study selected three representative areas: Langfang City in Hebei Province, Xichang City in Sichuan Province, and Kanti District in Myanmar, as depicted in Fig. 2. Langfang City, situated between 28°43' and 39° 52' north latitude and 116°40' and 117°56' east longitude, is characterized by its flat terrain, seasonal climate, expansive yet fragmented farmland distribution, and intricate boundary details. Xichang City, located between 27°48' and 28°43' north latitude and 101°25' and 102°22' east longitude, boasts a diverse climate, abundant rainfall, and fertile land suitable for various crops. The farmlands here are systematically arranged on level ground with distinct boundaries, facilitating the identification of different farmland segments. In contrast, the Kanti District in Myanmar, spanning from 23°20' to 25°30' north latitude and 94°44' to 97°14' east longitude, features terraced farmlands that are dispersed and meticulously adapted to the undulating terrain [7].

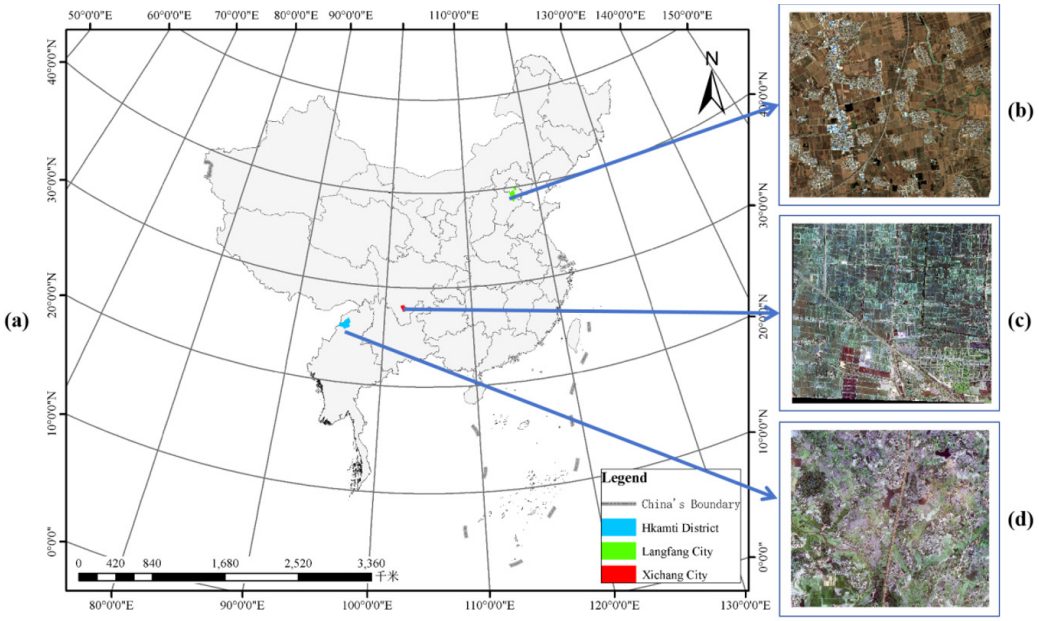


Fig 2. The study area is depicted in the accompanying sketch map. Location A represents the three areas under investigation. The farmland display maps for Langfang City in Hebei Province and Xichang City in Liangshan Yi Autonomous Prefecture, Sichuan Province are shown in B and C respectively. Section D illustrates the farmland distribution in Hkamti District, Sagaing Province, Myanmar.

2.2 Images and Preprocessing

In this research, a Jilin-1 Satellite image with a spatial resolution of 0.67 meters was chosen for Xichang City and Hkamti District, while a Gaofen-2 satellite image with a spatial resolution of 1 meter was selected for Langfang City. All images from the three study areas utilized blue, green, and red bands. The labeled vectors and images were segmented into 512x512 pixel sections using a sliding window technique. Subsequently, the data slices were cleaned to eliminate invalid sample pairs. For instance segmentation, the Coco type[8] was employed. In total, the data volumes for the three study areas were 85, 170, and 335, respectively. The datasets were partitioned into training, validation, and test sets at a ratio of 6:2:2, respectively. Additionally, the number of training samples was augmented through horizontal flipping.

