FL-DBENet: Double-branch encoder network based on segment anything model for farmland segmentation of large very-high-resolution optical remote sensing images

Wenqing Feng[1,*], Fangli Guan[1], Chenhao Sun[2], Wei Xu[1,3]

[1] School of Computer Science, Hangzhou Dianzi University, Hangzhou, P.R. China
(corresponding author, e-mail: wq_feng@hdu.edu.cn)
[2] Electrical & Information Engineering School, Changsha University of Science & Technology, Changsha, P.R. China
[3] Information System and Management College, National University of Defense Technology, Changsha, P.R. China

Keywords: Farmland extraction, remote sensing images, vision transformer, Segment Anything Model, SegFormer.

Abstract

Extracting farmland from very-high-resolution optical remote sensing images is a challenging task. Although deep learning algorithms have been extensively applied to farmland extraction, their performance remains limited due to the scarcity of labeled farmland samples and restricted generalization capabilities. The recent introduction of the Segment Anything Model (SAM), based on the Vision Transformer (ViT) architecture, has brought transformative advancements to remote sensing image analysis for farmland extraction. This paper introduces FL-DBENet, a farmland extraction network that builds on SAM’s strengths. FL-DBENet features a general-specialized double-branch encoder network: the general branch leverages SAM’s robust edge detection to capture precise farmland boundaries, while the specialized branch incorporates the lightweight SegFormer encoder to provide SAM with targeted prompts on farmland features. To further streamline the model, we integrate a Low-Rank Adaptation (LoRA) module into SAM’s image encoder, reducing training parameters and computational demands. Additionally, a prompt mixer module is developed to integrate diverse features effectively. Extensive evaluations on the GID dataset and the ultra-high resolution, ultra-rich context (URUR) dataset demonstrate that FL-DBENet achieves superior performance in both qualitative and quantitative assessments for farmland extraction tasks.

1. Introduction

Farmland is a vital monitoring target in remote sensing, closely linked to societal and economic development through its quantity, quality, and spatial distribution. As urbanization and industrialization accelerate in China, farmland resources face increasing encroachment. Rapid and accurate extraction of farmland information is therefore essential to support sustainable agricultural development and ensure national food security (Hong et al., 2024; Sun et al., 2022).

Farmland extraction from multi-source remote sensing data is a fundamental topic in remote sensing image interpretation. As image resolution has advanced, extraction methods have evolved significantly and can be broadly classified into four types: multi-feature-based methods, traditional machine learning methods, object-based methods, and deep learning methods. Multi-feature-based methods rely on manually crafted rules and low-level features like spectral, texture, and shape information for farmland extraction. While these methods offer a foundational approach, they struggle in complex backgrounds and under varying lighting conditions. Traditional machine learning methods (Jia et al., 2019), including Support Vector Machines (SVM), Extreme Learning Machines (ELM), Decision Trees (DT), and Random Forests (RF), automate feature selection to a degree by incorporating multi-dimensional remote sensing features (Sun et al., 2022; Zhang et al., 2023). However, they still depend on manual feature design and face limitations with high-dimensional and complex data. Object-based methods segment images into distinct semantic object regions, allowing feature classification at the object level. This approach reduces the “salt-and-pepper effect” common in pixel-level extraction and performs well on medium- to high-resolution imagery. Yet, challenges remain for object-based classification in very-high-resolution (VHR) imagery due to sensor diversity, imaging environment variability, complex scene targets, and dispersed farmland changes. In recent years, deep learning methods have revolutionized farmland extraction, progressing from Convolutional Neural Networks (CNNs) to advanced architectures like Transformer and Mamba, significantly enhancing accuracy (Yan et al., 2024). Public datasets such as GID (Tong et al., 2020), URUR (Ji et al., 2023), BLU (Ding et al., 2021), and LoveDA (Wang et al., 2021) have further accelerated deep learning advancements. However, current deep semantic segmentation networks often lose edge details in large-scale farmland extraction tasks, impacting overall accuracy. Farmland boundaries tend to be irregular, and narrow ridges particularly in southern regions appear blurred in imagery, making them difficult to detect. High-level semantic feature learning in deep networks can lead to spatial detail loss, hindering the recovery of geometric features during upsampling and reducing edge extraction accuracy. Few deep models are explicitly designed for farmland extraction, and most still rely on general open-source computer vision models, underscoring the need for further improvements in extraction accuracy tailored to this domain.

With significant breakthroughs in visual foundational models, particularly the Segment Anything Model (SAM), in semantic segmentation tasks for remote sensing imagery, the field is undergoing a revolutionary transformation. These changes are primarily reflected in aspects such as zero-shot generalization capability, efficient extraction of object boundaries and detail capture, as well as advancements in unsupervised and weakly supervised learning. SAM, trained on the SA-1B natural image dataset with over a billion images,

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-G-2025-253-2025

typically requires manually selected prompts (such as points, rectangular boxes, or segmentation masks) to guide the segmentation of target objects (Kirillov et al., 2023; Lin et al., 2023; Zhang et al., 2024). Given the challenges of recognizing fuzzy field boundaries in existing farmland extraction research and SAM's advantages in generalization and boundary sensitivity, an interesting question arises: Can SAM help accurately identify farmland areas in remote sensing images? However, the challenges in applying SAM include its limited consideration of the characteristics of farmland in remote sensing imagery, such as varying shapes and sizes of farmland, and changes in seasons and terrain, making it difficult to accurately identify and segment farmland areas. Additionally, SAM may underperform when handling remote sensing images with different resolutions and scales, affecting the accuracy and robustness of farmland extraction results. Moreover, due to the lack of VHR remote sensing images in SAM's training dataset, its generalization capability in actual farmland extraction tasks is insufficient. A common method to improve SAM's performance in farmland extraction is full-parameter fine-tuning, but this requires significant training time and computational resources, especially for large-scale visual models. To address these challenges, some research has proposed lightweight alternatives, usually by freezing SAM's original parameters and only training a small number of additional network parameters. Based on this, we propose a farmland extraction network named FL-DBENet, which features a general-specialized dual-branch architecture designed to better adapt SAM for farmland extraction tasks. Our contributions are as follows:

First, FL-DBENet adopts a dual-branch encoder network architecture. In the general branch, it utilizes SAM's powerful edge detection capabilities to learn finer farmland boundaries, while in the specialized branch, it employs the lightweight semantic segmentation model SegFormer’s image encoder (Xie et al., 2021) to provide SAM with more farmland-specific prompts.

