ITC 1/55
Information Technology and Control
Vol. 55 / No. 1/ 2026
pp. 65-85
DOI 10.5755/j01.itc.55.1.42839

Extraction of Film-Mulched Tobacco Fields and Estimation of Tobacco Planting Area Based on Deep Learning and High-Resolution Remote Sensing Images

Received 2025/09/20 | Accepted after revision 2025/12/21

HOW TO CITE: Huang, F., Gao, R., Wang, L., Zhao, Q., Chi, G. (2026). Extraction of Film-Mulched Tobacco Fields and Estimation of Tobacco Planting Area Based on Deep Learning and High-Resolution Remote Sensing Images. *Information Technology and Control*, 55(1), 65-85. https://doi.org/10.5755/j01.itc.55.1.42839

Extraction of Film-Mulched Tobacco Fields and Estimation of Tobacco Planting Area Based on Deep Learning and High-Resolution Remote Sensing Images

Fenghua Huang*

Fujian Key Laboratory of spatial information perception and intelligent processing, Yango University, 350015 Fuzhou, China; e-mail: fhhuang@ygu.edu.cn (F.H.)

Fujian University Engineering Research Center of Spatial Data Mining and Application, Yango University, 350015 Fuzhou, China; e-mail: fhhuang@ygu.edu.cn (F.H.)

College of Artificial Intelligence, Yango University, Fuzhou 350015, China; 17850053605@163.com (L.W.)

Ronggang Gao

Academy of Digital China (Fujian), Fuzhou University, Fuzhou 350108, China; e-mails: 451334901@qq.com (R.G.), 245527004@fzu.edu.cn (Q.Z.)

Lin Wang

College of Artificial Intelligence, Yango University, Fuzhou 350015, China; 17850053605@163.com (L.W.)

Qianyu Zhao

Academy of Digital China (Fujian), Fuzhou University, Fuzhou 350108, China; e-mails: 451334901@qq.com (R.G.), 245527004@fzu.edu.cn (Q.Z.)e-mail: ellee@utem.edu.my

Guosheng Chi

Guangze Branch of Nanping Tobacco Company, Nanping 354100, China; e-mail: 18039788618@139.com (G.C.)

Corresponding author: fhhuang@ygu.edu.cn

In tobacco cultivation, the use of plastic film mulch in tobacco fields serves as a reliable indicator of farmers' intent to transplant seedlings from greenhouses to the fields. An important research challenge is how to efficiently and accurately identify film-mulched tobacco fields over large areas using high-resolution satellite remote sensing imagery prior to the transplanting stage. Such identification can support the estimation of actual tobacco planting area, thereby assisting tobacco management authorities in evaluating the fulfillment of macro-level planting targets and formulating regulatory policies. To address the limitations of conventional manual and machine learning methods, such as low efficiency and insufficient accuracy in extracting tobacco field boundaries from high-resolution remote sensing images, this study proposes an approach based on the Skip_Segformer semantic segmentation model. Specifically, a SKIPAT module was integrated into the encoder of the traditional SegFormer model to reduce the number of training parameters and save computational resources. Additionally, the decoder was enhanced with a Multi-level Feature Fusion (MFF) mechanism to better integrate features across different scales, thereby significantly improving the accuracy of film-mulched field boundary extraction. The experiment was conducted in Siqian Township, Guangze County, Nanping City, using Jilin-1 satellite imagery with a spatial resolution of 0.5 meters, acquired on March 8, 2022. At both full-regional and local scales, the Skip_Segformer model was compared with four other networks (DeeplabV3+, Hrnet, Pspnet, and SegFormer) in extracting film-mulched tobacco field patches. The results were compared to identify the optimal model. Experimental results demonstrate that the Skip_Segformer achieved the highest extraction accuracy and generalization capability among the compared models. It attained an extraction accuracy of 97% across Siqian Township, with a relative error of only 1.6% in the estimated tobacco planting area, significantly outperforming the other four models. The proposed method shows strong feasibility and applicability for large-scale extraction of film-covered tobacco fields and estimation of planting area, effectively supporting tobacco administration departments in total planting area monitoring and providing a basis for local tobacco planting planning.

KEYWORDS: Deep learning, Semantic Segmentation, Film-mulched Tobacco Fields Extraction, Skip_Sgformer network model

1. Introduction

Tobacco is one of the most important cash crops and a major agricultural product for export in China. The tobacco industry plays a significant role in increasing tax revenue for the country. Under China's strict tobacco monopoly system, local specialized administrative agencies coordinate and plan the cultivation, processing, and purchasing of tobacco. Tobacco planting area, which is directly correlated with yield, serves as a key indicator for macro-level management within the industry. Accurate monitoring of planting area helps relevant departments to control the total amount of tobacco production, thereby avoiding regulatory failure and market disorder. However, traditional methods for estimating crop planting area, such as manual regional surveys and hierarchical statistical reporting, are not only labor-intensive and time-consuming but also prone to underreporting, misreporting, and false reporting due to human interference. These approaches often fail to provide accurate and timely data to support scientific decision-making by tobacco administrative agencies, which can lead to issues such as excessive planting leading to regulatory violations or insufficient planting area causing waste of allocated quotas.

Moreover, the tobacco planting process generally includes stages such as greenhouse seedling cultivation, ridge formation in fields, plastic film mulching, transplanting, fertilization and maintenance, leaf picking, curing, and processing. Among these, plastic film mulching reflects the actual intent of farmers to transplant and grow tobacco. Estimating the area of film-mulched fields before transplanting can provide a reliable indication of the final actual planting area. Nevertheless, due to the extensive and widespread nature of tobacco cultivation, accurately extracting film-mulched fields over large regions and estimating their area remains a significant challenge. Satellite remote sensing imagery, known for its macroscopic, timely,

and objective characteristics, has been widely applied in agricultural statistics and surveys, agricultural zoning, land resource and land use research, crop growth monitoring, and yield estimation. In particular, high-resolution satellite imagery has become a major data source for estimating crop planting areas. Since film-mulched tobacco fields exhibit distinct spatial texture features, intelligent interpretation of high-resolution satellite imagery (with resolution ≤ 0.5 meters) can effectively extract field boundaries and covered areas, thereby supporting accurate estimation of tobacco planting area in the study region.

Compared to traditional machine learning methods (e.g., support vector machines, random forests), deep learning models which have advanced rapidly in recent years possess stronger feature learning capabilities and the ability to fit complex nonlinear functions. These models can classify each pixel in an image into specific semantic categories, enabling fine-grained image segmentation and understanding, and substantially improving the performance of semantic segmentation. Numerous studies in recent years have employed various deep network architectures for image semantic segmentation. Representative models include DeepLabv3 [1], HRNet (High-Resolution Network) [2], PSPNet (Pyramid Scene Parsing Network) [16], SETR (Segmentation Transformer) [3,24], T2T-ViT (Tokens-to-Token Vision Transformer) [22], SegFormer [21], and U-Net [10]. Among these, SETR, T2T-ViT, and SegFormer are Transformer-based models that offer certain advantages in image semantic segmentation tasks, such as capturing both global and local contextual information and performing accurate pixel-level semantic segmentation, making them well-suited for remote sensing image analysis. In the specific field of semantic segmentation of tobacco field remote sensing imagery, which is directly related to this study, the aforementioned deep learning models have also been applied. Notable examples include: Zhang et al. [23], who used GF-2 satellite imagery to create a small-sample tobacco dataset and employed U-Net for tobacco extraction, achieving over 3% improvement in accuracy and recall compared to other machine learning methods; Tian et al. [18], who utilized fused panchromatic and multispectral imagery from GF-1 and applied U-Net, PSPNet, and DeepLabv3+ for crop classification, providing a reference for accurately obtaining crop types, area, and spatial distribution in regions with complex planting structures; and Fu et al. [4], who proposed a method based on the DeeplabV3+ deep semantic segmentation model for precise extraction of tobacco planting area from UAV remote sensing imagery. By replacing the original atrous convolution structure with four classical lightweight backbone networks, they generally improved the accuracy of tobacco field semantic segmentation.

However, the accuracy and efficiency of existing models applied to semantic segmentation of tobacco remote sensing imagery still require further improvement. Moreover, there is relatively little research, both domestically and internationally, on directly using deep learning for extracting film-mulched tobacco fields and estimating planting area. In light of the above, this study aims to improve the SegFormer network by incorporating a SKIPAT module [19] and a Multi-level Feature Fusion (MFF) module, proposing a Skip_Segformer-based semantic segmentation model combined with high-resolution remote sensing imagery to achieve accurate extraction of film-mulched tobacco fields and high-precision estimation of actual tobacco planting area.