3. Methods

SAM employs Masked Auto Encoders (MAE) for pre-training[9],. By segmenting the image into smaller blocks and masking certain blocks, the model effectively learns the image's feature representation. Once the training phase is finalized, SAM exclusively depends on the encoder to extract features. This ensures that the same image retains consistent features regardless of varying prompts. Such a characteristic renders SAM exceptionally proficient in tasks with limited samples.

This study addresses the challenges in farmland instance segmentation by proposing two schemes. These schemes utilize SAM's image encoder and convolution block for feature fusion, and integrate with the existing instance segmentation networks, Mask2Former and Mask R-CNN. The selection of these two networks is based on their robust performance in instance segmentation tasks, their ability to effectively identify and segment complex objects, and their flexibility and adaptability. By substituting the backbone of these two instance segmentation networks with the SAM encoder, which is based on the ViT structure, the feature representation learned by SAM can be fully leveraged. This substitution also reduces the sample learning amount and computational cost. Furthermore, SAM enhances its feature encoder's efficiency in extracting key image features by standardizing image size, and integrating convolution feature extraction and Transformer processing. The convolution operation captures local details, while the Transformer optimizes the capture of global features, thereby improving the accuracy of instance segmentation.

3.1 Splicing SAM Models

The network architecture of SAM and Mask2Former is designed for farmland instance segmentation. Upon preprocessing, the image data is fed into the high-precision Vit_h image encoder within SAM to extract both low-level and abstract features. These extracted features are then transformed into high-resolution pixel embeddings via the pixel encoder before being inputted into the Transformer decoder. The self-attention mechanism is employed to discern global dependencies, subsequently generating a prediction mask and categorizing each instance. The loss function comprises three components: the instance segmentation cross-entropy loss, which gauges the discrepancy between the predicted and actual masks; the boundary IOU loss, which bolsters the model's ability to recognize object boundaries; and the occlusion perception loss, which addresses occlusion challenges[10].

In the Mask R-CNN model, we use the Vit_h image encoder to extract feature maps. Then, candidate regions are generated by the Region Proposal Network (RPN), which includes bounding boxes and foreground probabilities to reduce computational cost and improve detection accuracy[2]. For each candidate region, an ROI Align is extracted and mapped to a fixed-size feature map, which is used for instance segmentation through the segmentation head to provide semantic information. The loss function consists of three parts: cross-entropy classification loss measures the similarity between the true and predicted farmland types; smoothed L1 regression loss measures the difference between the real farmland area and the predicted bounding box; and average binary cross-entropy mask loss measures the accuracy of mask coverage, so that the model can pay attention to important farmland areas.[11].

Mask2Former and Mask R-CNN exhibit similarities in the areas of feature extraction, loss calculation, and instance segmentation. Both models employ SAM's Vit_h encoder and utilize specific loss functions to optimize performance. They generate outputs of target categories and prediction masks, thereby facilitating efficient farmland segmentation, as illustrated in Fig. 3.