Secondly, to reduce the number of trainable parameters, we introduce a Low-Rank Adaptation (LoRA) module (Hu et al., 2021) into SAM's image encoder, effectively lowering the computational cost. Additionally, we design a prompt mixer module to integrate different features. Thanks to the limited number of trainable parameters, the FL-DBENet algorithm significantly reduces computational resource requirements and improves training efficiency.

Finally, comprehensive ablation experiments demonstrate that FL-DBENet outperforms existing farmland extraction models without significantly increasing computational costs.

The structure of this paper is as follows: Section 2 provides a detailed introduction to the FL-DBENet farmland extraction algorithm and its experimental details; Section 3 presents the experimental results; Section 4 discusses the findings; and Section 5 concludes with the main conclusions of this paper.

2. Methodology

2.1 FL-DBENet framework

Farmland extraction falls under the category of semantic segmentation, which refers to separating objects from the background in an image to help computers better understand its content. Suppose we have a satellite image $\mathbf{X} \in \mathbb{R}^{H\times W\times B}$, with dimensions $H\times W$, containing C bands. The task of farmland extraction from the satellite image is to design a method f to identify the farmland areas U (if they exist) in the image and represent it as $\mathrm{U}: f(\mathbf{X}) \rightarrow \mathrm{U}$. Figure 1 presents the proposed FL-DBENet framework, designed for farmland extraction from remote sensing images. This framework follows a general-specialized dual-branch architecture. In the general branch, for a given satellite image $\mathbf{X}$ as input, the image is fed into the image encoder Φ_{SAM} of SAM (with large pre-trained parameters), and the output is F_{SAM}. Leveraging SAM's powerful edge detection capabilities, it effectively captures subtle features of farmland boundaries, allowing for learning finer farmland contours and shapes, thereby improving overall segmentation accuracy. Meanwhile, in the specialized branch, a lightweight semantic segmentation model, SegFormer, is used as the image encoder. The image $\mathbf{X}$ is fed into SegFormer’s image encoder Φ_{Seg} (with small learnable parameters), and the output is F_{Seg}. SegFormer not only provides efficient feature extraction but also offers SAM more targeted farmland feature prompts, making it more precise and efficient in handling farmland-related tasks. The design of this general-specialized dual-branch architecture further optimizes the detail and accuracy of farmland detection. Additionally, to further reduce the number of trainable parameters, a LoRA module is introduced into SAM’s image encoder. LoRA approximates the original weight matrix by introducing low-rank matrices in specific layers, effectively reducing computational complexity and memory overhead while maintaining the model’s representational ability and accuracy. Finally, we designed a prompt mixer module Φ_{Mix} to integrate feature prompts from different levels, ensuring that multi-scale information is captured while enhancing the complementarity and robustness of feature representations. Thanks to these innovative designs, the FL-DBENet algorithm significantly reduces the number of trainable parameters and lowers the demand for computational resources while maintaining model performance.

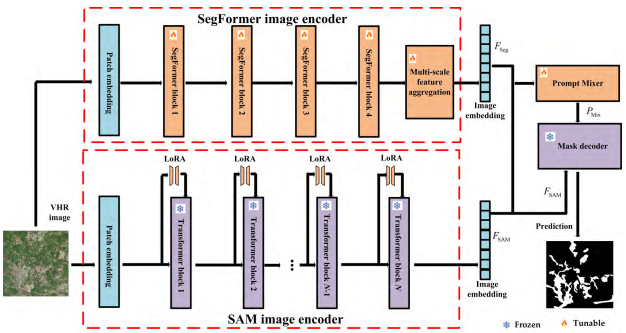

Figure 1. Pipeline of the proposed FL-DBENet framework.

2.1.1 SegFormer feature extraction

In this paper, the lightweight MiT-B0 backbone (Xie et al., 2021) is selected as the encoder for SegFormer. It employs a hierarchical vision transformer with a pyramid structure, designed to extract multi-level features at various scales. Specifically, at each stage l, the feature extraction through the Transformer module can be represented as:

$$F_l = \mathrm{TransformerStage}_l(F_{l-1}) \quad (1)$$

where the features extracted at each stage F_l have different resolutions, and F_{l-1} is the output of the previous stage, with $F_0 = \mathbf{X}$ representing the initial input image. SegFormer gradually extracts a multi-scale feature set $\{F_1, F_2, \cdots, F_l\}$ through its multi-layer Transformer encoder, where each feature

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-G-2025-253-2025 | © Author(s) 2025. CC BY 4.0 License.

captures image information at different scales. Since F_{Seg} contains multi-scale features, FL-DBENet applies a Φ_{Agg} fusion layer to aggregate these features, represented as:

$$F_{\text{Seg}} = \Phi_{\text{Agg}}(\Phi_{\text{Seg}}(\mathbf{X})) \tag{2}$$

2.1.2 SAM feature extraction

SAM, with its exceptional object segmentation capabilities, has demonstrated outstanding performance in land feature extraction from remote sensing images, significantly improving the accuracy and robustness of interpretation, especially in scenarios with complex feature structures and blurred boundaries. FL-DBENet inherits SAM's powerful segmentation performance and further optimizes the segmentation boundaries of farmland areas. In the FL-DBENet framework, all parameters of SAM's image encoder are frozen, and a trainable bypass channel is designed for each Transformer module. As shown in the LoRA module in Figure 2, these bypasses first compress the Transformer features into a low-rank space and then reproject them to align with the feature channels output by the frozen Transformer modules. Compared to fine-tuning all parameters in SAM, LoRA allows updating only a small number of parameters for the task of farmland extraction, which not only reduces computational costs but also minimizes the complexity of model deployment and storage during fine-tuning, while maintaining excellent segmentation performance.