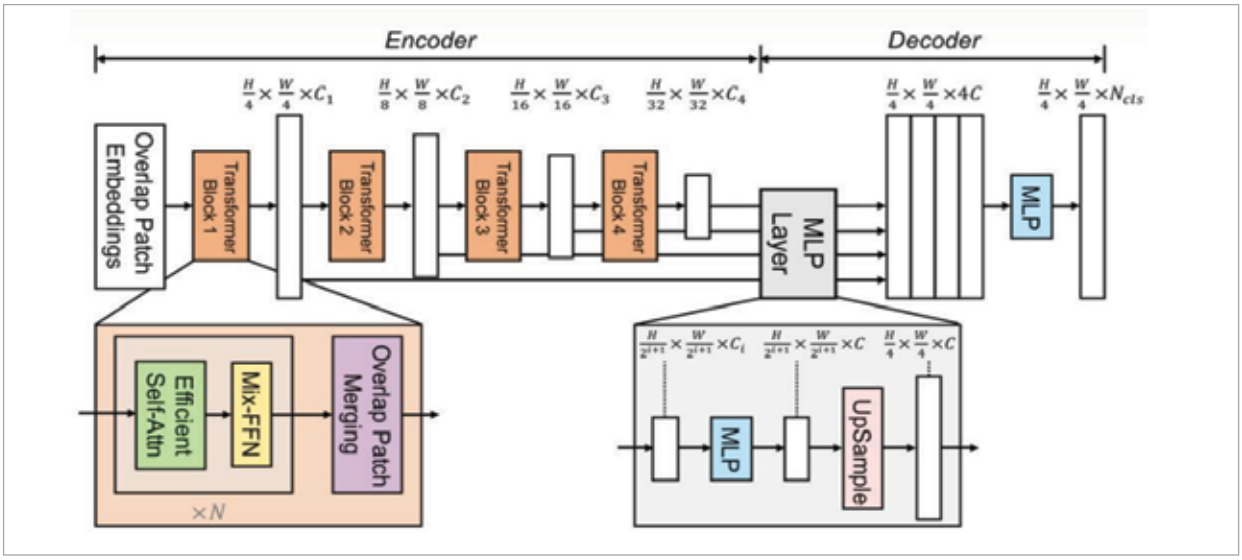
2. Methods

This paper proposes a method for extracting plastic-film-covered tobacco fields based on the Skip_Segformer semantic segmentation model and high-resolution remote sensing imagery. The method can improve the accuracy of boundary extraction while reducing computational complexity, thereby enhancing the effectiveness of field identification and the precision of planting area estimation.

2.1. SegFormer Network Architecture

SegFormer is a convolutional neural network architecture designed for image segmentation tasks. It integrates Transformer and convolutional operations, offering high computational efficiency and strong modeling capability. The network structure is illustrated in Figure 1 [21].

Figure 1
Structure of the SegFormer network model.



The SegFormer model consists of two main components: an encoder and a decoder. The encoder incorporates a Transformer architecture adapted from Vision Transformer (ViT) for segmentation tasks. It replaces the standard patch embeddings with an Overlap Patch Embedding (OPE) structure to extract and downsample features from input images. The encoder also includes multiple Efficient Multi-head Self-Attention (EMSA) layers and Mix Feed Forward (MixFFN) layers to capture rich detailed and semantic features. The decoder employs a multi-level feature fusion mechanism. It processes feature maps output from the encoder's four stages (with resolutions of 1/4, 1/8, 1/16, and 1/32 of the original image size) using a simple Multilayer Perceptron (MLP) and convolutional operations for upsampling and fusion. The final output is a high-resolution feature representation and segmentation result. The encoder used in this study is an improved version based on MiT-B0. The values of key hyperparameters of the encoder in SegFormer are listed in Table 1.

The meanings of the hyperparameter metrics in Table 1 are as follows:

- **Embed_dims:** The encoding length per feature point.
- **Num_layers:** The number of repetitions of EMSA and MixFFN in the TransformerBlocks across the four stages.
- **Num_heads:** The number of heads in the EMSA module at each stage. The product of Num_heads and Embed_dims gives the output channel number for each stage: 64, 128, 320, and 512, respectively.
- **Patch_sizes:** Kernel sizes of the convolutional layers in OPE at each stage.
- **Strides:** Sampling stride of OPE.
- **Sr_ratios:** Downsampling ratios for (K, V) input in each stage.
- **Mlp_ratio:** When multiplied by Embed_dims, it yields the elevated channel dimension in MixFFN, which is 256 across the all four stages.

Table 1
The values of key hyperparameters of the encoder in SegFormer.

Hyperparameter items	Item Values
Embed_dims	64
Num_layers	[3, 4, 18, 3]
Num_heads	[1, 2, 5, 8]
Patch_size	[7, 3, 3, 3]
Strides	[4, 2, 2, 2]
Sr_ratios	[8, 4, 2, 1]
Mlp_ratio	4
Skip_layer	2

2.2. SKIPAT Block

SKIPAT is a lightweight framework that can be integrated into various Transformer architectures to improve the computational efficiency of Vision Transformers (ViTs) [7, 11]. Depending on the specific model structure, SKIPAT can skip the Multi-Head Self-Attention (MSA) operation in one or more layers of the Transformer [6]. Although ViT employs computationally expensive self-attention mechanisms at every layer, the self-attention operations across different layers are often highly similar, leading to substantial redundant computation. As illustrated in Figure 2 [6, 7, 11], SKIPAT approximates the attention computation in subsequent layers by reusing self-attention results from previous layers, thereby significantly reducing computational overhead.

To maintain performance while reusing self-attention, SKIPAT introduces a concise parameterized function that enhances computational speed while preserving capacity compared to the original Transformer architecture. Specifically, starting from a certain layer, the method reuses the Efficient Multi-Head Self-Attention (EMSA) module from the previous layer. By employing a newly designed parameterized function (see Figure 3 [6, 7, 11]), it skips the EMSA module in the next layer and merges it with the Multi-Layer Perceptron (MLP) module in the subsequent layer, progressively constructing the output of the Transformer layers. Unlike the typical feature propagation in standard Transformer blocks, SKIPAT uses a lightweight parameterized function to directly skip certain layers, effectively transmitting features to later network stages. This approach can be applied to each Transformer block within the Segformer encoder.

Figure 2
SKIPAT architecture.

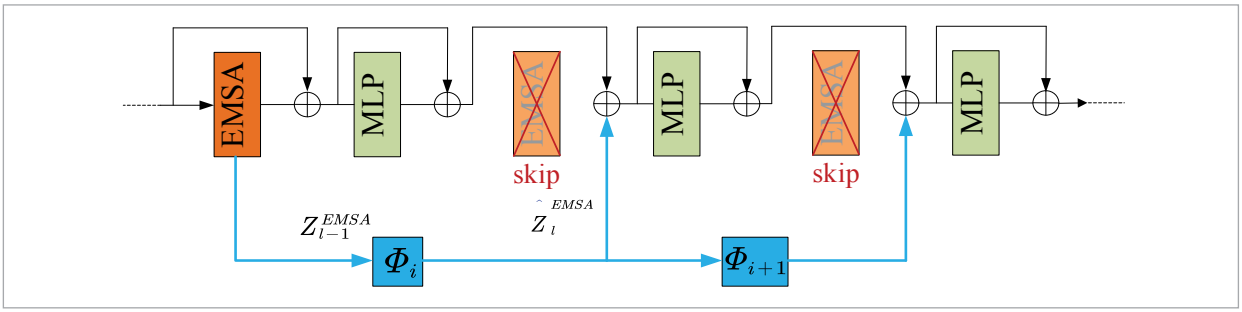
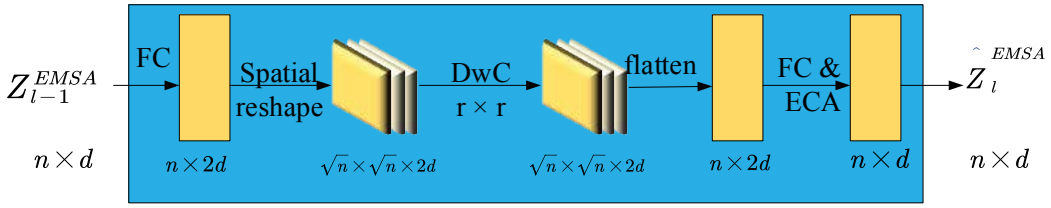


Figure 3
Schematic diagram of SKIPAT parameterized function.



In Figures 2-3, the calculation process of SKIPAT parameterized function is shown as Formula (1):

$$\hat{Z}_l^{EMSA} = ECA\left(FC_2\left(DwC\left(FC_1\left(Z_{l-1}^{EMSA}\right)\right)\right)\right). \quad (1)$$

The SKIPAT blocks are introduced into each layer of transformer block, and the modified calculation flow is represented as Formulas (2)-(3):

$$Z_l \leftarrow \Phi\left(Z_{l-1}^{EMSA}\right) + Z_{l-1} \quad (2)$$

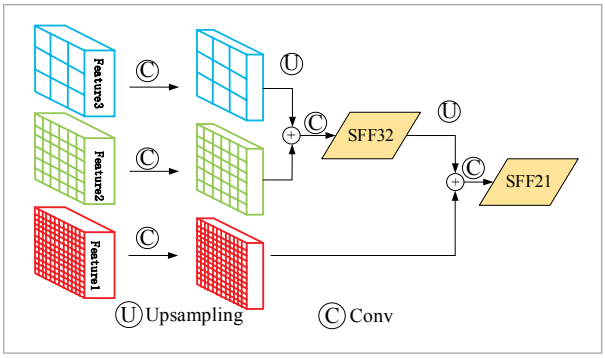
$$Z_l \leftarrow \mathrm{MLP}(Z_l) + Z_l. \quad (3)$$

2.3. Multi-level Feature Fusion Module

In semantic segmentation tasks, boundary delineation has remained a persistent challenge. Since such tasks involve pixel-level classification and rely heavily on the pixel-wise reconstruction of shallow-level features, greater emphasis must be placed on extracting these shallow features. Inspired by Yuan Liu et al. [13], who proposed a multi-scale feature fusion strategy that significantly improved segmentation accuracy, and by SERT [17], which selectively integrates features from different layers, this paper introduces a multi-scale feature fusion module. Unlike the original SegFormer decoder, which uses an MLP for feature fusion, our method adopts a stepped feature fusion (SFF) approach. This mechanism progressively merges features from deeper layers into shallower ones, effectively combining high-level semantic information with low-level spatial details, as illustrated in Figure 4 [13, 17].

Furthermore, drawing inspiration from ASPP [15], we propose a Multiple Effective Channel Attention (MECA) module. This component aims to capture semantic dependencies among feature channels and recalibrate feature maps adaptively. The overall structure of the MECA module is depicted in Figure 5 [15].