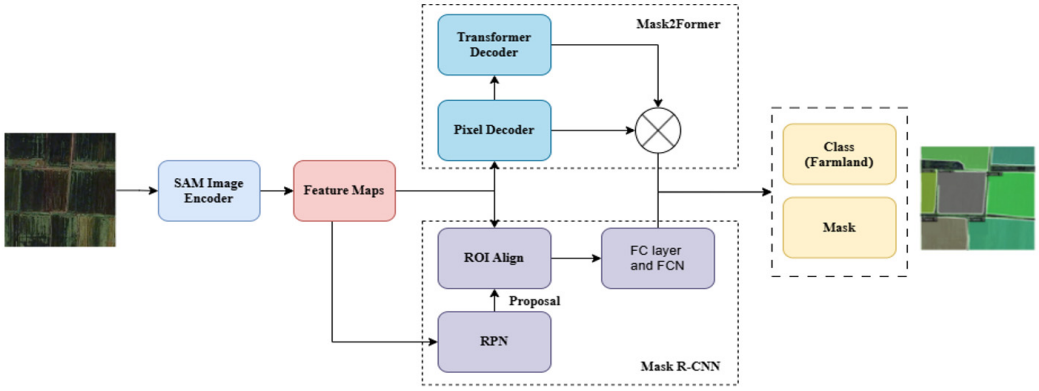


Fig 3. The structure of the splicing model with SAM

3.2 SAM-based Model Improvements

Fig. 4 illustrates an enhancement that amalgamates the interactive features of the SAM model with a learnable mask generation mechanism, thereby significantly improving farmland type recognition and instance segmentation. Through the training of a SAM-oriented Prompter, the model's generalization capacity is augmented, facilitating more effective inference of object categories and instance masks.

In scenarios with limited data samples, overfitting is a common issue. To enhance generalization, the Efficient Channel Attention (ECA) module is introduced, focusing on task-relevant channel information and preventing the loss of information typically seen in traditional dimensionality reduction methods. When integrated with the Region Proposal Network (RPN), the model can identify potential farmland areas, align visual features with spatial data, and generate candidate boxes. The Region of Interest (ROI) pooling is then used to capture visual features for each candidate box. These features are processed through three heads: a semantic head for category recognition, a localization head for instance-prompt association, and a prompt head for embeddings required by the SAM mask decoder.

The architecture of the model employs a pre-trained ViT_h model to extract and consolidate farmland features, subsequently generating input prompts that incorporate spatial and contextual information. These prompts are integrated with image features via cross-attention, thereby strengthening feature relationships prior to the final segmentation conducted by the SAM mask decoder. The loss function encompasses RPN binary classification, semantic and localization losses, as well as the SAM decoder segmentation loss [12] [13]. This approach facilitates balanced classification, localization, and segmentation while simultaneously reducing model complexity and data dependency.

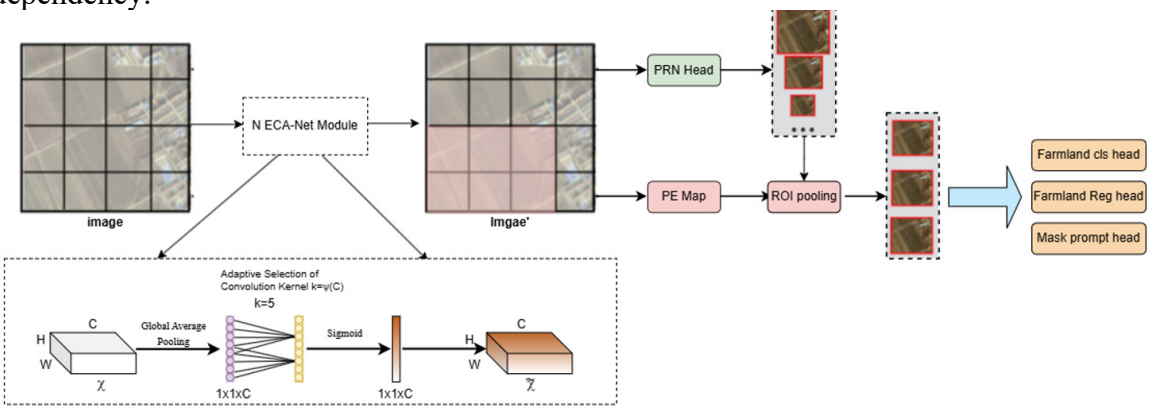


Fig 4. The structure of SAM + ECA + Prompter model

4. Result and Analysis

4.1 Experimental Settings

The enhanced network model is executed within the PyTorch framework, utilizing a 24GB GTX 3090 GPU. The AdamW optimizer is employed, with an initial learning rate set at 2e-4. To mitigate overfitting due to limited sample sizes, a cosine annealing strategy dynamically adjusts the learning rate and augments both L1 and L2 regularization constraints. Notably, the SAM+ECA+Prompter model exclusively incorporates the training prompt component of SAM, while other parameters remain frozen. Given that each study area exhibits distinct characteristics and distributions, training and evaluation are conducted independently for each dataset, facilitating an examination of the model's performance across various regions and farmland types. Each enhanced model is trained with a batch size of 4, over 200 rounds. To assess the impact of sample size on the model, training was performed with 50, 100, and 200 samples respectively. Experimental outcomes validate the model's accuracy using the AP formula (1) and AR formula (2) from the COCO dataset.