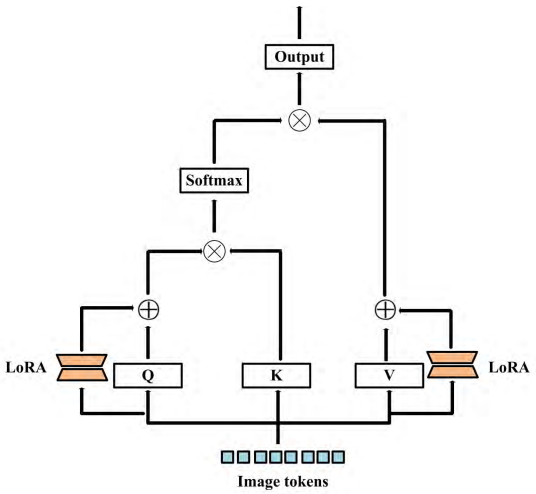

Figure 2. The LoRA design adopted in FL-DBENet. Adding LoRA to the *Q* layer and *V* layer of the transformer.

Assuming we are processing a given encoded token sequence $E \in \mathbb{R}^{B\times N\times C_{\text{in}}}$ and the output token sequence $\widehat{E} \in \mathbb{R}^{B\times N\times C_{\text{out}}}$ from the projection layer $W \in \mathbb{R}^{C_{\text{out}}\times C_{\text{in}}}$, LoRA posits that the updates to the projection layer *W* should be gradually stable, thus employing low-rank approximation to describe this progressive update process. Following this strategy, in the FL-DBENet framework, we first freeze the Transformer layers to keep *W* unchanged and then add a bypass to achieve the low-rank approximation. This bypass consists of two linear layers $L_A \in \mathbb{R}^{r\times C_{\text{in}}}$ and $L_B \in \mathbb{R}^{C_{\text{out}}\times r}$, where $r \ll \min\{C_{\text{in}}, C_{\text{out}}\}$. Therefore, the handling of the updated layer $\widehat{W}$ can be described as follows:

$$\begin{aligned} \widehat{E} &= \widehat{W}E \\ \widehat{W} &= W + \Delta W = W + L_B L_A \end{aligned} \tag{3}$$

Similarly, the processing strategy for multi-head self-attention will transform into the following steps:

$$\begin{aligned} \text{Attention}(Q,K,V) &= \text{Softmax}\left(\frac{QK^T}{\sqrt{C_{\text{out}}}} + B\right)V \\ Q &= \widehat{W}_q E = W_q E + L_{B_q} L_{A_q} E \\ K &= W_k E \\ V &= \widehat{W}_v E = W_v E + L_{B_v} L_{A_v} E \end{aligned} \tag{4}$$

where W_q, W_k, and W_v are frozen projection layer parameters from SAM, while L_{B_q}, L_{A_q}, L_{B_v} and L_{A_v} are trainable LoRA parameters (Wang et al., 2024). Finally, the satellite image **X** is input into the improved SAM image encoder, and the output features are F_{SAM}. The entire process can be represented as:

$$F_{\text{SAM}} = \Phi_{\text{SAM}}(\mathbf{X}) \tag{5}$$

2.1.3 Prompt generation and mask decoder

Prompt Generation. Since F_{Seg} and F_{SAM} both aggregate abstract semantic information specific to farmland areas in remote sensing images, FL-DBENet models both as semantic prompts. Specifically, we designed a hybrid module Φ_{Mix} to generate prompts by fusing these two types of prompts, which can be represented as:

$$P_{\text{Mix}} = \Phi_{\text{Mix}}(F_{\text{Seg}}, F_{\text{SAM}}) = \text{MLP}([F_{\text{Seg}}; F_{\text{SAM}}]) \tag{6}$$

In the above expression, Φ_{Mix} does not simply add features F_{Seg} and F_{SAM} together but concatenates them, placing them in a higher-dimensional space. Subsequently, the concatenated features pass through a projection layer, which reduces the dimensionality of the features, transforming the high-dimensional representation into a lower-dimensional form, thus achieving a more refined feature synthesis effect.

Mask Decoder. Finally, based on the generated hybrid prompt P_{Mix} and the pre-trained mask decoder from SAM, FL-DBENet identifies farmland areas in the remote sensing image, a process that can be represented as:

$$\text{U} = \Phi_{\text{SAM-Mask}}(F_{\text{SAM}}, P_{\text{Mix}}) \tag{7}$$

where elements in U indicate whether specific pixels belong to the farmland area.