Figure 4
Overall pipeline of the SFF module.



2.4. Skip-Segformer Network

The proposed Skip-Segformer network consists of an encoder and a decoder network, with the overall architecture illustrated in Figure 6.

Figure 5
Overall workflow of the MECA module.

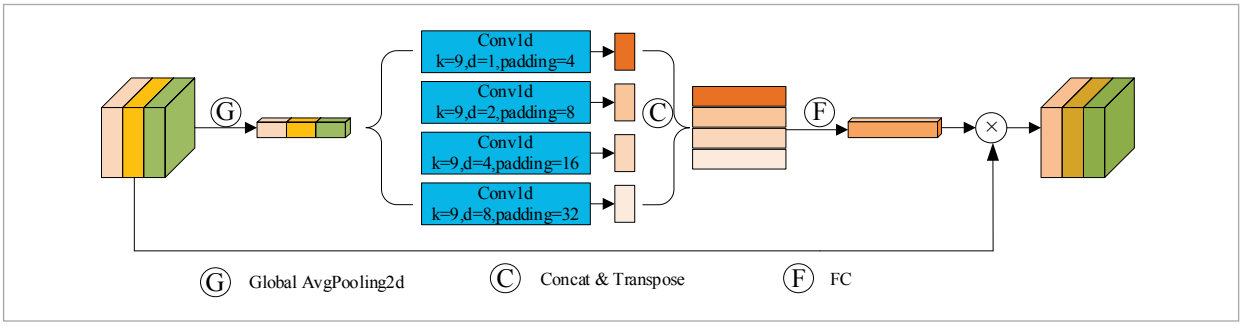
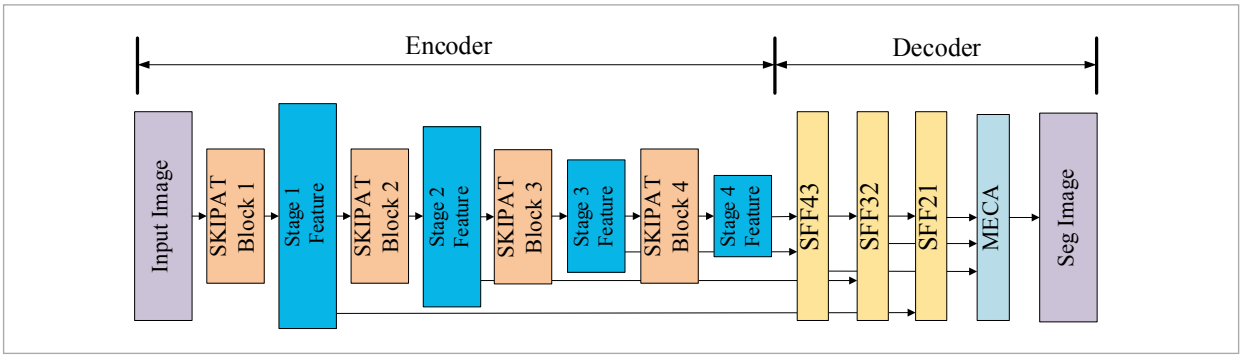


Figure 6
Architecture of the Skip-Segformer network.



The encoder network is based on the MiT-B0 backbone from SegFormer, augmented with the incorporation of SKIPAT blocks. During multi-scale feature extraction, starting from the second block, the model skips the EMSA computation in each transformer block to mitigate redundancy among successive EMSA operations. Instead, the EMSA output from the previous layer is reused via a parameterized SKIPAT function. The decoder network deviates from the original SegFormer design. It first performs stepwise fusion of multi-scale features from the encoder, propagating from deeper to shallower layers. These fused features are then upsampled and enhanced through the MECA module, which recalibrates the feature mappings, ultimately yielding the prediction.

The detailed structure of the Skip-Segformer network is summarized in Table 2. Skip-Segformer network is optimized primarily through the integration of SKIPAT blocks in the encoder and both SFF and MECA modules in the decoder. In Table 2, the "SKIPAT" column indicates whether the output from a preceding EMSA or SKIPAT block is used as input to the current layer, and the "SFF Level" specifies the features involved in the current stepped feature fusion.

Table 2
Architecture of the Skip-Segformer network.

	Modules	Layers	Operation layer Category	SKIPAT block	SFF level
Encoder	SKIPAT block1	Layer1	Mix-FNN(Attn)		
		Layer2	Mix-FNN(Attn)		
		Layer3	Mix-FNN(SKIPAT)	x:Attn(2)	
	SKIPAT block2	Layer4	Mix-FNN(Attn)		
		Layer5	Mix-FNN(Attn)		
		Layer6	Mix-FNN(SKIPAT)	x:Attn(5)	
		Layer7	Mix-FNN(SKIPAT)	x: SKIPAT(6)	
	SKIPAT block3	Layer8	Mix-FNN(Attn)		
		Layer9	Mix-FNN(Attn)		
		Layer10	Mix-FNN(SKIPAT)	x:Attn(9)	
		Layer11	Mix-FNN(SKIPAT)	x: SKIPAT(10)	
		Layer12	Mix-FNN(SKIPAT)	x: SKIPAT(11)	
		Layer13	Mix-FNN(SKIPAT)	x: SKIPAT(12)	
	SKIPAT block4	Layer14	Mix-FNN(Attn)		
		Layer15	Mix-FNN(Attn)		
		Layer16	Mix-FNN(SKIPAT)	x: Attn(15)	
Decoder	SFF43	Layer17	Conv		
		Layer18	Upsample		
		Layer19	DSC(x1,x2)		x1: Upsample (18),x2: Mix-FNN(13)
	SFF32	Layer20	Conv		
		Layer21	Upsample		
		Layer22	DSC(x1,x2)		x1: Upsample (21),x2: Mix-FNN(7)
	SFF21	Layer23	Conv		
		Layer24	Upsample		
		Layer25	DSC((x1,x2)		x1: Upsample (24),x2: Mix-FNN(3)
	MECA	Layer26	ECA		

2.5. Loss Function Construction

The primary task of this study is semantic segmentation involving two classes (background and film-mulched tobacco fields) with a significant imbalance between positive and negative samples. Therefore, a combination of Focalloss and Diceloss is adopted to measure the similarity between predictions and ground truth labels. Focalloss [20] was designed to address class imbalance, while Diceloss, first introduced in V-Net [5, 9, 12], is a region-based loss function where the training loss and gradient of a pixel depend not only on its own label and prediction, but also on those of other pixels. It is well-suited for semantic segmentation and performs robustly under severe class imbalance, as it emphasizes mining foreground regions [8, 14]. However, Diceloss can be unstable during training, particularly when segmenting small objects. To mitigate this issue, we integrate Focalloss with Diceloss. The specific loss function (l_{sum}) used in this paper is a combination of focalloss loss function ($l_{Focaloss}$) and diceloss loss function ($l_{Diceloss}$). The relevant calculation process is shown in Formulas (4)-(7):

$$l_{sum} = l_{Focalloss} + l_{Diceloss} \tag{4}$$

$$l_{Focalloss} = \begin{cases} -(1-\hat{p})^{\gamma} log(\hat{p}) & if\ y = 1 \\ -\hat{p}^{\gamma} log(1-\hat{p}) & if\ y = 0 \end{cases} \tag{5}$$

Let $p_t = \begin{cases} \hat{p} & if\ y = 1 \\ 1 - \hat{p} & otherwise \end{cases}$, then $l_{Focalloss}$ can be unified as:

$$l_{Focalloss} = -(1-p_t)^{\gamma} log(p_t) \tag{6}$$

$$l_{Diceloss} = 1 - \left(2 * \sum_i^N \hat{y}_i * y_i\right) / \left(\sum_i^N \hat{y}_i^2 + \sum_i^N y_i^2\right), \tag{7}$$

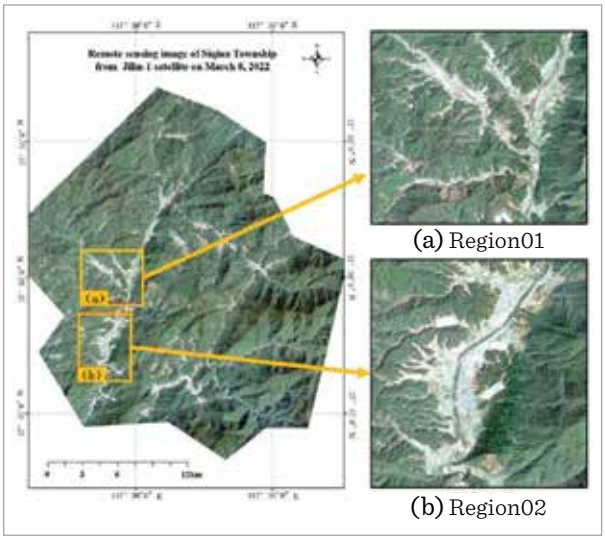
where p_t helps alleviate the positive-negative sample quantity imbalance, and γ controls the sample quantity imbalance between easy and hard examples. Here, $\hat{p}$ denotes the predicted probability of the pixel belonging to the film-mulched tobacco field class, and, $\gamma \in [0,5]$, γ can be empirically set to 2. N represents the total number of pixels in the considered region, $\hat{y}_i$ is the predicted probability of the i-th pixel, $\hat{y}_i \in [0,1]$, and y_i is the corresponding ground truth label, $y_i = 0$ *or* 1.

