remote sensing

MDPI

Article

Crop Classification from Drone Imagery Based on Lightweight Semantic Segmentation Methods

Zuojun Zheng [1,2], Jianghao Yuan [1,2,3], Wei Yao [2], Hongxun Yao [4], Qingzhi Liu [5] and Leifeng Guo [2,*]

[1] College of Information Science and Technology, Hebei Agricultural University, Baoding 071001, China
[2] Institute of Agricultural Information, Chinese Academy of Agricultural Sciences, Beijing 100081, China
[3] Academy of National Food and Strategic Reserves Administration, Beijing 100039, China
[4] School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China
[5] Information Technology Group, Wageningen University and Research, 6700 HB Wageningen, The Netherlands
* Correspondence: guoleifeng@caas.cn

Abstract: Technological advances have dramatically improved precision agriculture, and accurate crop classification is a key aspect of precision agriculture (PA). The flexibility and real-time nature of UAVs have led them to become an important tool for acquiring agricultural data and enabling precise crop classification. Currently, crop identification relies heavily on complex high-precision models that often struggle to provide real-time performance. Research on lightweight models specifically for crop classification is also limited. In this paper, we propose a crop classification method based on UAV visible-light images based on PP-LiteSeg, a lightweight model proposed by Baidu. To improve the accuracy, a pyramid pooling module is designed in this paper, which integrates adaptive mean pooling and CSPC (Convolutional Spatial Pyramid Pooling) techniques to handle high-resolution features. In addition, a sparse self-attention mechanism is employed to help the model pay more attention to locally important semantic regions in the image. The combination of adaptive average pooling and the sparse self-attention mechanism can better handle different levels of contextual information. To train the model, a new dataset based on UAV visible-light images including nine categories such as rice, soybean, red bean, wheat, corn, poplar, etc., with a time span of two years was created for accurate crop classification. The experimental results show that the improved model outperforms other models in terms of accuracy and prediction performance, with a MIoU (mean intersection ratio joint) of 94.79%, which is 2.79% better than the original model. Based on the UAV RGB images demonstrated in this paper, the improved model achieves a better balance between real-time performance and accuracy. In conclusion, the method effectively utilizes UAV RGB data and lightweight deep semantic segmentation models to provide valuable insights for crop classification and UAV field monitoring.

Keywords: precision agriculture; UAV remote sensing; deep learning; lightweight; crop classification

check for updates

Citation: Zheng, Z.; Yuan, J.; Yao, W.; Yao, H.; Liu, Q.; Guo, L. Crop Classification from Drone Imagery Based on Lightweight Semantic Segmentation Methods. *Remote Sens.* **2024**, *16*, 4099. https://doi.org/10.3390/rs16214099

Academic Editor: Guido D'Urso

Received: 11 July 2024
Revised: 18 October 2024
Accepted: 30 October 2024
Published: 2 November 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

With progress in science and technology, the arrival of Industry 4.0 is pushing agriculture toward Agriculture 4.0, and precision agriculture (PA) is one of its hallmarks [1]. The continuous maturation of technologies such as remote sensing, artificial intelligence, Internet of Things (IoT) and big data has facilitated the development of precision agriculture technologies. Among all agricultural technologies, precise crop classification is crucial, with applications ranging from monitoring crop acreage and growth to crop yield estimation and precise crop management. Therefore, realizing high-precision crop classification has become a hot research topic today.

High-precision crop classification typically requires large amounts of agricultural data to train deep learning (DL) models. Rapid access to crop data has long been a challenge for research organizations and industrial companies worldwide. The emergence of satellite

remote sensing seems to have broken through this bottleneck, as satellites can be used to acquire large amounts of crop information [2]. However, satellite remote sensing is limited by cloud cover and cannot provide high-resolution images. The development of UAV remote sensing technology has brought about a change in data acquisition methods. UAV remote sensing has many advantages such as high flexibility, low cost, real-time performance, and the ability to carry different sensors to acquire high-resolution images [3].

In order to fully utilize the rich feature details in the high-resolution remote sensing data acquired by UAVs and to improve the classification accuracy, researchers have tried object-oriented classification methods such as machine learning (ML) algorithms like support vector machines [4], random forests [5], and BP neural networks [6]. Although the overall process of different segmentation methods varies, all of them require careful selection of the optimal segmentation parameters, which mainly relies on repeated experimental comparisons. Therefore, these traditional solutions are generally complex, cumbersome, and have relatively low classification accuracy. DL algorithms have emerged as effective solutions for many agricultural applications, enabling agricultural growers and owners to make accurate decisions at the right time [7]. In contrast to classical machine learning algorithms, DL models can be trained in an end-to-end manner, learning directly from the raw input data to the target output without the need to manually design complex feature transformations or intermediate steps. This simplifies the model design and training process, leading to better utilization of data. As a result, the application of DL algorithms in crop classification tasks has increased significantly in recent years [8].

Researchers have proposed several effective image segmentation techniques that can be used for crop type classification. J. Long et al. [9] pioneered the CNN-based image segmentation technique, known as Full Convolutional Networks (FCNs), which marked the birth of semantic segmentation. However, the maximum pooling layer of FCNs reduces the spatial resolution and may lead to the loss of valuable information about different crops or plants. Therefore, more sophisticated algorithms such as U-Net [10], SegNet [11], and DeepLab [12] have been proposed in recent years, which improve the crop classification results.

Kattenborn, T et al. [13] tested how to combine a U-Net model with training data obtained from visual interpretation to achieve stable and fine-grained mapping of vegetation species and communities in high-resolution UAV data. The results showed that the method could segment and map vegetation species and communities with an accuracy of 84%. Yang et al. [14] used FCN-Alex Net and SegNet semantic segmentation models combined with UAV visible-light images to map rice pests. The results show an F1 score of 0.8, which reduces the computation time by 10–15 times compared to the traditional maximum likelihood classification (MLC). Morales et al. [15] used an end-to-end trainable convolutional neural network (CNN) based on the DeepLab V3+ architecture for semantic segmentation of drone RGB images of palm tree canopies in the Amazon rainforest, with an overall accuracy of 98.14%. Zhong et al. [16] proposed a deep convolutional neural network framework with Conditional Random Field Classifier (CNNCRF) for accurate classification of multiple crops using high-spatial-resolution hyperspectral images acquired by UAVs. The overall accuracy was more than 85%. Anul Haq, M et al. [17] introduced a novel CNNLVQ model for weed detection with an overall accuracy of 99.44%. In conclusion, many researchers have demonstrated that DL-based semantic segmentation models combined with UAV remote sensing images can accurately classify various crops.