2.2 Training loss

Similar to SAM, in the FL-DBENet model proposed in this paper, we employ a mask prediction strategy that combines three loss functions: focal loss L_{focal}, dice loss L_{dice}, and mean squared error loss L_{mse}, which are linearly combined with a weight ratio of 1:1:1. Focal loss L_{focal} addresses the imbalance between farmland samples and background regions, dice loss L_{dice} optimizes the segmentation boundaries of farmland areas, while mean squared error loss L_{mse} minimizes prediction errors. These three loss functions enhance the model's performance from different dimensions. Additionally, SegFormer uses cross-entropy loss L_{Seg} to strengthen the model's classification accuracy. The overall loss function is defined as:

$$\begin{aligned} L_{\text{SAM}} &= L_{\text{focal}} + L_{\text{dice}} + L_{\text{mse}} \\ L &= \lambda L_{\text{SAM}} + L_{\text{Seg}} \end{aligned} \tag{8}$$

where λ is a hyperparameter used to adjust the weights of the general and specialized modules during the training process. This design ensures the model's segmentation performance

while reducing computational costs and improving deployment efficiency.

3. Experimental analyses

3.1 Experimental datasets

To assess the FL-DBENet algorithm, we utilized two publicly available datasets, GID (Tong et al., 2020) and the ultra-high resolution with ultra-rich (URUR) context dataset (Ji et al., 2023). GID is a large-scale VHR remote sensing land cover dataset based on China's Gaofen-2 satellite data. The GID dataset is divided into two parts: a large-scale classification set (GID-5) and a fine land cover set (GID-15). GID-5 includes 5 land cover categories: built-up, farmland, forest, meadow, and water, consisting of 150 pixel-level labeled Gaofen-2 satellite remote sensing images. Among them, 105 images are used for training, 15 for validation, and 30 for testing. In this paper, we only used the RGB bands of the Gaofen-2 satellite remote sensing images. The original Gaofen-2 images have a size of 6800×7200 and were pixel-level annotated by experts in the field of remote sensing interpretation. We did not use the labels from the fine land cover set (GID-15), but instead used the labeled images from the GID-5 classification system, with farmland set as the foreground region and the other four categories set as background regions. The original images were divided into non-overlapping regions of size 1024×1024. As a result, the training set consisted of 4,410 images, the validation set of 630 images, and the test set of 1,260 images. Examples of the GID dataset are shown in Figure 3, where white represents farmland areas and black represents background areas.

The URUR dataset contains 3,008 VHR images, each sized 5120×5120, with 3 bands, covering a wide range of complex scenes (from 63 cities), and includes rich background diversity (1 million instances across 8 categories) with fine-grained annotations (around 80 billion manually annotated pixels). The training, validation, and test sets consist of 2,157, 280, and 571 VHR images, respectively, with an approximate ratio of 7:1:2. All images were manually annotated in detail, including pixel-level fine classification into 8 types: building, farmland, greenhouse, woodland, bareland, water, road, and others. We set farmland as the foreground region, and the other 7 categories as background regions. The original training and validation set images were resampled to 1024×1024 for model training. For model testing, the test set images were first resampled to 1024×1024 for inference, but the accuracy evaluation was still calculated based on the original 5120×5120 resolution. Examples of the URUR dataset are shown in Figure 4, where white represents farmland areas and black represents background areas.

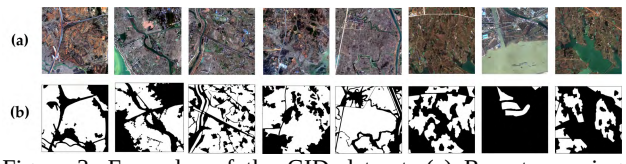

Figure 3. Examples of the GID dataset. (a) Remote sensing images. (b) Ground truth.

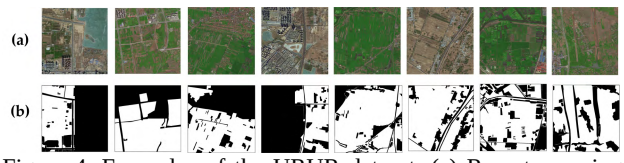

Figure 4. Examples of the URUR dataset. (a) Remote sensing images. (b) Ground truth.

3.2 Implementation details and evaluation metrics

We implemented FL-DBENet using the PyTorch framework and conducted experiments on a workstation equipped with a 12th Generation Intel Core i9-12900K @ 3.19 GHz processor, 64.00 GB RAM, and an NVIDIA GeForce RTX A6000 graphics card. In our experiments, we used the ViT-B backbone from SAM and the lightweight MiT-B0 encoder from SegFormer. The hyperparameter λ in the experiment was fixed at 0.1, and the AdamW optimizer and cosine annealing learning rate strategy were applied, with an initial learning rate of 0.0005. The batch size was set to 4, and the full training process spanned 200 epochs. During the experiments, we utilized LoRA to fine-tune the frozen Q and V projection layers of the Transformer blocks. The rank of LoRA was set to 4 to improve both efficiency and performance.

We compared the performance of our algorithm to state-of-the-art (SOTA) methods using four metrics: precision, recall, F1 score, and intersection over union (IoU). These metrics are defined as follows:

$$\text{precision} = \frac{TP}{TP + FP} \tag{9}$$

$$\text{recall} = \frac{TP}{TP + FN} \tag{10}$$

$$\text{F1} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{11}$$

$$\text{IoU} = \frac{TP}{TP + FP + FN} \tag{12}$$

where TP denotes true-positive values, TN denotes true-negative values, FP denotes false-positive values, and FN denotes false-negative values.

3.3 Comparison to SOTA approaches

In this paper, we conducted a comparative evaluation of our proposed method against several standard semantic segmentation algorithms. The comparison includes models with diverse backbone networks and architectures, such as advanced UNet variants like UNet++ (Peng et al., 2019), ResUNet (Zhang et al., 2018), and MMUU-Net (Gao et al., 2020), as well as widely recognized segmentation models, including PSPNet (Zhao et al., 2017), HRNetV2-W48 (Sun et al., 2019), and DeepLabv3+ (Chen et al., 2018). We also evaluated the SAM-Adapter method (Chen et al., 2023), which fine-tunes the SAM image encoder, and the SAMUS method (Lin et al., 2023), which incorporates a CNN-branch image encoder alongside a cross-branch attention module. This comparative evaluation aims to demonstrate the performance and robustness of our approach across a range of model architectures and segmentation frameworks.