3. Research Areas and Sample Preparation

3.1. Experimental Areas

The experiment was conducted in Siqian Township, Guangze County, Nanping City, Fujian Province. Located in the northern part of Guangze County, Siqian Township is one of the major tobacco-producing areas in the county, covering a total area of approximately 433 km². Two representative regions (Region01 and Region02) were selected as experimental areas, whose specific locations are illustrated in Figure 7.

Figure 7
Spatial distribution and satellite imagery of the experimental areas.



As shown in Figure 7, both experimental regions (Region01 and Region02) exhibit complex topography. However, Region01 contains more fragmented and irregularly shaped tobacco fields, whereas Region02 features a higher proportion of regular tobacco fields distributed along river valleys. Some typical film-mulched tobacco field scenes are shown in Figure 8.

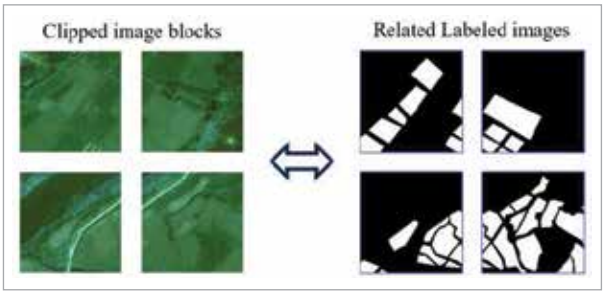
Figure 8
Examples of typical film-mulched tobacco fields.

3.2. Data Sources and Sample Collection

This experiment utilized remote sensing imagery from the "Jilin-1" Wide-Band 01B satellite, acquired on March 8, 2022, covering the entire area of Siqian Township, Guangze County. The imagery consists of one panchromatic band with a spatial resolution of 0.5 meters and four multispectral bands with a spatial resolution of 2 meters. Preprocessing steps including radiometric, topographic, and atmospheric corrections were applied to the original satellite images. Subsequently, pan-sharpening was performed to fuse the panchromatic and multispectral bands, generating RGB images with a spatial resolution of 0.5 meters.

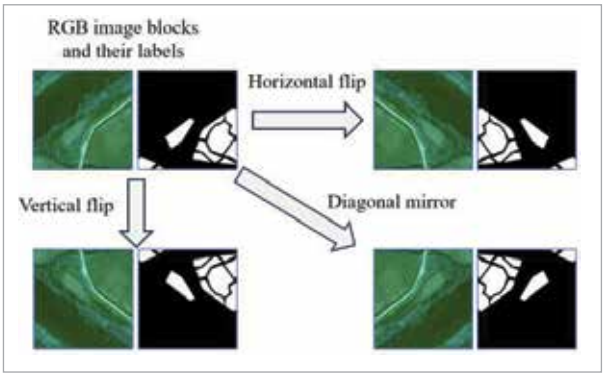
Field surveys were first carried out using handheld GPS devices to manually geolocate several typical film-mulched tobacco fields. The original annotated sample set was created using ArcGIS 10.5. Each sample image contained pixels categorized into two classes: background (labeled as 0) and foreground, i.e., film-mulched tobacco fields (labeled as 1). Due to the large size of the original imagery and limited computational resources, directly inputting the entire high-resolution remote sensing image into a deep neural network for training would require substantial computational power. Therefore, the images and corresponding label data were divided into patches and processed in parallel. Using a sliding window of size 256×256 pixels with a 30% overlap along both the X and Y axes, a large number of square image patches were generated. These patches served as the basic input units for the Skip_Segformer network, enhancing the learning efficiency of the deep neural network and preventing the disruption of feature integrity during training. Examples of clipped film-mulched tobacco field image patches and their corresponding label data are shown in Figure 9.

Figure 9
Examples of clipped image patches of film-mulched tobacco fields and corresponding label data.



As deep learning-based extraction of film-mulched tobacco fields requires a large number of labeled samples, and the manually annotated original samples were limited, data augmentation was applied to increase the number of samples. In this paper, sample augmentation was performed by rotating the image patches. Specifically, horizontal flipping, vertical flipping, and diagonal mirroring were applied, increasing the number of labeled image patches to four times the original amount. Figure 10 illustrates sample image patches and their labels before and after augmentation.

Figure 10
Examples of image patches and labels before and after augmentation.



The augmented sample set was divided into training, validation and test sets. The detailed distribution is presented in Table 3.

Table 3
Division of the sample set.

Dataset	Category	Number of image blocks	Block size
Tobacco01 (Region01)	train	2156	256×256
	val	308	256×256
	test	616	256×256
Tobacco02 (Region02)	train	3732	256×256
	val	534	256×256
	test	1066	256×256

4. Experimental Environment and Evaluation Metrics

4.1. Experimental Environment Configuration

The experiment was conducted under a 64-bit Ubuntu operating system. The deep learning framework used was PyTorch 1.7, with Python 3.7 as the programming language and CUDA version 10.1. The hardware configuration included an Intel(R) Xeon(R) CPU E5-2620 v4 @ 2.10GHz, 160 GB of RAM (5×32 GB), one NVIDIA GeForce RTX 2060 GPU with 8 GB of VRAM, and a 1 TB hard disk drive. Identification results of film mulched tobacco fields and output imagery were visualized using ArcGIS 10.5. The experimental environment settings for training the Skip_Segformer model are summarized in Table 4. During model training, the AdamW optimizer was used for parameter optimization, with an initial learning rate of 1e-4, a minimum learning rate of 1e-6, a weight decay of 0.01, and a Cosine learning rate schedule. The Skip_Segformer model was trained for 300 epochs.

Table 4
Parameter settings for Skip_Segformer training.

Parameter Items	Item values
Input size	256×256
Train size	256×256
Test size	1024×1 024
Epoch	300
Batch Size	4
Optimizer	Adamw
Learning strategy	Cosine
Data augmentation	Random cut; Random flip

4.2. Evaluation Metrics

This paper employs a confusion matrix and related metrics to evaluate the accuracy of film-mulched tobacco fields extraction based on the Skip_Segformer model. The definition of the confusion matrix is provided in Table 5.

Table 5
Confusion Matrix.

	Number of the pixels in tobacco fields predicted to be film-mulched	Number of the pixels in tobacco fields predicted to be non-film-mulched
Number of the pixels in actual film-mulched tobacco fields	TP	FN
Number of the pixels in actual non- Film-mulched tobacco fields	FP	TN

There are six relevant metrics used in the actual extraction of film mulched tobacco fields in this paper. They are defined as follows (where the positive class refers to film-mulched tobacco fields, and the negative class refers to the background):

1 **Precision(P)**

P represents the proportion of correctly predicted positive samples out of all samples predicted as positive. It is calculated as shown in Formula (8):

$$P = \frac{TP}{TP + FP} \tag{8}$$

2 **Recall (*R*)**

R denotes the proportion of correctly predicted positive samples out of all actual positive samples. It is calculated as shown in Formula (9):

$$R = \frac{TP}{TP + FN} \tag{9}$$

3 **Overall Accuracy (*OA*)**

OA is the ratio of all correctly classified samples (i.e., the sum of the diagonal elements in the confusion matrix) to the total number of samples (i.e., the sum of all elements in the confusion matrix). It is a fundamental metric for evaluating the overall performance of a classification model. In this paper, *OA* specifically reflects the proportion of correctly classified pixels (both film-mulched and non-film-mulched tobacco fields) relative to the total number of pixels. It is computed as shown in Formula (10):

$$OA = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

4 **Intersection over Union (*IoU*)**

IoU measures the ratio between the intersection and the union of the real samples and predicted samples for all classes. Specifically, it is the number of correctly predicted positive samples divided by the sum of three quantities: the true positives, the false negatives, and the false positives. It is calculated as shown in Formula (11):

$$IoU = \frac{TP}{TP + FP + FN} \tag{11}$$

5 **mean Intersection over Union (*mIoU*)**

mIoU is the average of the *IoU* values calculated for each class. The computation is given in Formula (12), where *n* represents the number of classes:

$$mIoU = \frac{\sum_i^n IoU}{n} \tag{12}$$

6 **F1-Score (*F*1)**

*F*1 value is the harmonic average value of *P* and *R*. It is a metric of comprehensive consideration of classification accuracy, especially suitable for finding the balance point between *P* and *R*. Its calculation is shown in Formula (13). Generally, the high *F*1 value requires that precision and Recall are both relatively high. It can be calculated as shown in Formula (13):

$$F1 = \frac{2 \times P \times R}{P + R} \tag{13}$$

5. Experimental Results and Analysis

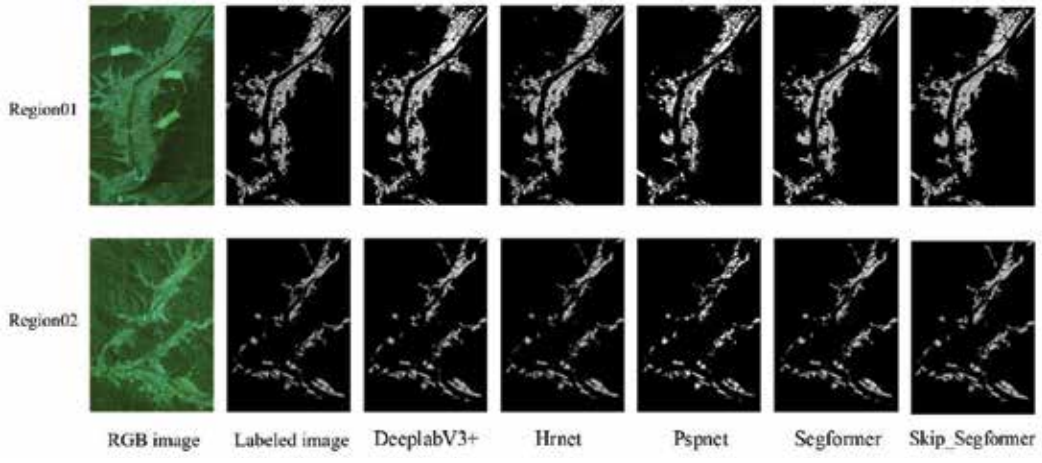
To evaluate the performance of the proposed Skip_Segformer-based method for extracting film-mulched tobacco fields, comparative experiments were conducted using other four representative deep neural network models (DeeplabV3+, Hrnet, Pspnet, and SegFormer) along with the proposed Skip_Segformer network model. The experiments were based on the division of training, validation and test sets shown in Table 3 and the network model training parameters specified in Table 4. To more effectively analyze the extraction performance of different models, comparisons and evaluations were carried out at two different scales: across the entire experimental area and within several representative local regions.