In recent years, with the advancement of computer hardware and the continuous improvement in remote sensing image resolution, there is a gradual increase in the demand for real-time performance. The high-precision semantic segmentation models such as U-Net, SegNet, and DeepLab mentioned above can achieve accurate classification of crops. However, all of these solutions are based on a large amount of data and long-term training. Moreover, these complex algorithmic structures and solutions require the support of high-performance computational hardware, which highly increases the deployment and usage cost. To cope with this challenge, lightweight semantic segmentation models [18] are

proposed for achieving real-time classification. In this paper, we propose a lightweight semantic segmentation model that integrates pyramid pooling and self-attention mechanisms for precise identification of crops in drone-visible-light data. In addition, we have established a drone visible-light dataset comprising nine categories, including rice, soybeans, red beans, wheat, corn, poplar trees, roads, coniferous trees, and background. Furthermore, we compared our improved model with several other lightweight models that currently exhibit good performance. Experimental results demonstrate that, in comparison to other lightweight models, the model proposed in this paper achieves superior accuracy and prediction performance while maintaining its lightweight characteristics.

The remainder of this paper consists of four chapters. Section 2 introduces the research area and the self-built UAV visible-light dataset. Section 3 details the proposed spatial pyramid pooling module, the lightweight model, experimental results, and analysis, and Section 4 provides discussion and analysis. Finally, Section 5 summarizes the conclusions and outlines future work.

2. Material and Methods

2.1. Study Area and UAV Parameters

The study area is located in Suibin County, Hegang City, Heilongjiang Province, China. The area is located in the delta where the Heilongjiang River meets the Songhua River (latitude: 47°35′37.74″N, longitude: 132°0′40.54″E) (shown in Figure 1). The region has a continental cold-temperate climate, with an average annual temperature of about 3 °C. Rainfall is abundant, with an average annual precipitation of about 502.5 mm. The main soil types are meadow soils and brown forest soils, which are fertile and well vegetated. Agriculture is the main industry in the region, with rice, soybeans, corn, wheat and some cash crops being grown; in 2009, the total production of grains and beans amounted to 370,000 tons, with a per capita income of USD 2814. The region has more than 6000 sets of large agricultural machinery and equipment, and the level of comprehensive mechanization has reached more than 96 per cent.

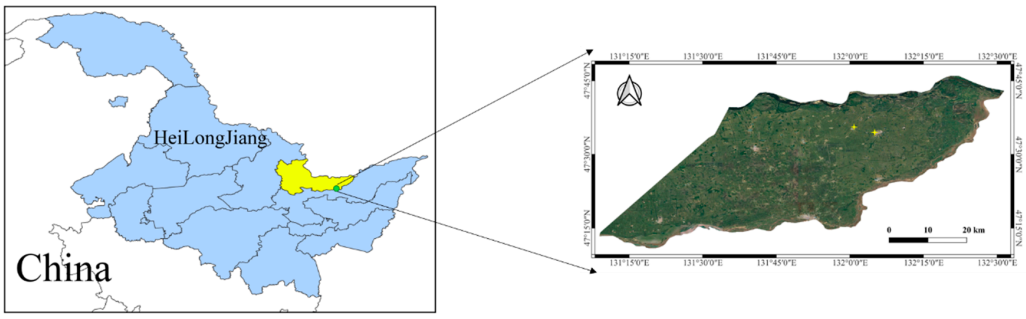
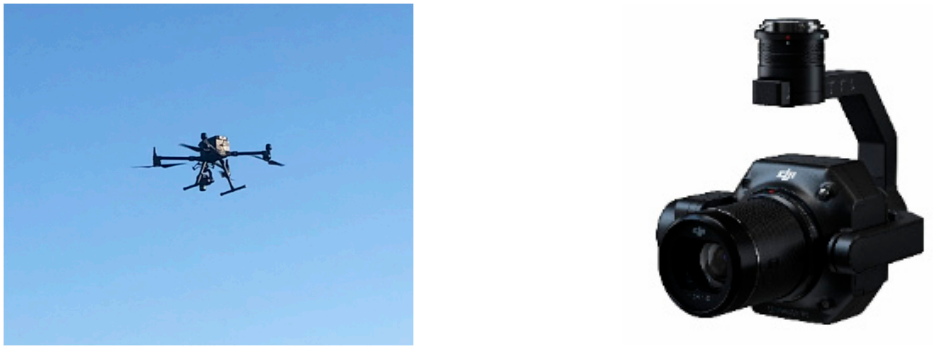


Figure 1. Schematic of the location of the study area.

The data collection in this study was conducted using the DJI M300 RTK industrial-grade surveying and inspection UAV (Figure 2a), equipped with the DJI Zen muse P1 camera (Figure 2b). The specific parameters of the UAV and camera are detailed in Table 1.

Table 1. Technical specifications of the UAV and camera.

Device	Technical Parameters	Specific Values
UAV	Total Weight (with Battery and Lens)	7.1 kg
	Maximum horizontal flight speed (automatic mode)	17 m/s
	Maximum flight time	55 min
	Symmetrical motor wheelbase	895 mm
Lens	Airborne footage	DJI ZenithP1\35 mm\FOV 63.5°
	Image size	3:2 (8192 × 5460)
	Effective pixel	45 million
	Aperture range	f/2.8-f/16



(**a**) (**b**)

Figure 2. (**a**) Schematic diagram of the UAV. (**b**) Schematic diagram of the onboard camera. The drones and sensors are manufactured at DJI Innovation Technology Co. in Shenzhen, China.

The data collection for this study was conducted under clear weather conditions with plenty of sunlight and low wind speeds. The lateral overlap rate was set at 70%, the longitudinal overlap rate at 80%, and the flight speed was configured at 17 m per second.

2.2. UAV Visible Light Dataset

In this study, a new UAV visible light dataset was constructed and collected from July to September 2022 and July to September 2023, respectively. The data collection area included a large-scale farmland planting area and a rice science and technology planting area. The two experimental fields are approximately 8 km apart. Specific details about the dataset will be provided in the following sections.

2.2.1. Large-Scale Farmland Cultivation Areas

The total area of extensive farmland cultivation is approximately 200 hectares. The main crops include rice, soybeans, corn, wheat, and red beans. Poplar trees are planted as a separating zone between different crops, and coniferous forests are planted as windbreaks. The distribution of crops is shown in Figure 3. The specific growth cycles of the five crops are shown in Figure 4.