$$AP = \sum_{i=1}^{n-1}(r_{i+1}-r_i)P_{inter(r_i+1)} \quad (1)$$

$$AR = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{K}\sum_{j=1}^{K}\max_{d\in D_i}\left(IoU\left(d, g_{i,j}\right) \geq t\right) \quad (2)$$

4.2 Analysis of the Effect of Sample Size on Improved Model Accuracy

Table 1. The AP (%) and AR (%) values of the three methods in the study area of Xichang City under different sample size conditions

Model	Study Quantity	AP_{Mask}	$AP^{50}{}_{Mask}$	$AP^{75}{}_{Mask}$	AR^{100}	AR_{medium}	AR_{large}
SAM+ Mask2Fromer	50	0.512	0.686	0.608	0.282	0.184	0.422
	100	0.628	0.732	0.683	0.383	0.290	0.489
	200	0.681	0.848	0.752	0.708	0.643	0.810
SAM+ Mask R-CNN	50	0.494	0.679	0.602	0.343	0.225	0.515
	100	0.644	0.721	0.667	0.464	0.384	0.568
	200	0.546	0.837	0.778	0.653	0.564	0.790
SAM+ECA Prompter	50	0.498	0.715	0.553	0.398	0.279	0.581
	100	0.519	0.785	0.628	0.529	0.466	0.625
	200	0.771	0.875	0.788	0.753	0.664	0.903