3.4 Experimental results

3.4.1. Performance comparison for the GID dataset

To validate the effectiveness of the proposed FL-DBENet model for farmland extraction tasks on the GID dataset, we conducted a comparative analysis of its extraction results against recent SOTA farmland extraction models. Table 1 presents the performance of various models on the GID dataset for farmland extraction. As shown in Table 1, the FL-DBENet model achieved a Precision of 83.55%, Recall of 79.88%, F1 score of 81.67%, and IoU of 69.02%. Compared to other models, FL-DBENet improved the F1 score by 1.26%–7.37% and the IoU by 1.79%–9.91%. UNet++ had the highest Precision at 87.65%, followed by FL-DBENet, but UNet++ had

the lowest F1 score and IoU, which were 7.37% and 9.91% lower than FL-DBENet, respectively. ResUNet had the highest Recall at 82.55%, followed by FL-DBENet at 79.88%, but ResUNet’s F1 score and IoU were 4.74% and 6.51% lower than FL-DBENet, respectively. Both SAM-Adapter and SAMUS models are based on efficient parameter fine-tuning of SAM and performed better than traditional CNN-based SOTA methods on farmland extraction in the GID dataset. In contrast, FL-DBENet combines the strengths of SAM and SegFormer, demonstrating improvements across all metrics. Compared to SAM-Adapter and SAMUS, FL-DBENet improved Precision by 1.36% and 2.3%, Recall by 1.74% and 0.3%, F1 score by 1.56% and 1.26%, and IoU by 2.19% and 1.79%, respectively. To visually illustrate the results, Figure 5 shows the partial extraction results of farmland from the GID dataset using these nine methods. Overall, the FL-DBENet model produced clear farmland boundaries that closely matched the actual edges. As shown in Figure 5, the FL-DBENet model exhibited fewer holes and noise in the farmland regions, with no missing large plots and sharp edges. Small plots had clear contours with minimal deformation. In complex edges and gap regions, FL-DBENet showed significant improvement over other methods, particularly in edge extraction, effectively reducing jagged edges and holes in the farmland areas. Based on the quantitative and qualitative analysis results, the proposed FL-DBENet model demonstrates significant advantages in addressing boundary blurring issues in farmland extraction, providing an effective technical reference for further optimization of farmland extraction techniques.

Methods	Backbone	GID			
		Precision	Recall	F1	IoU
UNet++	Res-50	**87.65**	64.48	74.30	59.11
ResUNet	Res-50	72.03	**82.55**	76.93	62.51
MMUU-Net	Res-50	80.39	75.46	77.85	63.73
PSPNet	Res-50	81.42	75.43	78.31	64.35
DeepLabv3+	Res-50	**83.38**	75.18	79.07	65.38
HRNetV2-W48	HRNet	82.54	76.28	79.28	65.68
SAM-Adapter	ViT-B	82.19	78.14	**80.11**	**66.83**
SAMUS	ViT-B	81.25	**79.58**	**80.41**	**67.23**
FL-DBENet	ViT-B and MiT-B0	**83.55**	**79.88**	**81.67**	**69.02**

Table 1. Performance comparison for the GID dataset. All values are percentages. Bold red text indicates highest, bold blue text indicates second-highest, and bold black text indicates third-highest performances.