5.1. Comparison and Analysis of Extraction Results over the Entire experimental Area

For the entirety of Experimental Region 1 (Region01) and Experimental Region 2 (Region02), the five aforementioned models were applied to extract film-mulched tobacco fields. The results are illustrated in Figure 11.

As shown in Figure 11, the Pspnet model exhibited noticeable omission errors, particularly in extracting small tobacco fields. The three models of DeeplabV3+, Hrnet and SegFormer have their own advantages and disadvantages in the problem of missed extraction and false extraction. However, DeeplabV3+, Hrnet, Pspnet and SegFormer models showed varying degrees of boundary loss or adhesion in the extracted film-mulched fields. In contrast, the proposed Skip_Segformer network effectively mitigated these issues, producing clearer and more complete field boundaries. Moreover, in the

Figure 11
Comparison of film-mulched tobacco field extraction results obtained by different network models over the two experimental areas.



areas with numerous small and fragmented fields, Skip_Segformer achieved higher efficiency and accuracy in extraction compared to the other four models. The quantitative comparison of extraction accuracy among the five models for the two regions is summarized in Tables 6-7.

Table 6
Comparison of extraction accuracy among different network models in Region01.

	DeeplabV3+	Hrnet	Pspnet	SegFormer	Skip_Segformer
Precision	0.84	0.88	0.76	0.89	0.90
Recall	0.82	0.85	0.88	0.94	0.94
OA	0.95	0.97	0.96	0.98	0.98
IOU	0.73	0.77	0.69	0.84	0.86
mIOU	0.86	0.87	0.83	0.91	0.92
F1	0.83	0.86	0.82	0.91	0.92

Table 7
Comparison of extraction accuracy among different network models in Region02.

	DeeplabV3+	Hrnet	Pspnet	SegFormer	Skip_Segformer
Precision	0.83	0.85	0.76	0.90	0.92
Recall	0.82	0.83	0.86	0.93	0.95
OA	0.98	0.97	0.96	0.98	0.98
IOU	0.77	0.76	0.68	0.83	0.87
mIOU	0.87	0.86	0.83	0.90	0.93
F1	0.82	0.84	0.81	0.91	0.93

As can be seen from the above comparison in Tables 6-7, in Region01, where tobacco fields are relatively regular, the SegFormer model outperformed DeeplabV3+, Hrnet and Pspnet across all metrics (P, R, IoU, mIoU and F1-score). Specifically, compared to DeeplabV3+, the values of the aforementioned five metrics of SegFormer were increased by 5%,12%,11%,5% and 8%, respectively; Compared to Hrnet, the values of the aforementioned five metrics of SegFormer were increased by 1%, 9%, 7%, 4% and 5%, respectively; Compared to Pspnet, the values of the aforementioned five metrics of SegFormer were increased by 13%, 6%, 15%, 8% and 9%, respectively. After integrating the SKIPAT and MFF modules into SegFormer, the improved model (Skip_Segformer) exhibited a significant increase in accuracy compared with the original SegFormer. The values of the aforementioned five metrics of Skip_Segformer were increased were increased by 1%, 2%, 1% and 1%, respectively, while the Overall Accuracy (OA) remained at 0.98.

In Region02, Skip_Segformer outperformed the other four models across all metrics except for OA, which remained comparable to SegFormer at 0.98. Compared to the baseline SegFormer, the values of the aforementioned five metrics of Skip_Segformer were increased by 2%, 2%, 4%, 3% and 2%, respectively. Moreover, relative to DeeplabV3+, Hrnet and Pspnet, Skip_Segformer showed substantial improvements in the above five metrics except OA: the values of P were increased by 9%, 7% and 16%, respectively; the values of R were increased by 13%, 12% and 9%, respectively; the values of IoU were increased by 10%, 11% and 19%, respectively; the values of mIoU were increased by 6%, 7% and 10%, respectively; the values of F1-score were increased by 11%, 9% and 12%, respectively.

In summary, the improved SegFormer network (Skip_Segformer) achieved significantly higher accuracy in extracting film-mulched tobacco fields compared to other four network models (DeeplabV3+, Hrnet,Pspnet and original SegFormer).

5.2. Comparison and Analysis of Extraction Results in Representative Local Regions

To analyze the experimental outcomes of the five network models at a finer scale, two representative local regions (LR1 and LR2) were selected from Experimental Region 1 and Experimental Region 2, respectively, for detailed comparative analysis. The extraction results are presented in Figures 12-13.

As shown in Figures 12-13, the five models tested under the same experimental conditions and using the same dataset yielded different results in terms of both visual outcome and accuracy in extracting film-mulched tobacco fields in the two local regions. In local region LR1 (Figure 12), all five models achieved relatively high accuracy in extracting densely distributed and regularly shaped tobacco fields. Specifically: In subfigure LR1-DeeplabV3+, the model performed well overall but produced some false extractions, misclassifying non-field features (e.g., woodland and grassland) as tobacco fields. Some tobacco field edges

Figure 12

Comparison of detailed extraction results of five models in local region LR1.

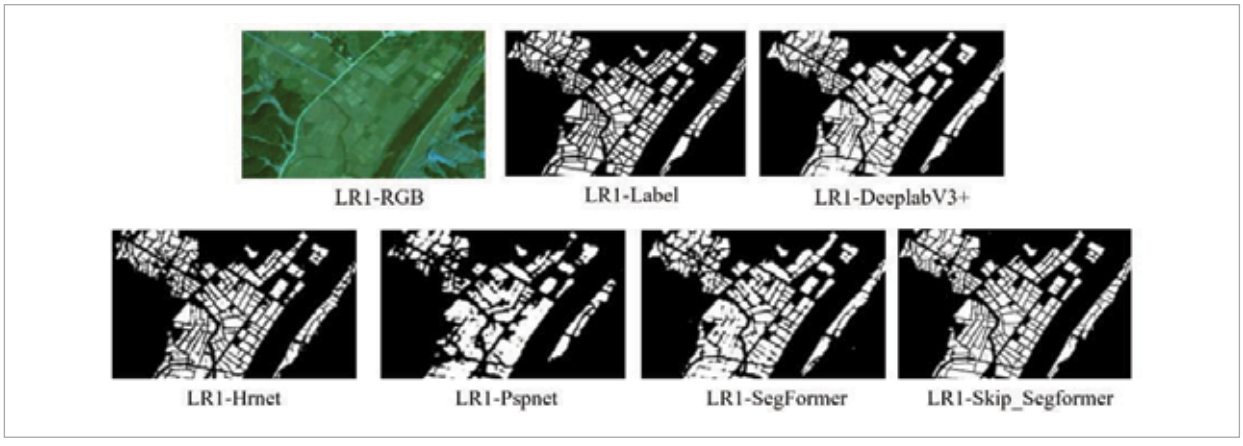
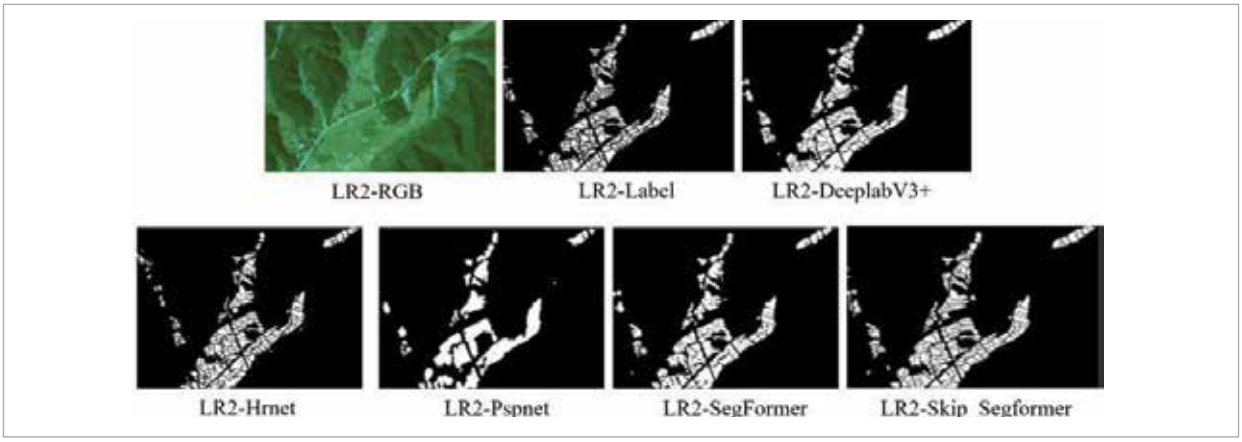


Figure 13
Comparison of detailed extraction results of five models in local region LR2.



also showed signs of overfitting; in subfigure LR1-Hrnet, Hrnet extraction resulted in certain omissions and false extractions. Although speckle-like noise was present, the extracted boundaries were relatively clear with minimal adhesion; subfigure LR1-Pspnet exhibited more severe false extractions and omissions compared to the previous two models, with significant boundary adhesion being the most notable issue; in subfigure LR1-SegFormer, SegFormer outperformed Pspnet, with fewer false positives and omissions, complete boundary extraction, and absence of speckle noise. However, its overall boundary extraction quality was slightly inferior to that of DeeplabV3+ and Hrnet; subfigure LR1-Skip_Segformer shows the proposed Skip_Segformer had no obvious false extractions or omissions, and delivered complete and clear boundary extraction.