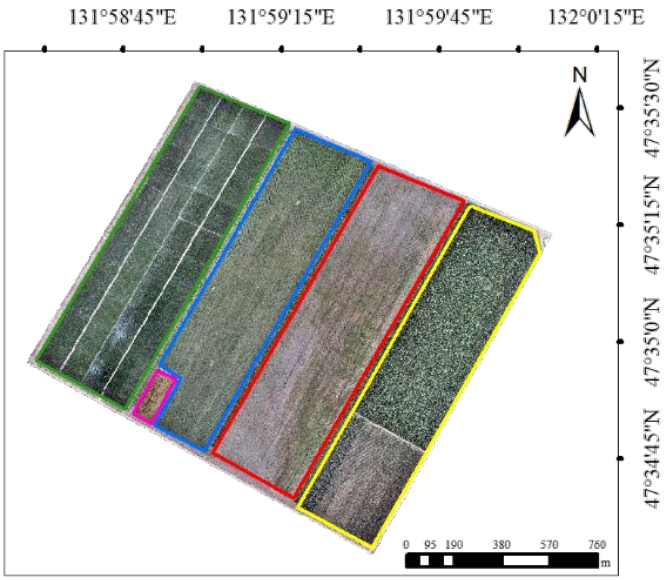


Figure 3. Crop distribution map of the field experiment zone. Rice planting area. Soybean planting area. Wheat planting area. Red bean planting area. Corn planting area.

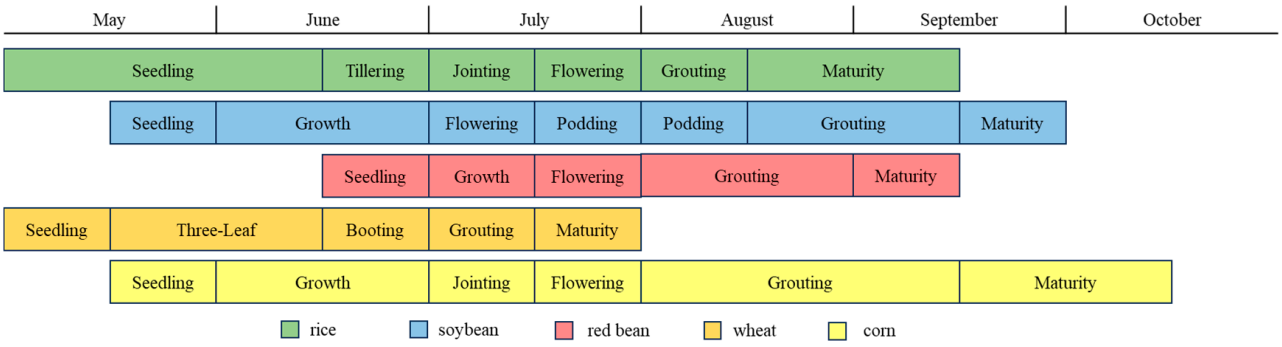


Figure 4. Phenological stages of crops.

UAV RGB images of the farmland were collected on 6 and 14 July 2022, and on 11 July 2023. The specific phenological stages of each crop in the farmland at the time of collection are detailed in Table 2.

Table 2. Phenological Stages of Crops.

Crop	Phenological Stage
Rice	Jointing and tasseling stage
Red bean	Late nutritional growth and development
Wheat	Grouting period
Soybean	Flowering and podding period
Corn	Late phenological stage

On 6 and 14 July 2022, we collected RGB images at a flight altitude of 40 m for wheat planting areas. For rice planting areas, the camera was tilted at a 45 degrees angle, and RGB images were captured at a flight altitude of 30 m. For corn planting areas, RGB images were captured at a flight altitude of 40 m. Due to tiny visual changes in the appearance of corn within a week, data for 14 July were not collected. Finally, we adjusted the flight altitude to 450 m and captured RGB images of the entire farmland. We used images obtained under this height, which clearly lacked clarity, to test the model's ability to handle such images. On 11 July 2023, we collected RGB images of the entire farmland at a flight altitude of 160 m. The specific data are presented in Table 3.

Table 3. Data acquisition in the Daejeon experimental area.

Acquisition Area	Camera Angle	Aerial Heights/m	GSD/cm	Date	Number of Images
Wheat growing area	Top shot 90°	40	0.5	6 July 2022 14 July 2022	425 426
Experimental area as a whole	Top shot 90°	450	5.625	6 July 2022 14 July 2022	362 431
Maize growing area	Top shot 90°	40	0.5	6 July 2022	968
Rice-growing areas	Tilt 45°	30	0.375	6 July 2022 14 July 2022	303 305
Experimental area as a whole	Top shot 90°	160	2	11 July 2022	2111

2.2.2. Data Acquisition at Rice Technology Park

The science and technology park (STP) is a small-scale experimental area for rice multi-modal planting technology, aiming to test rice yield under different planting modes and different varieties. Therefore, we can obtain rice images of multiple varieties to enhance the richness of the data. Moreover, the rice growth cycle in the STP was consistent with that in the large-scale farmland planting area. The STP was only 8 km away from the large-scale

farmland planting area, and the climatic conditions and soil and water conditions were highly similar. In this study, we took RGB images on 9 July 2023 at a flight altitude of 30 m in STP. A total of 44 different rice varieties, 4 different fertilization levels, and 4 different planting densities were included. The crop distribution in the park is shown in Figure 5, and the specific data are detailed in Table 4.



Figure 5. Distribution of science and technology parks: ☐ 44 varietal growing areas; ☐ 4 fertilizer application rates and planting density zones.

Table 4. Data on science and technology parks.

Acquisition Area	Camera Angle	Aerial Heights/m	GSD/cm	Date	Number of Images
Rice technology park	Top shot 90°	30	0.375	9 July 2023	1941

2.3. Data Pre-Processing

The data sample categories include rice, soybean, wheat, red bean, maize, poplar forest, coniferous forest, roads, and background (unused fields, dirt roads, paddy ridges, and canals), for a total of nine categories (Figure 6b). Among them, canals may contain both water and dry portions. There is no difference between the dry portion and the unused field (as shown in Figure 6a). Therefore, the drains are classified as background.

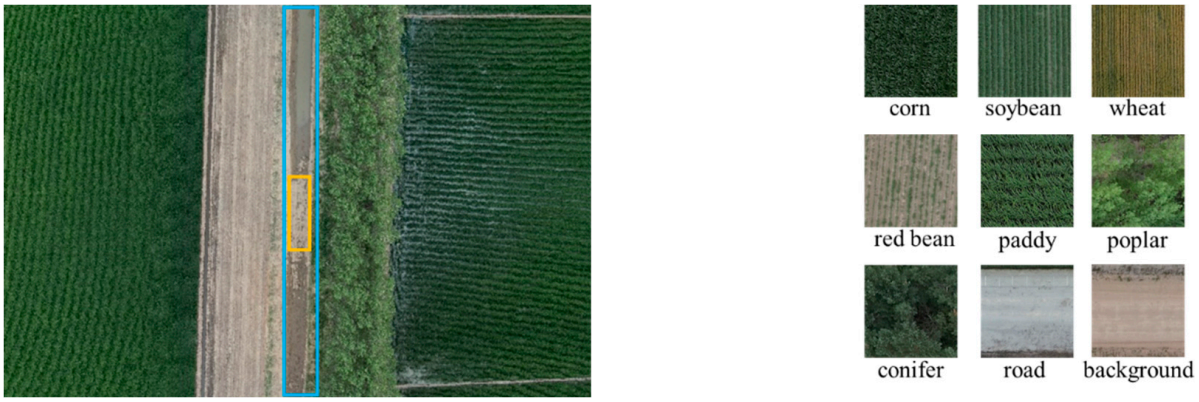


Figure 6. Sample data set. (**a**) Complete Aqueduct. Dry drains. (**b**) Examples of study area samples.