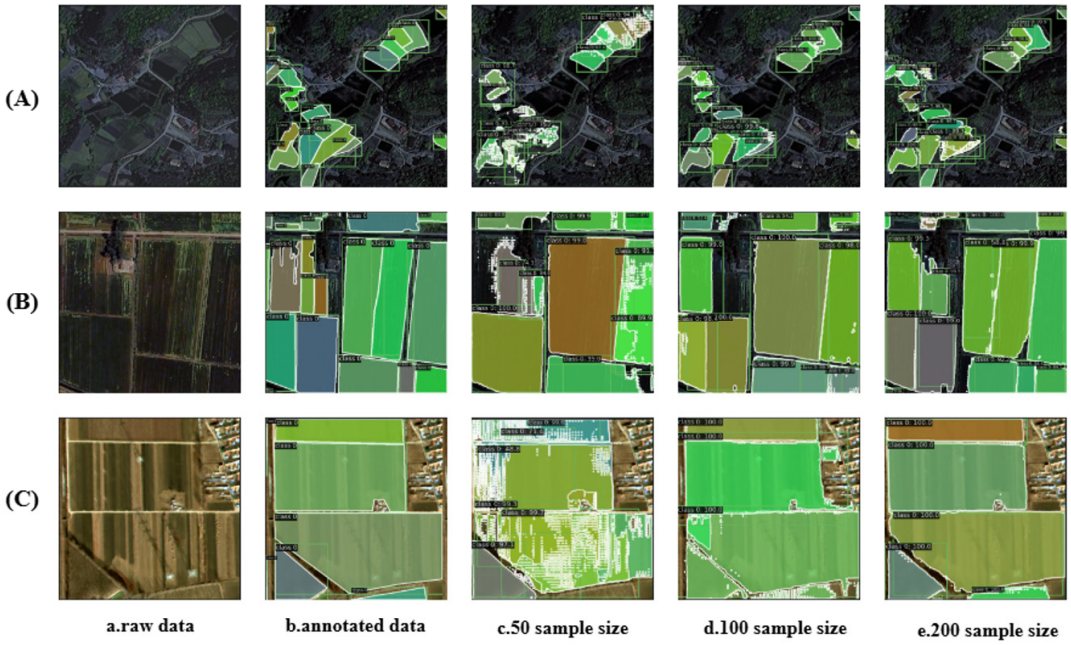


Fig 5. The segmentation results of three methods in 100 sample farmland examples

Table. 1 demonstrates that as the sample size increases, the accuracy of the three methods within the same study area significantly improves. Despite a smaller sample size of 50 being less than ideal for conventional training, as illustrated in Fig. 5, the results remain satisfactory. When the sample size is increased to 100 or 200, the model's accuracy approaches that of traditional big data training models, with the highest index reaching 0.903. This suggests that the sample size plays a crucial role in determining the model's accuracy. Furthermore, even with a smaller sample size, the model can still fulfill the requirements of conventional tasks. When the IOU threshold is set to 50 and the farmland area is large, the accuracy of the three models reaches its peak performance.

4.3 Accuracy Analysis in Three Areas with Limited Samples

Table 2. The AP(%) and AR(%) values of the three methods under the conditions of 100 sample sizes and different study areas.

Model	Study Area	AP_{Mask}	$AP^{50}{}_{Mask}$	$AP^{75}{}_{Mask}$	AR^{100}	AR_{medium}	AR_{large}
SAM+ Mask2Fromer	Xichang	0.628	0.732	0.683	0.383	0.290	0.489
	Hkamti	0.614	0.715	0.672	0.347	0.310	0.421
	Langfang	0.623	0.722	0.689	0.426	0.371	0.554
SAM+ Mask R-CNN	Xichang	0.644	0.721	0.667	0.464	0.384	0.568
	Hkamti	0.626	0.703	0.665	0.297	0.334	0.367
	Langfang	0.634	0.711	0.682	0.343	0.129	0.432
SAM+ECA Prompter	Xichang	0.519	0.785	0.628	0.529	0.466	0.625
	Hkamti	0.504	0.750	0.617	0.295	0.298	0.495
	Langfang	0.530	0.778	0.624	0.358	0.152	0.447

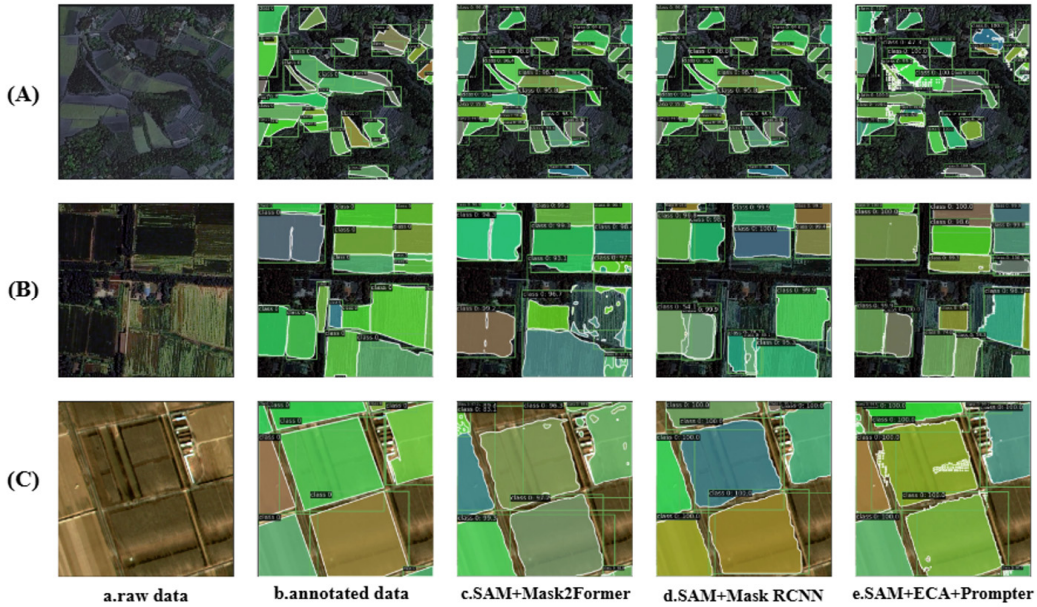


Fig 6. SAM+ECA+Prompter segmentation results of 100 samples of farmland in three study areas