Similarly, in local region LR2, it can be seen from Figure 13 that the above five different networks have their own advantages and disadvantages in the extraction of film mulched tobacco fields: in subfigure LR2-DeeplabV3+, DeeplabV3+ was prone to false extractions in areas with small and densely distributed fields, with unsatisfactory boundary delineation; in subfigure LR2-Hrnet,it can be seen Hrnet tended to omit small fields in the upper-left dense region but achieved better boundary extraction than DeeplabV3+;in subfigure LR2-Pspnet, Pspnet extraction showed severe boundary adhesion, capturing only the approximate location and shape of clustered fields; in subfigure LR2-SegFormer,SegFormer model performed well in recognizing non-dense and irregular fields, achieving results comparable to DeeplabV3+ with fewer omissions and relatively clear boundaries though still inferior to Hrnet in boundary quality; in subfigure LR2-Skip_Segforme, the proposed Skip_Segformer network model excelled in identifying non-dense and irregular fields. In fragmented areas with small field patches, it almost provided complete extraction with clear boundaries.

5.3. Generalization Capability Analysis of the Skip_Segformer Model

To further evaluate the accuracy and generalization capability of Skip_Segformer, particularly its performance in extracting film-mulched tobacco fields over large areas with limited training samples, we compared it with four other network models (DeeplabV3+, Hrnet, Pspnet, and SegFormer). To enhance the generalizability of the recognition model, the training dataset was expanded by combining the training samples from both Experimental Region 1 and Experimental Region 2. Each of the above five original networks was retrained using this merged dataset to produce updated models. These new models were then applied to extract film-mulched tobacco fields across the entire experimental area (covering Siqian Township). The training parameters remained consistent with those listed in the earlier parameter table (Table 4). The extraction results obtained from the five retrained models are presented in Figures 14-18. This evaluation aimed to validate the practicality and accuracy of the proposed algorithm, ensuring its suitability for large-scale ex-

traction of film-mulched fields and estimation of tobacco planting areas. Due to the large area of Siqian Township and the scattered distribution of relatively small tobacco fields, detailed information from the large-scale extraction is not easily visible at a glance. Therefore, six representative sub-regions (A-F) were selected to clearly illustrate the differences among the extraction results from the five models.

Figure 14 shows the extraction results of film-mulched tobacco fields across the entire Siqian Township using the DeeplabV3+ model. Overall, the extraction performance is acceptable. In sub-regions A, B, and C, where the film-mulched fields are densely distributed and irregularly shaped, DeeplabV3+ extraction produces blurred boundaries and noticeable adhesion between adjacent fields. Some central plots in sub-regions A and B exhibit internal holes. In sub-region D, which contains no actual film-mulched tobacco fields, but DeeplabV3+ model incorrectly extract some ambiguous patches there, indicating obvious false extraction. In sub-regions E and F, which contain regular and dense distributed fields, the boundaries are relatively clear. However, the model fails to fully extract the boundaries of the central dense fields in sub-region F, and omits some fragmented fields in the lower-right part of the same sub-region. These observations suggest that the DeeplabV3+ model lacks sufficient accuracy and generalization capability for extracting film-mulched tobacco fields across Siqian Township. Figure 15 presents

Figure 14
The results of film-mulched tobacco fields extraction based on Deeplabv3+.

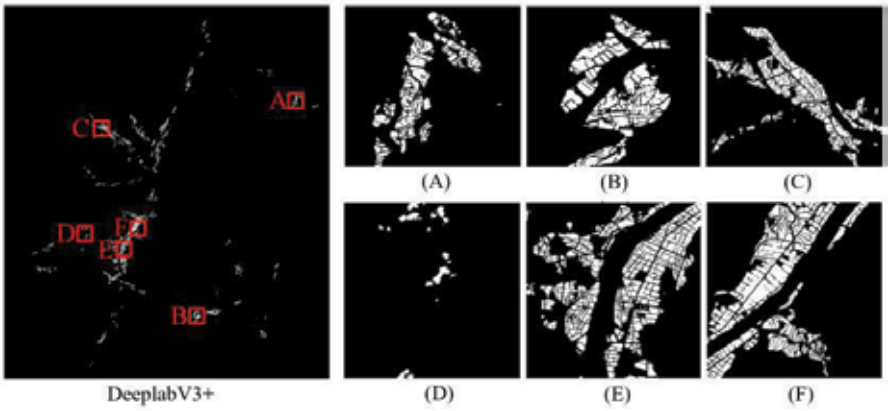


Figure 15
The results of film-mulched tobacco fields extraction based on Hrnet.

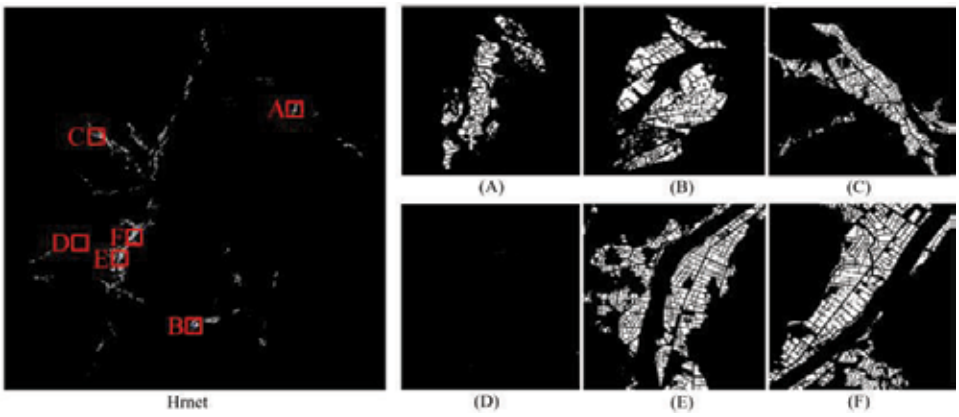


Figure 16
The results of film-mulched tobacco fields extraction based on Pspnet.

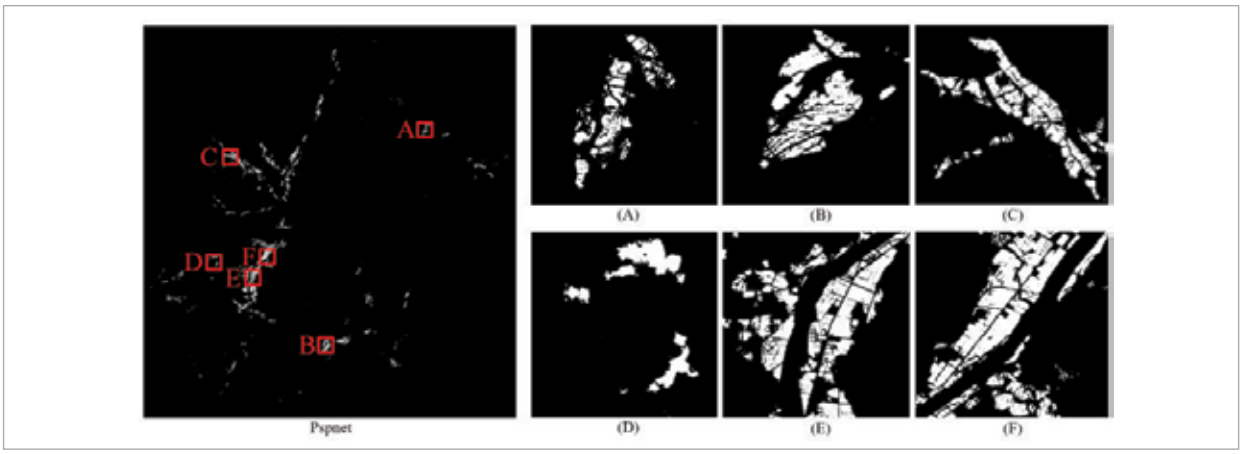


Figure 17
The results of film-mulched tobacco fields extraction based on SegFormer.