In this study, the rice science and technology park dataset and the data collected on 11 July 2023 from a large-scale agricultural planting area were selected as the training set for the model. The remaining data were used for testing. Since the dataset of the science and technology park mainly consists of rice samples, the rice portion of the data from the large-scale farmland planting area was singled out for testing in order to avoid an over-representation of rice samples.

The repetition rate of neighboring images was 70–80%. To prevent excessive repetition of pixel points in the training set, one pixel point was selected every four images along the same longitudinal flight line. For longitudinal flight lines, one is selected every other image, as shown in Figure 7, where the red and yellow dashed lines represent the selected flight lines. To avoid overrepresentation of crop samples, pure crop images were removed (as shown in Figure 8a). Then, semantic annotation was performed using the LabelMe plug-in (shown in Figure 8b), followed by splitting the original and labeled images into small images of 512 × 512-pixel size (shown in Figure 8c). Finally, 23,681 small images are obtained, which are divided into a training set and validation set in the ratio of 8:2. During the training process, data preprocessing techniques such as RandomDistort, RandomPaddingCrop, and RandomHorizontalFlip are used to reduce the occurrence of overfitting. Table 5 lists the proportion of pixel points in each sample of the training set. The remaining images are used for model prediction.

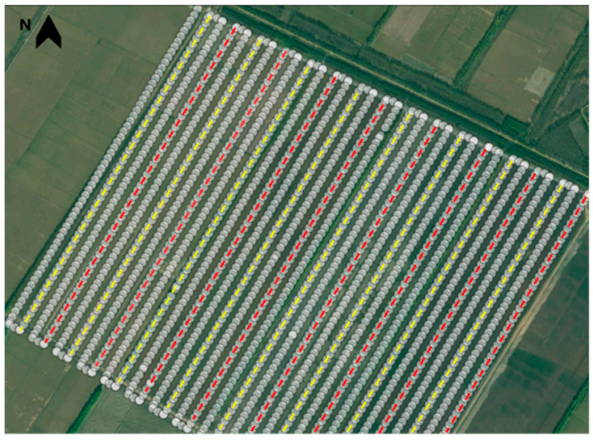


Figure 7. Schematic diagram of UAV flight lines.

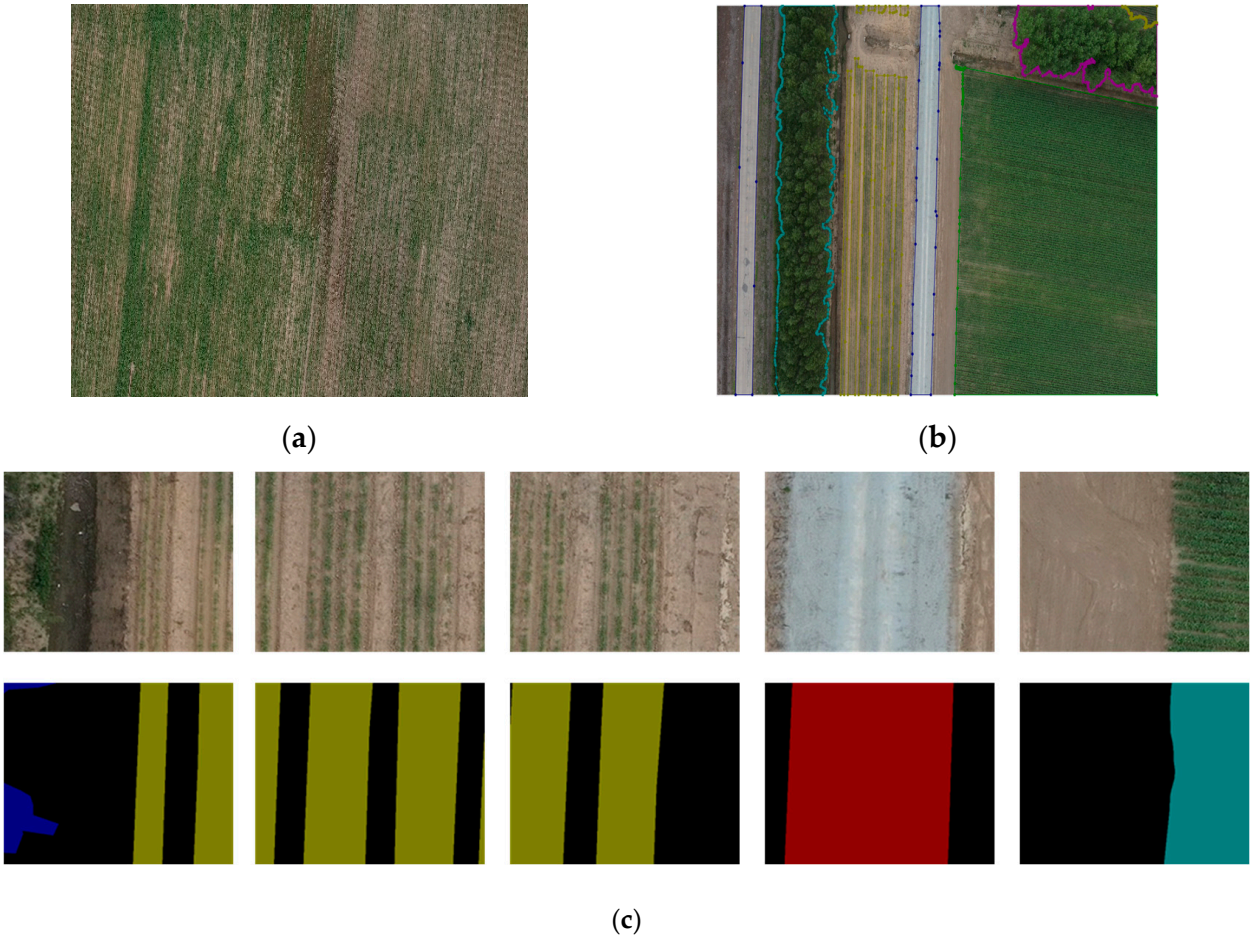


Figure 8. Sample dataset. (**a**) Red-bean-growing area. (**b**) Labeling example. (**c**) Example of a small picture.