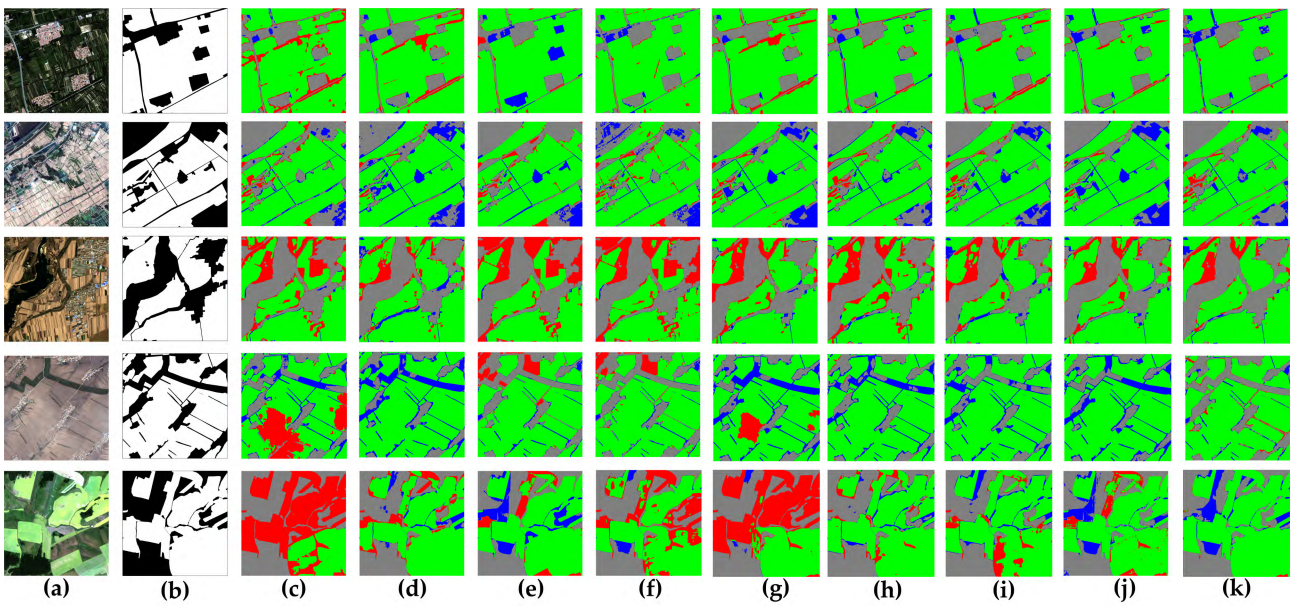

Figure 5. Visual comparisons of the different SOTA models applied to the GID dataset. (a)Input images; (b)Ground truths; (c)UNet++; (d)ResUNet; (e)MMUU-Net; (f)PSPNet; (g)DeepLabv3+; (h)HRNetV2-W48; (i)SAM-Adapter; (j)SAMUS; (k)FL-DBENet. Grey: *TN* pixels; Green: *TP* pixels; Blue: *FP* pixels; Red: *FN* pixels.

3.4.2. Performance Comparison for the URUR dataset

To quantitatively evaluate the effectiveness of the proposed method, we tested models including UNet++, ResUNet, MMUU-Net, PSPNet, DeepLabv3+, HRNetV2-W48, SAM-Adapter, SAMUS, and FL-DBENet on the URUR dataset, and measured their Precision, Recall, F1 score, and IoU, as shown in Table 2. The proposed FL-DBENet model achieved the highest scores in Recall, F1 score, and IoU, with values of 90.30%, 88.28%, and 79.02%, respectively. Although the Precision of FL-DBENet was slightly lower than that of SAM-Adapter and SAMUS, it was still significantly better than traditional CNN-based models. Compared to other models, FL-DBENet improved Recall by 5.1%–8.52%, F1 score by 2.71%–4.62%, and IoU by 4.24%–7.11%. SAMUS achieved the highest Precision at 87.59%, followed by SAM-Adapter at 86.74%, but their Recall values were 8.52% and 5.87% lower than FL-DBENet’s, respectively; their F1 scores were 3.7% and 2.71% lower; and their IoU values were 5.73% and 4.24% lower, respectively. The quantitative analysis shows that FL-

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-G-2025-253-2025 | © Author(s) 2025. CC BY 4.0 License.

DBENet exhibits superior overall performance. Figure 6 displays some visualized results of the nine methods in extracting farmland plots from the URUR dataset. As shown in Figure 6, the extraction results of FL-DBENet feature clear boundaries and high separation from non-farmland areas, effectively distinguishing farmland boundaries with fewer misclassifications and omissions. Although SAM-Adapter also performed well in farmland identification, its boundary handling was slightly rougher compared to FL-DBENet, with more fragmentation. Especially in complex backgrounds, FL-DBENet outperformed traditional CNN-based networks in farmland extraction, with clearer boundary definition, showcasing its stronger capability in boundary recognition. The quantitative and qualitative experimental results further validate the superiority of the proposed FL-DBENet model.

Methods	Backbone	URUR			
		Precision	Recall	F1	IoU
UNet++	Res-50	84.13	84.35	84.24	72.77
ResUNet	Res-50	83.03	84.89	83.95	72.34
MMUU-Net	Res-50	85.19	82.19	83.66	71.91
PSPNet	Res-50	83.54	84.39	83.96	72.36
DeepLabv3+	Res-50	83.04	**85.20**	84.11	72.57
HRNetV2-W48	HRNet	84.12	**84.99**	84.56	73.17
SAM-Adapter	ViT-B	**86.74**	84.43	**85.57**	**74.78**
SAMUS	ViT-B	**87.59**	81.78	**84.58**	**73.29**
FL-DBENet	ViT-B and MiT-B0	**86.35**	**90.30**	**88.28**	**79.02**

Table 2. Performance comparison for the URUR dataset. All values are percentages. Bold red text indicates highest, bold blue text indicates second-highest, and bold black text indicates third-highest performances.