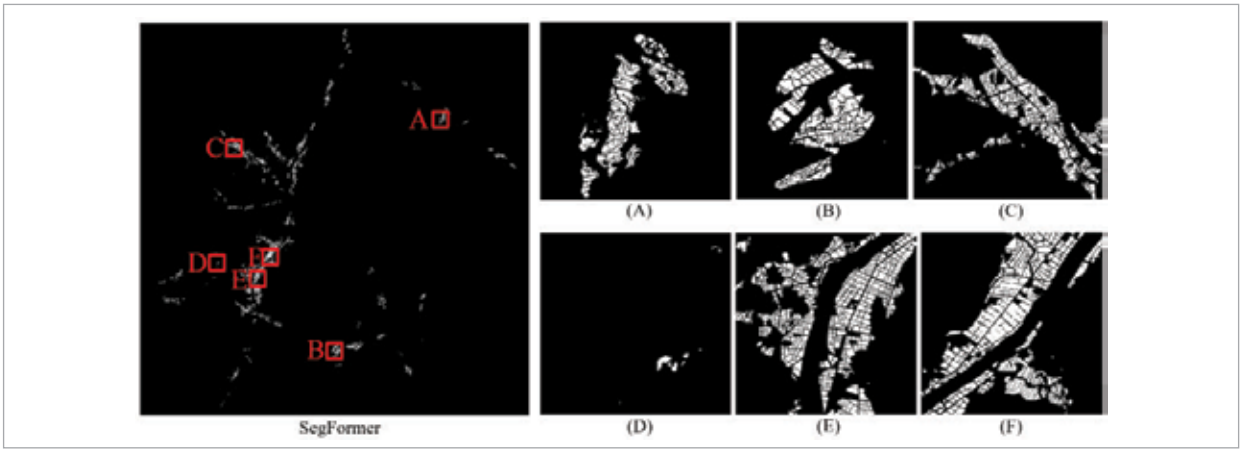
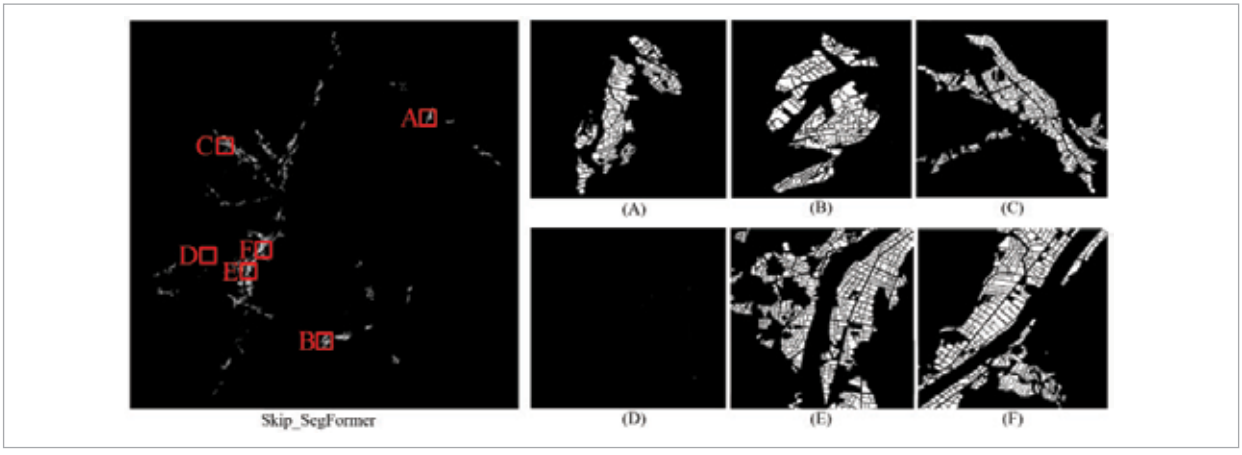


Figure 18
The results of film-mulched tobacco fields extraction based on Skip_Segformer.



the overall extraction results obtained by the Hrnet model. Compared with DeeplabV3+, Hrnet produces more false extractions and noticeable salt-and-pepper noises, along with several omission errors. In sub-region A, scattered speckles appear due to misclassification between bare land and film-mulched fields. Similar errors occur in sub-regions E and F. Nevertheless, the boundaries of regular and dense fields in sub-regions E and F are extracted completely. Overall, the Hrnet model also demonstrates limited accuracy and generalization capability in extracting film-mulched tobacco fields. Figure 16 illustrates the extraction results from the Pspnet model. In sub-region D, the model misclassifies paddy fields as film-mulched tobacco fields. In sub-regions E and F, which contain regular and dense fields, Pspnet model only predicts approximate shapes and fails to accurately distinguish boundaries between adjacent fields, sometimes resulting in complete misclassification. Additionally, many uncultivated weedy areas are falsely extracted, appearing as speckled noise in the output. Clearly, the Pspnet model also performs poorly in terms of accuracy and generalization in extracting film-mulched tobacco fields. Figure 17 displays the results from the SegFormer model. Its performance is superior to the previous three models. In sub-regions E and F, with dense tobacco fields, it produces fewer false extractions. It also avoids significant boundary adhesion in sub-region C, which contains dense irregular fields. However, it still misclassifies some paddy fields in sub-region D as tobacco fields, and produces minor errors in the central dense fields in sub-region F. Overall, the SegFormer model shows improved but still limited generalization capability in extracting film-mulched tobacco fields. Figure 18 demonstrates the results from the proposed Skip_Segformer model. In sub-regions E and F, the boundaries are clearly extracted, and even narrow ridges between adjacent fields are well distinguished with little adhesion. In sub-regions A, B, and C, which contain irregularly shaped fields, the boundaries are well preserved. These results indicate that the Skip_Segformer model exhibits better accuracy and generalization capability compared to the other four models.

The quantitative comparison of the five retrained models is summarized in Table 8. Except for the Overall Accuracy (OA), which is 1% lower than that of SegFormer and Pspnet, Skip_Segformer outperforms the other four models across all metrics. In particular, the F1-score of Skip_Segformer is 6%, 9%, 13%, and 3% higher than that of DeeplabV3+, Hrnet, Pspnet, and SegFormer, respectively. This confirms that Skip_Segformer achieves the highest extraction accuracy in the large-scale extraction of plastic-mulched tobacco fields in Siqian Township.

In summary, the proposed Skip_Segformer model exhibits superior generalization capability and extraction accuracy compared to the other four models, demonstrating strong potential for large-scale extraction of film-mulched tobacco fields and estimation of tobacco planting areas.

5.4. Estimation of Tobacco Planting Area

Although the extracted area of film-mulched tobacco fields does not exactly equal the actual tobacco planting area due to partial false extractions and omissions, the two are generally close. Therefore, it is feasible to estimate the total tobacco planting

Table 8

Comparison Comparison of extraction accuracy among different models across Siqian Township.

Metrics	DeeplabV3+	Hrnet	Pspnet	Segformer	Skip_Segformer
Precision	0.86	0.82	0.71	0.86	0.88
Recall	0.83	0.80	0.84	0.89	0.92
OA	0.97	0.96	0.98	0.98	0.97
IOU	0.75	0.71	0.65	0.80	0.82
mIOU	0.85	0.84	0.81	0.89	0.90
F1	0.84	0.81	0.77	0.87	0.90

area by calculating the total area of extracted film-mulched fields. Based on the previously obtained extraction results from the "Jilin-1" satellite remote sensing image over Siqian Township on March 8, 2022, we conducted a comprehensive statistical analysis of the pixel counts for each field patch. The total areas of film-mulched fields predicted by the five different models (DeeplabV3+, Hrnet, Pspnet, SegFormer, and Skip_Segformer) were calculated and compared with the actual tobacco planting areas of Siqian Township in 2022, which officially reported by the local government. The results are presented in Table 9. As shown in Table 9, the deviation rate between the tobacco planting area estimated using the Skip_Segformer extraction results and the actual area (4190 mu) is only 0.16%, which is significantly lower than those of the other four network models. Moreover, the consistency between the estimated planting area and the extraction accuracy further validates the reliability of the Skip_Segformer model. These results demonstrate that Skip_Segformer offers strong feasibility and applicability for large-scale extraction of film-mulched tobacco fields and estimation of tobacco planting area. It can effectively support tobacco management departments in monitoring total planting area and provide a decision-making basis for local tobacco planting planning.

Table 9

Comparison of predicted tobacco planting areas in Siqian Township based on different models.

Model	actual total area of to bacco planting /mu	predicted total area of film mulched tobacco field/mu	Deviation rate /%
DeeplabV3+	4190	4272.147	1.96%
Hrnet	4190	3676.624	12.25%
Pspnet	4190	4868.828	16.20%
Segformer	4190	4270.997	1.93%
Skip_Segformer	4190	4196.525	0.16%w

6. Discussion

First, in this work, since mulching tobacco fields is a necessary step before transplanting tobacco seedlings in the tobacco planting process, the total area of the film-mulched tobacco fields is linearly correlated with the actual tobacco planting area and their values are relatively close. Therefore, through multiplying the former by a correction factor, the latter can be estimated more accurately. The tobacco yield is not only related to its planting area, but also to the growth status of tobacco. In this work, the Skip_Sgformer model was mainly used for estimating the area of film-mulched tobacco fields. Its analysis results cannot reflect the growth status of tobacco, so it cannot be used alone to estimate the final yield of tobacco fields.

Second, complex surrounding environments such as other crops or weeds around the tobacco fields, as well as other crop greenhouses, may affect the accuracy of extracting film-mulched tobacco fields. This can be solved by overlaying permanent tobacco field mask layers to remove the impact of other crop greenhouses, and removing the impact of other crops or weeds through extracted and analyzed vegetation indices.

Third, the Skips_Sgformer model in this work is specifically designed for extracting from film-mulched tobacco fields, and has certain reference value for extracting from non-film-mulched tobacco fields or other agricultural film-mulched farmlands. However, the structure of the Skips_Sgformer model needs to be adjusted appropriately according to specific scenarios. In addition, due to the differences in the spatial distribution and plant categories of different tobacco fields, the Skips_Sgformer model proposed in this work has good generalization ability for extracting film-mulched tobacco fields in other regions, and its network structure has great reference value. However, in complex scenarios, the Skips_Sgformer model still needs to be fine-tuned and re-optimized according to local specific scenarios.

Finally, the construction and testing of the Skip_Sgformer model mainly focused on the county-scale spatial range, achieving good accuracy and generalization in extracting film-mulched tobacco fields. However, the accuracy and generalization of the model have not been tested in a larger spatial scale. The model will be trained and tested in larger spatial scale in the future work to achieve better generalization performance.