Table 5. Percentage of sample pixel points.

Serial Number	Classification	Pixel Count	Percentage
0	background	636,663,905	10.25%
1	road	62,928,541	1.01%
2	rice	3,889,723,226	62.64%
3	red bean	371,221,364	5.98%
4	coniferous tree	61,888,531	1.00%
5	wheat	51,259,191	0.83%
6	soybean	447,241,100	7.20%
7	poplar tree	177,057,377	2.85%
8	corn	511,421,695	8.24%

The data were collected in mid to late July, a critical period for the growth and development of most crops. Except for red beans and poplar trees, the leaves of the other crops can cover the land. At this time, red beans are in the early stage of development, with relatively small leaves, leading to some land being exposed between plants (Figures 6b and 8a). This can easily be confused with background samples (idle fields). The edges between poplar trees and coniferous trees are highly cluttered, and there is often exposed land between the tree leaves. Roads, trees, and wheat samples are significantly less numerous compared to other crop samples, while the proportion of rice samples is excessively high. All these factors pose significant challenges for semantic segmentation.

2.4. Lightweight Crop Classification Model

2.4.1. Basic PP-Liteseg Model

The PP-Liteseg model introduces a Flexible Lightweight Decoder (FLD) to reduce the computational overhead of conventional decoders. To enhance the feature representation, the model proposes the Unified Attention Fusion Module (UAFM), which utilizes spatial and channel attention to generate weights and then fuses the input features with these weights. In addition, the Simple Pyramid Pooling Module (SPPM) is proposed to pool the global context at a low computational cost [19].

As shown in Figure 9, in traditional encoder-decoder architectures, the feature channels of the decoder remain the same. However, FLD gradually reduces the feature channels from high-level to low-level, alleviating the redundancy of the decoder, balancing the computational costs of the encoder and decoder, and making the model more lightweight.



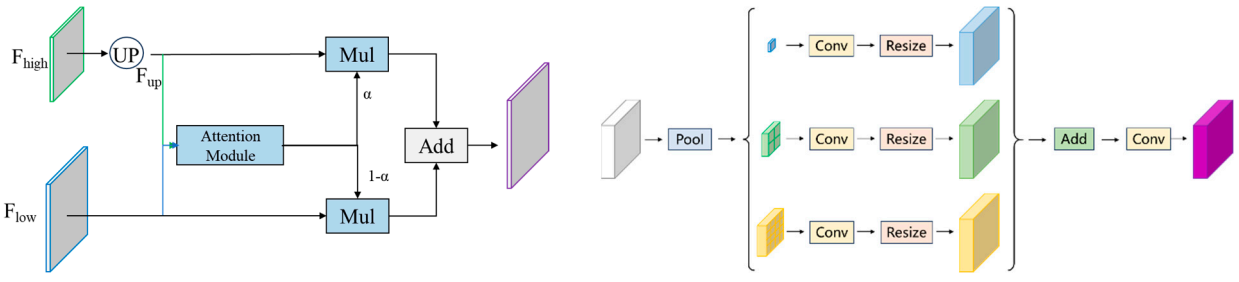
(**a**) (**b**)

Figure 9. (**a**) Traditional encoder-decoder architecture. (**b**) Encoder with lightweight decoder (FLD).

The Unified Attention Fusion Module (UAFM) (Figure 10a) utilizes an attention module to generate weights α, which are then fused with the input features through Mul and Add operations. The specific process can be expressed by Equation (1).

$$\begin{aligned} F_{up} &= \text{Upsample}\left(F_{high}\right), \\ \alpha &= \text{Attention}\left(F_{up}, F_{low}\right), \\ F_{out} &= F_{up} \cdot \alpha + F_{low} \cdot (1 - \alpha) \end{aligned} \tag{1}$$

where F_{high} represents the output of the deep module, F_{low} denotes the output of the encoder, F_{up} is the upsampled feature, and α signifies the weights generated by the attention module.



(**a**) (**b**)

Figure 10. (**a**) Structure of Unified Attention Fusion Module (UAFM). (**b**) Structure of Simple Pyramid Pooling Module (SPPM).

The Simple Pyramid Pooling Module (SPPM) (Figure 10b) first utilizes the pyramid pooling module to fuse input features; then, after the output features, convolution and up sampling operations are performed. Finally, the up sampled features are convolved to generate refined features. Compared to the original PPM, SPPM reduces the intermediate and output channels, removes the shortcut, and replaces the concatenation operation with an addition operation.

The PP-Liteseg model achieves a class IoU of 72.0% and fps of 273.6 on the cityscapes dataset [19], demonstrating excellent trade-offs between accuracy and speed. However, further refinement of this network architecture is required to achieve superior results in specific application domains. To achieve better results, we will introduce improved methods in the following sections.

2.4.2. Methods for Enhancing PP-Liteseg Model

The unified attention fusion module (UAFM) combines spatial and channel attention modules. However, when handling relationships between non-adjacent regions in images, it may unevenly handle the importance of different features. This can lead to some features being ignored or assigning excessively large weights to certain features. While the simple pyramid pooling module (SPPM) is more lightweight than traditional PPM, it may lead to insufficient multi-scale information, inadequate semantic information, and decreased accuracy. Particularly when dealing with complex scenes and multi-scale objects, it can affect the model's performance.

Combining the characteristics of our dataset with the aim of enhancing model performance, we opt for self-attention (SA) [20] and spatial pyramid pooling (SPP).

2.4.3. A Novel Spatial Pyramid Pooling Module

In traditional convolutional neural networks (CNNs), pooling layers are commonly used to reduce the size of feature maps and extract higher-level features. However, conventional pooling layers typically require input images to have fixed sizes, limiting the network's adaptability to inputs of varying sizes. To address this challenge, He, K et al. [21] proposed the spatial pyramid pooling (SPP) module, as illustrated in Figure 11. Its core idea is to partition the input feature map into a hierarchical grid and perform pooling operations on each grid, followed by concatenating all pooling results to generate a fixed-length feature representation. Consequently, regardless of the input image's size, the SPP module can produce feature vectors of fixed sizes, enhancing the network's robustness to inputs of different dimensions.

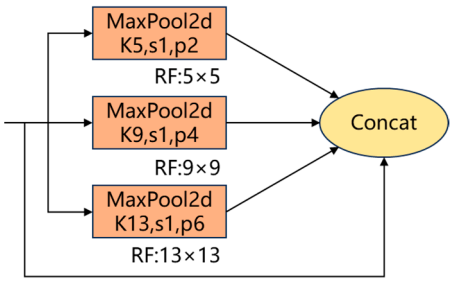


Figure 11. Structure of spatial pyramid pooling (SPP).