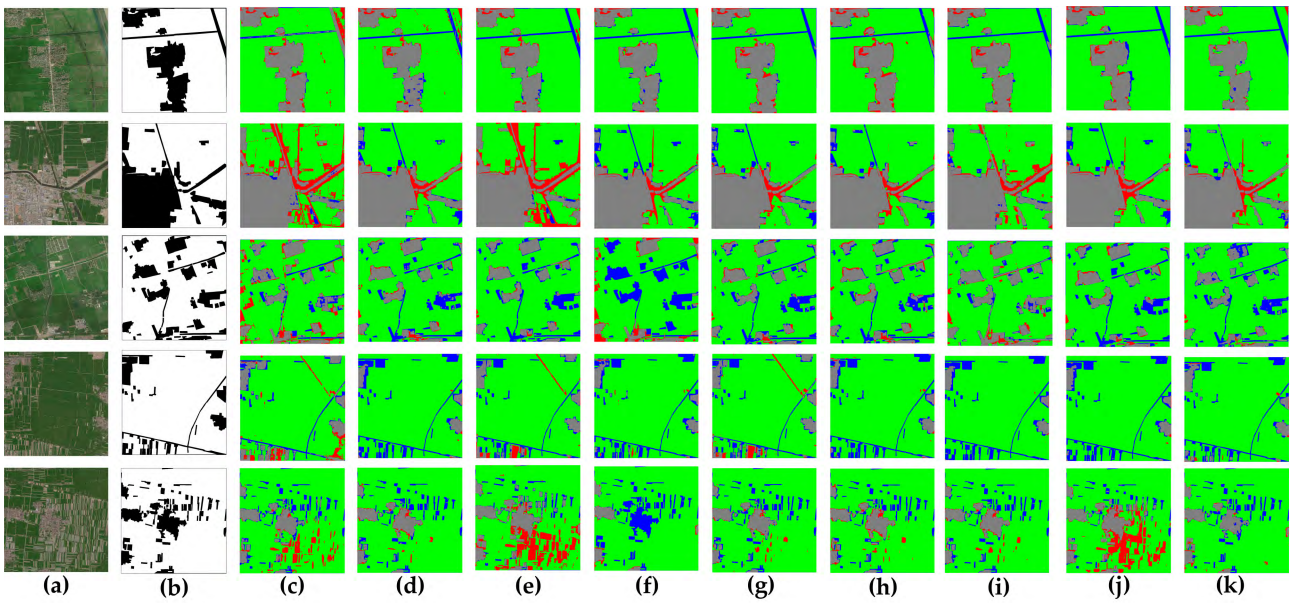

Figure 6. Visual comparisons of the different SOTA models applied to the URUR dataset. (a)Input images; (b)Ground truths; (c)UNet++; (d)ResUNet; (e)MMUU-Net; (f)PSPNet; (g)DeepLabv3+; (h)HRNetV2-W48; (i)SAM-Adapter; (j)SAMUS; (k)FL-DBENet. Grey: TN pixels; Green: TP pixels; Blue: FP pixels; Red: FN pixels.

3.4.3. Ablation experiment

To evaluate the effectiveness of the main components in the FL-DBENet network, we conducted ablation studies on the GID and URUR datasets. Specifically, we experimented with two key components of the network (i.e., the SAM image encoder, the SegFormer image encoder) and their combination. Detailed results of the ablation study are shown in Table 3. As seen in Table 2, each component in the FL-DBENet network contributes positively to the overall results. When only the SAM image encoder is used, fine-tuned with the LoRA component, the model achieves an F1 score of 77.98% and an IoU of 63.91% on the GID dataset; on the URUR dataset, it achieves an F1 score of 84.54% and an IoU of 73.12%. When only the SegFormer image encoder is used, the model achieves an F1 score of 77.71% and an IoU of 63.55% on the GID dataset; on the URUR dataset, it achieves an F1 score of 83.70% and an IoU of 71.97%. When both components are combined, the model's F1 score on the GID dataset improves by 3.69% to 3.96%, and IoU increases by 5.11% to 5.47%; on the URUR dataset, the F1 score improves by 3.74% to 4.58%, and IoU increases by 5.90% to 7.05%. The experimental results further confirm the necessity of the general-specialized dual-branch encoder for feature extraction.

https://doi.org/10.5194/isprs-annals-X-G-2025-253-2025

SAM	SegFormer	GID		URUR	
		F1	IoU	F1	IoU
✓	×	77.98	63.91	84.54	73.12
×	✓	77.71	63.55	83.70	71.97
✓	✓	**81.67**	**69.02**	**88.28**	**79.02**

Table 3. Ablation experiment of FL-DBENet network with and without the use of SAM image encoder and SegFormer image encoder. All values are percentages. × indicates excluded steps during the training process, while ✓ denotes their inclusion. Bold black text indicates highest performances.

4. Discussion

In recent years, foundational models like BERT (Koroteev et al., 2021), GPT (Achiam et al., 2023), CLIP (Radford et al., 2021), and SAM have emerged, showcasing remarkable performance in natural language processing and computer vision, and often surpassing traditional deep learning networks. Pre-trained on large-scale datasets, these models learn rich feature representations with strong generalization capabilities across various downstream tasks, significantly reducing dependency on extensive labeled data. Building on SAM’s success in image segmentation, this paper explores its application to farmland extraction from remote sensing images. To this end, we propose FL-DBENet, a dual-branch architecture specifically designed for farmland extraction in remote sensing data, and evaluate its performance on the GID and URUR datasets, comparing it to several other SOTA methods. The advantages of FL-DBENet are mainly reflected in two key areas:

1) Enhanced Semantic Understanding with Transformer-Based Encoders: As a Transformer-based model, FL-DBENet demonstrates notable advantages in F1-score and IoU metrics compared to traditional CNN models. While structures like buildings or roads generally have clear boundaries, farmland often features blurred edges with irregular and narrow field ridges. Leveraging the fine-grained feature capturing of the Transformer’s self-attention mechanism, FL-DBENet more accurately detects these subtle boundaries, whereas CNN-based models tend to capture broader semantic abstractions, resulting in less precise delineation.

2) Improved Spatial Feature Extraction through a Dual-Branch Architecture: FL-DBENet combines SAM’s image encoder with SegFormer’s multi-scale feature extraction capabilities, enabling robust handling of farmland parcels with varying shapes and scales. This general-specialized approach enhances FL-DBENet’s adaptability across diverse image datasets, with experiments showing superior performance over traditional SAM fine-tuning methods, particularly in complex farmland extraction scenarios.

Despite its excellent performance in farmland extraction on the GID and URUR datasets, FL-DBENet has some limitations. For other land types with features similar to farmland, its multi-scale feature extraction may offer limited performance gains, suggesting a need for further optimization for specific datasets. Additionally, FL-DBENet’s effectiveness is influenced by dataset diversity, and future research could further assess its capabilities across a broader range of datasets.