7. Conclusion

This paper introduced the SKIPAT and MFF modules into the traditional SegFormer model, proposing a Skip_Segformer-based semantic segmentation approach using high-resolution remote sensing imagery for identifying film-mulched tobacco fields. Based on the extracted field patches, tobacco planting area was further estimated. The experimental results in Siqian Township, Guangze County, Nanping City, Fujian Province, demonstrated that Skip_Segformer outperforms four other network models (DeeplabV3+, Hrnet, Pspnet, and SegFormer) in both accuracy and generalization capability for film-mulched fields extraction and planting area estimation. In 2022, the extraction accuracy of Skip_Segformer across Siqian Township reached 97%, with a deviation rate of only 0.16% in planting area estimation, significantly exceeding the performance of the other four models. The proposed Skip_Segformer network produced clearer and more complete boundaries of film-mulched fields. Moreover, in the areas with small and fragmented fields, it achieved higher efficiency and accuracy in detecting small tobacco plots compared to the other models. These results confirm that Skip_Segformer is a feasible and applicable network model for large-scale extraction of film-mulched tobacco fields and estimation of planting area, capable of supporting tobacco administration departments in monitoring total planting area and facilitating local tobacco planning decisions. Nevertheless, despite its improved accuracy in boundary extraction and reduced computational complexity, the efficiency, accuary, and generalization ability of Skip_Segformer for even larger-scale film-mulched tobacco fields extraction and planting area estimation require further validation due to the complexity of field environments.In addtion, the preforemance of Skip_Segformer model in the non-film-mulched tobacco fields,the film-mulched farmlands of other crops or complex surrounding environments,need to be validated in the future work.

Funding

This research was funded by Fujian Province Science and Technology Plan Project (Guided Project, 2023N0021), the National Natural Science Foundation of China (NSFC, 41501451), China Postdoctoral Science Foundation (2015M571963).

References

1. Cui, W. H., Lan, Y. B., Li, J. Q., Yang, L., Zhou, Q., Han, G. T., Xiao, X., Zhao, J., Qiao, Y. L. Apple Yield Estima-tion Method Based on CBAM-ECA-Deeplabv3+ Image Segmentation and Multi-Source Feature Fusion. Sensors, 2025, 25(10), 3140-3159. https://doi.org/10.3390/s25103140
2. Cao, Z., Huang, Y., Ji, Z., Zhou, Y., Peng, Z. GF-ResFormer: A Hybrid Gabor-Fourier ResNet-Transformer Network for Precise Semantic Segmentation of High-Resolution Remote Sensing Imagery. IEEE Journal of Se-lected Topics in Applied Earth Observations and Remote Sensing, 2025, 18, 23779-23800. https://doi.org/10.1109/JSTARS.2025.3606691
3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X. H., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N. An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale. Proceedings of 2021 International Conference on Learning Representations (ICLR 2021), Online, May 4-8, 2021, 1-22. https://arxiv.org/abs/2010.11929
4. Fu, B. H., Huang, L. Extraction of Tobacco Planting Area from UAV Images Based on Deep Semantic Segmentation. Communications Technology, 2022, 55(02), 181-186. https://d.wanfangdata.com.cn/periodical/txjs202202007
5. Guo, X. Y., Yao, H. M., Liu, Y. A., Ng, M., Song, S. J. Deep Learning Approach for Microwave Imaging in Broad Frequency Band Based on Physics-Driven Loss and Deep Convolutional V-Net Structure. IEEE Microwave and Wireless Technology Letters, 2025, 35(9), 1264-1267. https://doi.org/10.1109/LMWT.2025.3575160

6. Han, K., Xiao, A., Wu, E. H., Guo, J. H., Xu, C. J., Wang, Y. H. Transformer in Transformer. Advances in Neural Information Processing Systems, 2021, 34, 15908-15919. https://doi.org/10.48550/arXiv.2103.00112
7. He, Q. Y., Sun, T. Integrating Linear Skip-Attention with Transformer-Based Network of Multi-Level Features Ex-traction for Partial Point Cloud Registration. IET Image Processing, 2025, 19(1), 1-18. https://doi.org/10.1049/ipr2.70055
8. Hou, Q., Zhou, D., Feng, J. Coordinate Attention for Effi-cient Mobile Network Design. Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2021), Nashville, USA, June 20-25, 2021, 13713-13722. https://doi.org/10.1109/CVPR46437.2021.01350
9. Hu, S. M., He, J. T., Gu, W. D. Visualization System of Human Brain Hippocampi Segmentation for MRI Images Based on V-Net. Proceedings of 2023 8th International Conference on Intelligent Computing and Signal Pro-cessing (ICSP 2023), Xi'an, China, April 21-23, 2023, 456-460. https://doi.org/10.1109/ICSP58490.2023.10248537
10. Jian, R. B., Cai, Z. Y., Yang, Z. S., Wang, W. Z., Liu, Y., Chen, J. F., Wang, M. L. Research on Image Segmentation Method of Field Wheat Harvest Boundary Based on U-Net. Journal of Henan Agricultural University, 2023, 57(03), 444-450. https://www.cnki.com.cn/Article/CJFDTotal-NNXB202303009.htm
11. Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., Shah, M. Transformers in Vision: A Survey. ACM Computing Surveys (CSUR), 2022, 54(10s), 1-41. https://doi.org/10.1145/3505244
12. Lin, H. B., Zhang, Y. H., Chen, X. F., Wang, H., Xia, L. Z. Research on Pulmonary Nodule Segmentation Algorithm Based on Improved V-Net. Proceedings of 2022 IEEE 6th Advanced Information Technology, Electronic and Au-tomation Control Conference (IAEAC 2022), Beijing, China, October 03-05, 2022, 194-198. https://doi.org/10.1109/IAEAC54830.2022.9929520
13. Liu, Y., Zhang, S. Y., Chen, J. C., Yu, Z. H., Chen, K., Lin, D. H. Improving Pixel-Based MIM by Reducing Wasted Modeling Capability. Proceedings of 2023 IEEE/CVF In-ternational Conference on Computer Vision (ICCV 2023), Paris, France, October 2-6, 2023, 5361-5372. https://doi.org/10.1109/ICCV51070.2023.00494
14. Miao, R., Li, Y., Zhou, K., Zhang, Y. N., Chang, R. R., Meng, G. Research on Improved Faster R-CNN Multi-Target Detection Model for Remote Sensing Imagery. Computer Engineering, 2024. https://doi.org/10.19678/j.issn.1000-3428.0068856
15. Qian, Z. Y., Cao, Y. H., Shi, Z. K., Qiu, L. Y. A Semantic Segmentation Method for Remote Sensing Images Based on Deeplab v3. Proceedings of 2021 2nd International Conference on Big Data & Artificial Intelligence & Soft-ware Engineering (ICBASE 2021), Zhuhai, China, Sep-tember 24-26, 2021, 396-400. https://doi.org/10.1109/ICBASE53849.2021.00080 https://doi.org/10.1109/ICBASE53849.2021.00080
16. Shen, Y. X., Sun, X. Y., Cui, J. J., Lu, Y. Application of Pyramid Scene Parsing Network in Leaf Segmentation for Wheat Stripe Rust. Proceedings of 2024 5th International Conference on Computer Vision, Image and Deep Learn-ing (CVIDL 2024), Zhuhai, China, April 19-21, 2024, 926-930. https://doi.org/10.1109/CVIDL62147.2024.10604070
17. Tayeb, A. M., Kim, T. H. UNestFormer: Enhancing De-coders and Skip Connections with Nested Transformers for Medical Image Segmentation. IEEE Access, 2024, 12, 190996-191009. https://doi.org/10.1109/ACCESS.2024.3516079
18. Tian, T., Wang, D., Wang, Z., Li, H. B. Precise Classifica-tion of Crops in Complex Planting Structure Area Based on Deep Learning Model. Chinese Journal of Agricultural Resources and Regional Planning, 2022, 43(12), 147-158. https://www.cnki.com.cn/Article/CJFDTotal-ZGNZ202212016.htm
19. Venkataramanan, S., Ghodrati, A., Asano, Y. M., Porikli, F., Habibian, A. Skip-Attention: Improving Vision Trans-formers by Paying Less Attention. 2023. https://doi.org/10.48550/arXiv.2301.02240
20. Wang, Z. Y., Xie, X. M., Yang, J. X., Shi, G. M. Soft Focal Loss: Evaluating Sample Quality for Dense Object Detec-tion. Neurocomputing, 2022, 480, 271-280. https://doi.org/10.1016/j.neucom.2021.12.102

21. Xie, E. Z., Wang, W. H., Yu, Z. D., Anandkumar, A., Al-varez, J. M., Luo, P. SegFormer: Simple and Efficient De-sign for Semantic Segmentation with Transformers. Ad-vances in Neural Information Processing Systems, 2021, 34, 12077-12090. https://arxiv.org/abs/2105.15203

22. Yuan, L., Chen, Y. P., Wang, T., Yu, W. H., Shi, Y. J., Jiang, Z. H., Tay, F. E., Feng, J. S., Yan, S. C. Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet. Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV 2021), Montreal, QC, Canada, October 10-17, 2021, 538-547. https://doi.org/10.1109/ICCV48922.2021.00060

23. Zhang, L., Liu, C. H., Shi, L. F., Zhang, Y. Tobacco Plant-ing Information Extraction Based on U-Net Neural Net-work. Agriculture and Technology, 2021, 41(22), 44-47. https://www.cnki.com.cn/Article/CJFDTotal-NYYS202122014.htm

24. Zheng, S. X., Lu, J. C., Zhao, H. S., Zhu, X. T., Luo, Z. K., Wang, Y. B., Fu, Y. W., Feng, J. F., Xiong, T., Torr, P. H., Zhang, L. Rethinking Semantic Segmentation from a Se-quence-to-Sequence Perspective with Transformers. Pro-ceedings of 2021 IEEE/CVF Conference on Computer Vi-sion and Pattern Recognition (CVPR 2021), Nashville, TN, USA, June 20-25, 2021, 6877-6886. https://doi.org/10.1109/CVPR46437.2021.00681