Based on the SPP module, Glenn Jocher, the author of YOLOv5, proposed a faster module called SPPF (Figure 12). This module adopts a simplified structure with only one max-pooling layer instead of multi-scale pooling operations. By employing this simpler structure, computational complexity is reduced, leading to improved forward propagation speed.

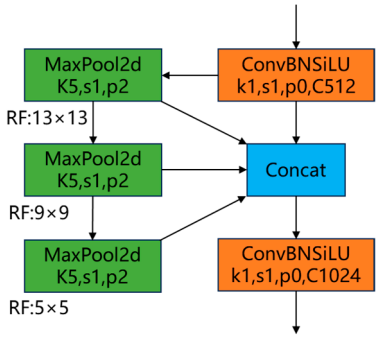


Figure 12. Structure of SPPF module.

Wang, C et al. [22] combined SPP with CSPC to design an SPPCSPC pooling module (Figure 13). CSPC is a technique aimed at improving feature propagation and network efficiency. In traditional CNNs, feature propagation occurs through serial connections, leading to bottlenecks in information flow. However, CSPC accelerates feature propagation by introducing inter-stage partial connections in the network, thereby better leveraging information between lower and higher-level features.

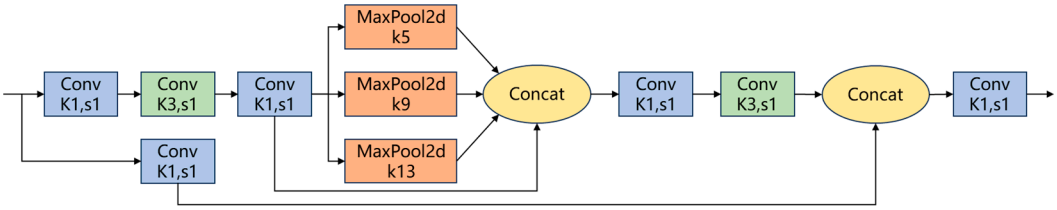


Figure 13. Structure of SPPCSPC module.

Inspired by the SPPF module, our team integrated CSPC with SPPF and replaced max pooling with adaptive average pooling (AAP), proposing a pooling module more suitable for lightweight semantic segmentation models, named SPPF-CSPC-A (Figure 14). This module effectively integrates and optimizes features of different scales by combining AAP, SPP techniques, and CSPC, thereby enhancing the feature extraction and contextual awareness capabilities of CNNs in semantic segmentation tasks.

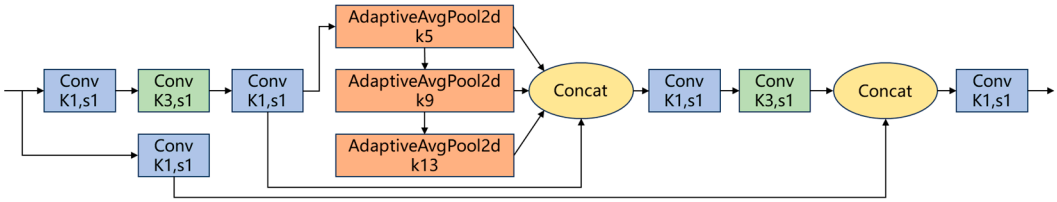


Figure 14. Structure of SPPF-CSPC-A module.

The SPPF-CSPC-A module comprises several key components: initially, multiple convolutional layers are employed for non-linear feature extraction and enhancement of the input feature maps. Subsequently, an adaptive pooling layer is utilized to resize multi-scale feature maps to a unified output size; the mathematical expression for adaptive average pooling is as follows:

$$y_c(i,j) = \frac{1}{h' \times w'} \sum_{m=0}^{h'-1} \sum_{n=0}^{w'-1} x_c(i \times h + m, j \times w + n) \quad (2)$$

In this context, x_c denotes the channel c of the input feature map, y_c represents the channel c of the output feature map, h' and w' are the height and width, respectively, of the output feature map, and h and w are the height and width of the input feature map.

Subsequently, diverse receptive field representations are obtained through spatial pyramid pooling techniques. Finally, integrating dynamic attention mechanisms adjusts the importance of features, further enhancing the accuracy and robustness of segmentation tasks. Overall, this module not only extends the feature representation capability of existing convolutional neural networks but also enhances the model's understanding of complex scenes, providing an effective method for addressing practical image analysis challenges.

2.4.4. Sparse Self-Attention

Due to the excessively trivial edges of poplar samples in our dataset (Figure 15), we employ self-attention (SA) to dynamically allocate attention weights between different positions. In this way, the model can better understand the relationship between object edges and surrounding environments, thus improving edge recognition accuracy. To maintain model lightweight, we opt for sparse self-attention (SSA) mechanism [23], a lightweight variant of self-attention mechanism aimed at reducing computational costs. Its core idea is to only consider elements that are relatively close in distance to the current position, thereby reducing computational complexity. In the implementation process, we follow these steps:

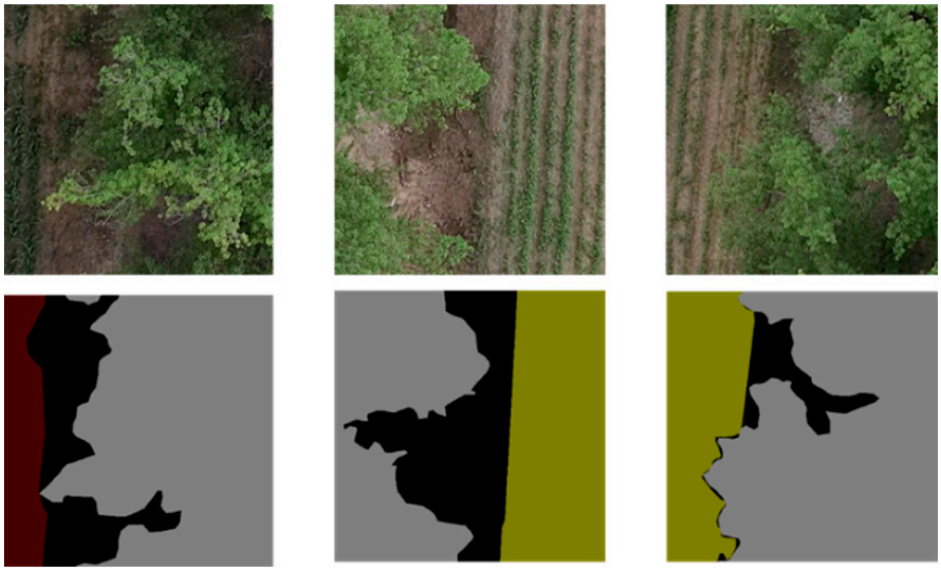


Figure 15. Original image and annotated image of poplar tree sample.