5. CONCLUSIONS

As artificial intelligence advances into VHR remote sensing interpretation, farmland extraction from such images is increasingly leveraging deep learning techniques, yielding substantial advancements alongside new challenges. Addressing these challenges, this paper introduces a specialized farmland extraction network, FL-DBENet, featuring a dual-branch encoder architecture that optimizes the foundational visual model SAM for farmland extraction. The universal-specialized architecture of FL-DBENet capitalizes on SAM’s powerful edge-detection capabilities in the universal branch, enhancing the precision of farmland boundary delineation. Simultaneously, the specialized branch integrates a lightweight SegFormer encoder to deliver farmland-specific features, refining SAM’s interpretation. Additionally, the integration of a LoRA module within SAM’s image encoder reduces computational overhead, while a prompt mixer module facilitates the effective synthesis of diverse feature representations. With its reduced trainable parameters, FL-DBENet not only minimizes computational demands but also significantly enhances training efficiency. Experimental evaluations on the GID and URUR datasets reveal that FL-DBENet achieves superior performance compared to existing state-of-the-art models in farmland extraction, all without incurring significant computational costs. This efficiency and precision highlight FL-DBENet’s potential as an effective solution for intelligent farmland extraction in VHR remote sensing applications.

Aknowledgements

This work was supported by the National Natural Science Foundation of China under Grant Nos. 42101358.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S. and Avila, R., 2023. Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

Chen, T., Zhu, L., Deng, C., Cao, R., Wang, Y., Zhang, S., Li, Z., Sun, L., Zang, Y. and Mao, P., 2023. Sam-adapter: Adapting segment anything in underperformed scenes. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 3367-3375).

Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.

Ding, L., Lin, D., Lin, S., Zhang, J., Cui, X., Wang, Y., Tang, H. and Bruzzone, L., 2021. Looking outside the window: Wide-context transformer for the semantic segmentation of high-resolution remote sensing images. arXiv preprint arXiv:2106.15754.

Gao, X., Liu, L. and Gong, H., 2020. MMUU-Net: A robust and effective network for farmland segmentation of satellite imagery. In Journal of physics: conference series (Vol. 1651, No. 1, p. 012189). IOP Publishing.

Hong, Q., Zhu, Y., Liu, W., Ren, T., Shi, C., Lu, Z., Yang, Y., Deng, R., Qian, J. and Tan, C., 2024. A segmentation network for farmland ridge based on encoder-decoder architecture in combined with strip pooling module and ASPP. *Frontiers in Plant Science*, 15, p.1328075.

Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. 2021. LoRA: Low-rank adaptation of large language models. arXiv 2021, arXiv:2106.09685.

Ji, D., Zhao, F., Lu, H., Tao, M. and Ye, J., 2023. Ultra-high resolution segmentation with ultra-rich context: A novel benchmark. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 23621-23630).

Jia, X., Khandelwal, A. and Kumar, V., 2019. Automated Monitoring Cropland Using Remote Sensing Data: Challenges and Opportunities for Machine Learning. arXiv preprint arXiv:1904.04329.

Koroteev, M.V., 2021. BERT: a review of applications in natural language processing and understanding. arXiv preprint arXiv:2103.11943.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.-Y., Dollar, P., Girshick, R. 2023. Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4015-4026.

Lin, X., Xiang, Y., Zhang, L., Yang, X., Yan, Z. and Yu, L., 2023. SAMUS: Adapting segment anything model for clinically-friendly and generalizable ultrasound image segmentation. arXiv preprint arXiv:2309.06824.

Peng, D.; Zhang, Y.; Guan, H. 2019. End-to-End Change Detection for High Resolution Satellite Images Using Improved UNet++. *Remote Sensing*, 11, 1382.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. and Krueger, G., 2021, July. Learning transferable visual models from natural language supervision. In International conference on machine learning (pp. 8748-8763). PMLR.

Sun, W., Sheng, W., Zhou, R., Zhu, Y., Chen, A., Zhao, S. and Zhang, Q., 2022. Deep edge enhancement-based semantic segmentation network for farmland segmentation with satellite imagery. *Computers and Electronics in Agriculture*, 202, p.107273.

Sun, K.; Zhao, Y.; Jiang, B.; Cheng, T.; Wang, J. 2019. High-resolution representations for labeling pixels and regions. arXiv 2019, arXiv:1904.04514.

Tong, X.Y., Xia, G.S., Lu, Q., Shen, H., Li, S., You, S. and Zhang, L., 2020. Land-cover classification with high-resolution remote sensing images using transferable deep models. *Remote Sensing of Environment*, 237, p.111322.

Wang, J., Zheng, Z., Ma, A., Lu, X. and Zhong, Y., 2021. LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation. arXiv preprint arXiv:2110.08733.

Wang, Y., Zhang, W., Chen, W. and Chen, C., 2024. BSDSNet: Dual-Stream Feature Extraction Network Based on Segment Anything Model for Synthetic Aperture Radar Land Cover Classification. Remote Sensing, 16(7), p.1150.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M. and Luo, P., 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. Advances in neural information processing systems, 34, pp.12077-12090.

Yan, S., Yao, X., Sun, J., Huang, W., Yang, L., Zhang, C., Gao, B., Yang, J., Yun, W. and Zhu, D., 2024. TSANet: A deep learning framework for the delineation of agricultural fields utilizing satellite image time series. *Computers and Electronics in Agriculture*, 220, p.108902.

Zhao, H.; Shi, J.; Qi, X.;Wang, X.; Jia, J. 2017. Pyramid scene parsing network. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.

Zhang, Z.; Liu, Q.; Wang, Y. 2018. Road extraction by deep residual u-net. *IEEE Geosci. Remote Sens. Lett*. 2018, 15, 749–753

Zhang, X., Huang, J., Ning T. 2023. Progress and Prospect of Cultivated Land Extraction from High Resolution Remote Sensing Images. *Geomatics and Information Science of Wuhan University*, 2023, DOI: 10.13203/j. whugis20230114.

Zhang, X., Liu, Y., Lin, Y., Liao, Q. and Li, Y., 2024, March. Uv-sam: Adapting segment anything model for urban village identification. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 20, pp. 22520-22528).