As demonstrated in Table. 2, when the sample size is held constant, the optimal accuracy of the three methods varies significantly upon changing the study area. The strip data from Xichang exhibits the highest accuracy, followed by that from Langfang City, with the terrace data from Hkamti District being the least accurate. As illustrated in Fig. 6, the blocky farmland data in Xichang is notably consistent and possesses distinct segmentation boundaries, enhancing model precision. While Langfang's distribution characteristics resemble those of Xichang, its color attributes are more intricate, rendering its accuracy slightly inferior to Xichang. The terrace data from Kangra yields the lowest accuracy due to its high data complexity, limited effective terrace data in the training samples,

and varied shapes. Despite satisfactory recognition performance for specific graphics, the overall accuracy remains suboptimal. Through rigorous evaluation of indicators and segmentation outcomes, it has been ascertained that 100 samples suffice for cost-effective and precise recognition model training.

5. Summary

Addressing the complexity of farmland scenes and the scarcity of high-quality labeled data, this study introduces a solution grounded in small sample learning. The SAM model is integrated with Mask2Former and Mask R-CNN, while the ECA + urging module is employed to tackle the issue of data scarcity in the segmentation of farmland crop instances. Furthermore, by creating a small sample dataset for quantitative comparison, the model's instance segmentation capability across various farmland types and research areas is validated. Experimental results demonstrate that the model can achieve high segmentation accuracy with a certain sample size, and maintains robust performance even with a limited sample size, thereby meeting practical requirements. Quantitative comparison experiments and visual analysis also confirm the model's scalability.

This study refines the conventional model to accommodate low-sample learning, offering quantitative comparative experimental guidance for future model optimization under limited sample conditions. Future research could delve into the application of meta-learning, transfer learning, and other techniques to extend to various crop or instance segmentation tasks, thereby augmenting the universality and practical potential of the model.

Acknowledgments

This research was funded by the Crop classification in high-resolution remote sensing imagery based on deep learning (YKY-2022-59), the Central government guide local science and technology development fund project Science and technology innovation base project (246Z7401G), the Supported by Open Fund of State Key Laboratory of Remote Sensing Science (Grant No. OFSLRSS202303), Fund of Guangxi Key Laboratory of Arable Land Conservation (23-026-12-24KF7).

References

[1] Yuan X, Shi J, Gu L. A review of deep learning methods for semantic segmentation of remote sensing imagery[J]. Expert Systems with Applications, 2021, 169: 114417.

[2] He K, Gkioxari G, Dollár P, et al. Mask r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2961-2969.

[3] Cheng B, Misra I, Schwing A G, et al. Masked-attention mask transformer for universal image segmentation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 1290-1299.

[4] Kirillov A, Mintun E, Ravi N, et al. Segment anything[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 4015-4026.

[5] Chen K, Liu C, Chen H, et al. RSPrompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024.

[6] Ministry of Land and Resources of the People's Republic of China. (2012). High-standard Basic Farmland Construction Standards: TD/T 1033-2012 [S].

[7] Zhao Z, Liu Y, Zhang G, et al. The Winning Solution to the iFLYTEK Challenge 2021 Cultivated Land Extraction from High-Resolution Remote Sensing Images[C]//2022 14th International Conference on Advanced Computational Intelligence (ICACI). IEEE, 2022: 376-380.

[8] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context[C]//Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer International Publishing, 2014: 740-755.

[9] He K, Chen X, Xie S, et al. Masked autoencoders are scalable vision learners[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 16000-16009.

[10] Chen Z, Zhou H, Lai J, et al. Contour-aware loss: Boundary-aware learning for salient object segmentation [J]. IEEE Transactions on Image Processing, 2020, 30: 431-443.

[11] Badrinarayanan V, Kendall A, Cipolla R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 39(12): 2481-2495.

[12] Wang Q, Wu B, Zhu P, et al. ECA-Net: Efficient channel attention for deep convolutional neural networks [C]// Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 11534-11542.

[13] Chen L C, Zhu Y, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 801-818.