① Given a sequence, $X = \{X_1, X_2, ..., X_n\}$, where X_i represents the allocation at position i.

② Calculate similar dimensions:

$$\text{Similarity}(x_i, x_j) = x_i \cdot x_j^T \tag{3}$$

Here, x_i and x_j represent the representation vectors at the i-th and j-th positions, respectively.

③ Calculate the attention weights:

$$\text{Attention}(x_i, x_j) = \begin{cases} \frac{\exp(\text{Similarity}(x_i, x_j))}{\sum_{k=1}^{n} \exp(\text{Similarity}(x_i, x_k))} & \text{if} |i-j| \leq max_\text{distance} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where, exp denotes the exponential function, $\sum_{k=1}^{n}$ signifies the summation over all positions, and max_distance is a parameter used to specify the maximum relative distance under consideration.

④ Weighted summation:

$$\text{SparseSelfAttention}(X) = \sum_{i=1}^{n} \text{Attention}(x_i, x_j) \cdot x_j \tag{5}$$

The improved PP-Liteseg structure proposed in this paper is depicted in Figure 16. The model adopts STDCNet [24] as its backbone network, which consists of 5 stages, each with a stride of 2, yielding feature maps at 1/32 of the original image size.

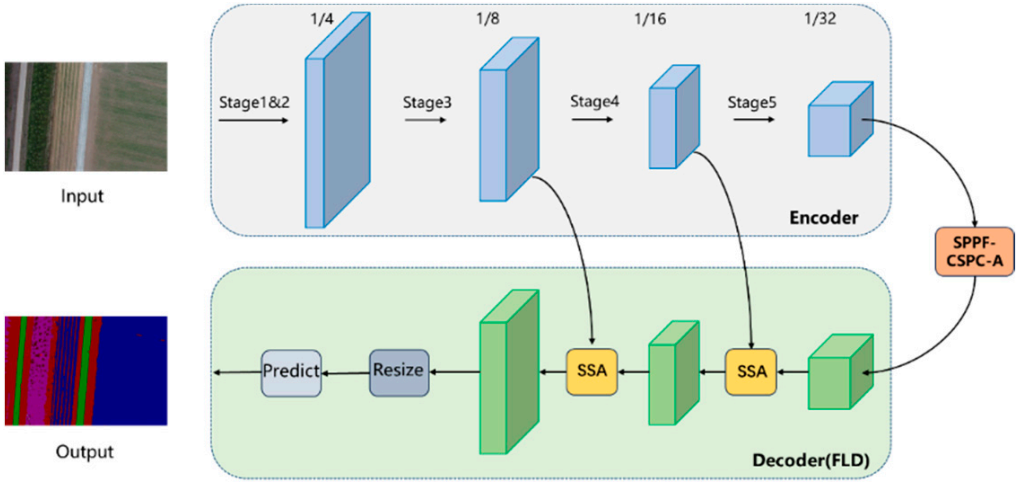


Figure 16. Crop classification model based on the improved PP-Liteseg.

2.5. Model Training Settings and Evaluation Metrics

This experiment utilizes an Intel i9-12700 CPU @ 3.20GHz and an NVIDIA GeForce RTX 3090 GPU with 32 GB of VRAM. Using PaddlePaddle as a deep learning framework, the specific hyperparameter settings are as shown in Table 6. We select accuracy (*Acc*), mean Intersection over Union (*MIoU*), and kappa coefficient to evaluate the precision of the model. FLOPs and FPS are chosen to assess whether the model is sufficiently lightweight.

Table 6. Hyperparameters.

Hyperparameters	Specific Parameters
Epoch	150
Optimizer	SGD
Lr_scheduler	PolynomialDecay
Learning_rate	0.01

Accuracy (*Acc*): the proportion of the number of correct predictions in positive and negative cases to the total number, expressed by the following formula:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

Among them, *TP* represents true positives, *FP* represents false positives, *FN* represents false negatives, and *TN* represents true negatives.

MIoU is the average value of each type of IoU, and IoU is the ratio of the intersection and concatenation of the predicted and true values, and *MIoU* is calculated by the following formula:

$$MIoU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k}p_{ij}+\sum_{j=0}^{k}p_{ji}-p_{ii}} \tag{7}$$

where p_{ij} represents predicting i as j, indicating false negatives (FN); p_{ji} represents predicting j as i, indicating false positives (FP); p_{ii} represents predicting i as i, indicating true positives (TP). Equation (6) is equivalent to the following:

The kappa coefficient [25] is a metric used for consistency testing and can also be used to measure the effectiveness of classification. Because for classification problems, the so-called consistency is whether the model prediction results and the actual classification results are the same. The kappa coefficient is calculated based on the confusion matrix, taking values between -1 and 1, usually greater than 0.

$$k = \frac{p_o - p_e}{1 - p_e} \tag{8}$$

Included among these, $p_o = ACC$, $p_e = \frac{\sum_i \text{Sum of elements in row } i \times \text{Sum of elements in column } i}{(\sum \text{All elements of the matrix})^2}$.

FLOPs refer to the number of floating-point operations required by a deep learning model during inference or training, used to measure the computational complexity of the model. For convolutional layers, we have

$$\text{FLOPs} = 2HW\left(C_{in}K^2 + 1\right)C_{out} \tag{9}$$

where H, W, and C_{in} represent the height, width, and number of channels of the input feature, K is the kernel width, and C_{out} is the number of output channels.

For fully connected layers, the method to calculate FLOPs is as follows:

$$\text{FLOPs} = (2I - 1)O \tag{10}$$

where I is the input dimension and O is the output dimension.

For semantic segmentation models, FPS refers to the number of images the model can process per second during image semantic segmentation tasks, reflecting the speed of the model when processing images in real-time or in batches.

As shown in Table 5, the dataset in this study exhibits significant class imbalance. To address this issue, we opt for OhemCrossEntropyLoss (Ohem) [26] as the loss function for model training. Ohem is specifically designed to mitigate class imbalance problems and is commonly employed in classification tasks. The underlying principle of Ohem is to only consider a subset of difficult-to-classify samples while disregarding easily classifiable ones when computing the loss. This allows the model to focus more on those challenging samples, thereby improving the issue of class imbalance. The specific implementation process is as follows:

$$\text{OhemCrossEntropyLoss} = -\frac{1}{M}\sum_{i=1}^{N}1(y_i \neq \text{ignore}) \cdot \log\left(\hat{y}_i^{(y_i)}\right) \tag{11}$$

where N represents the number of training samples; M denotes the number of difficult samples selected. y_i stands for the true label of the i-th sample. $\hat{y}_i$ represents the model's prediction for the i-th sample; $1(y_i \neq \text{ignore})$ is an indicator function, taking the value of 1 when the true label of the sample is not to be ignored, and 0 otherwise.

3. Results

3.1. Comparison of Models

In this study, several classical lightweight semantic segmentation models were selected for comparison. ENet [18] uses an early down sampling strategy to reduce the computational cost of processing large images and feature maps, with a category IoU of 80.4% on the cityscapes dataset, a class IoU of 58.3% and fps of 76.9. BiSeNetv1 [27] is a novel bilateral segmentation network that can quickly obtain a fairly large sensory field while preserving the spatial information of the original image, with a class IoU of 68.4% and fps of 105 on the cityscapes test dataset. BiSeNetv2 [28] introduces an effective context-guided fusion module for integrating information from spatial and contextual paths on top of the original bilateral segmentation network, resulting in faster inference with a class IoU of 72.6% and fps of 156 on the cityscapes dataset. ESPNetv2 [29] uses point-by-point and depth-by-expansion separable convolution to learn features from an expanded receptive field in a computationally tractable manner, with a class IoU of 66.4% and FLOPs of 2.7 B on the cityscapes dataset. The multi-branch network of Fast-SCNN [30] shares the computational cost of produce runtime fast segmented CNNs with class IoU of 68.0% and fps of 123.5 on the cityscapes dataset. STDCSeg [24] proposes a short-term densely connected module (STDC) for extracting deep features with scalable sensory fields and multiscale information on the cityscapes dataset with a class IoU of 77.0% and fps of 97. SegNeXt [31] is a novel convolutional attention network that uses cheap convolutional operations. The model introduces multi-scale convolutional attention in the encoder, and the decoder uses matrix decomposition for global airspace information modeling, with a class IoU of 78.0% and an fps of 25 on the cityscape dataset.

In addition, we selected several Transformer-based lightweight semantic segmentation models in recent years for benchmark comparison. TopFormer [32] proposed a new architecture combining CNNs and ViT and achieved a good trade-off between accuracy and computational cost in mobile semantic segmentation, with a class IoU of 75.0% and FLOPs of 11.2 G on the cityscapes dataset. SegFormer [33] improves the efficiency by introducing a hierarchical Transformer encoder and a lightweight full MLP decoder, with class IoU of 77.7% and FLOPs of 13.7 G on the cityscapes dataset. RtFormer [34] proposed a pairwise resolution module consisting of two types of attention and the feedforward networks arranged in a stepped layout, with a class IoU of 79.3% and fps of 110.0 on the cityscapes dataset. The performance of all the aforementioned models on the cityscapes dataset is presented in Table 7.

Table 7. The performance of lightweight models on the cityscapes dataset.

Model Category	Model	MIoU	FLOPs	FPS
Classical lightweight semantic segmentation models	Fast-SCNN	68.00%	2.7 B	123.5
	ENet	58.30%		76.9
	ESPNetv2	66.40%		
	BiSeNetv1	68.40%		105
	BiSeNetv2	72.60%		156
	STDCSeg	77.00%		97
	SegNeXt	78.00%		25
Transformer-based lightweight semantic segmentation models	TopFormer	75.00%	11.2 G	
	SegFormer	77.70%	13.7 G	
	RtFormer	76.30%		110
	PP-Liteseg	78.20%		102.6

We train all the aforementioned models on the same hardware device and framework, and adjust the hyperparameters of each model to the optimal parameters suitable for our dataset. The FPS is predicted by each model for 6400 512 × 512-pixel UAV visible-light images. The training results of all these models on our dataset are shown in Table 8.

Table 8. Training results of different models on the dataset of this paper.

Models	FOLPs	FPS	MIoU	Acc	Kappa
ENet	2.8 G	11.20	73.39%	96.58%	94.15%
BiSeNetv1	5.7 G	15.20	87.13%	97.05%	95.65%
ESPNetv2	2.8 G	12.70	91.11%	98.21%	96.95%
Fast-scnn	1 G	16.70	86.65%	97.38%	95.50%
BiSeNetv2	8.1 G	14.90	88.02%	97.84%	97.18%
STDCSeg	23.5 G	12.20	90.59%	97.67%	96.04%
SegNeXt	49.5 G	5.40	94.27%	98.89%	98.55%
TopFormer	1.6 G	12.00	81.31%	94.80%	93.21%
SegFormer	69.9 G	4.60	93.15%	98.70%	98.30%
RtFormer	16.9 G	11.10	93.64%	98.88%	98.21%
PP-Liteseg	9 G	13.30	92.00%	98.34%	97.84%
Ours	**9 G**	**12.30**	**94.79%**	**99.02%**	**98.72%**

Table 8 shows the training validation results of different semantic segmentation models on the training set. It is easy to see that the Transformer-based model has an eye-catching performance. However, the newly introduced traditional CNNs model SegNeXt is still not inferior, indicating that the traditional CNNs model can still achieve high-precision object classification. The model incorporating the proposed module in this paper achieved a 2.79% improvement in overall classification accuracy over the original model while maintaining lightweight characteristics, achieving the highest overall classification accuracy. In terms of prediction speed, our model is able to predict 512 × 512-pixel UAV visible light images at 12.3 FPS, which is in the middle of the range and meets the real-time requirement. Table 9 shows the accuracy of each category using the PA-Liteseg-SSA model. As mentioned before, the poplar and conifer trees with messy edges and the complex background are the important reasons negatively affecting the accuracy.

Table 9. Accuracy of each category.

Classification	Class IoU	Class Precision	Class Recall
Background	90.20%	95.02%	94.67%
Road	91.28%	93.98%	96.95%
Rice	99.49%	99.85%	99.64%
Red bean	97.93%	98.63%	99.28%
Coniferous tree	83.49%	88.86%	93.25%
Wheat	99.74%	99.87%	99.87%
Soybean	98.89%	99.42%	99.46%
Poplar tree	92.43%	96.27%	95.86%
Corn	99.68%	99.79%	99.89%

3.2. Prediction Results for the Dataset in the Daejeon Experimental Area

The visualization of the effect graphs provides a more intuitive representation of the specific performance of the models. Among the models described in the previous section, we choose the better-performing models SegNeXt and RtFormer, the original model PP-Liteseg, and the model PA-Liteseg-SSA for the comparison of prediction effects. Figure 17 shows the classification effects of the above four models, and the original image is the overall map of the experimental area taken on 11 July 2023 at 160 m aerial height.

First, as shown in the multiple prediction maps in Figure 17, the similarity between different crops resulted in a considerable number of misclassifications. For example, rice is misclassified as soybean, soybean is misclassified as red bean, and corn is misclassified as wheat. In addition, the original PP-Liteseg model exhibits significant pretzel noise as shown in Figure 17c. The model proposed in this paper integrates a global average pool and a self-attention mechanism, which can better utilize the global context information, effectively reduce the pretzel noise, and enhance the robustness of the model.

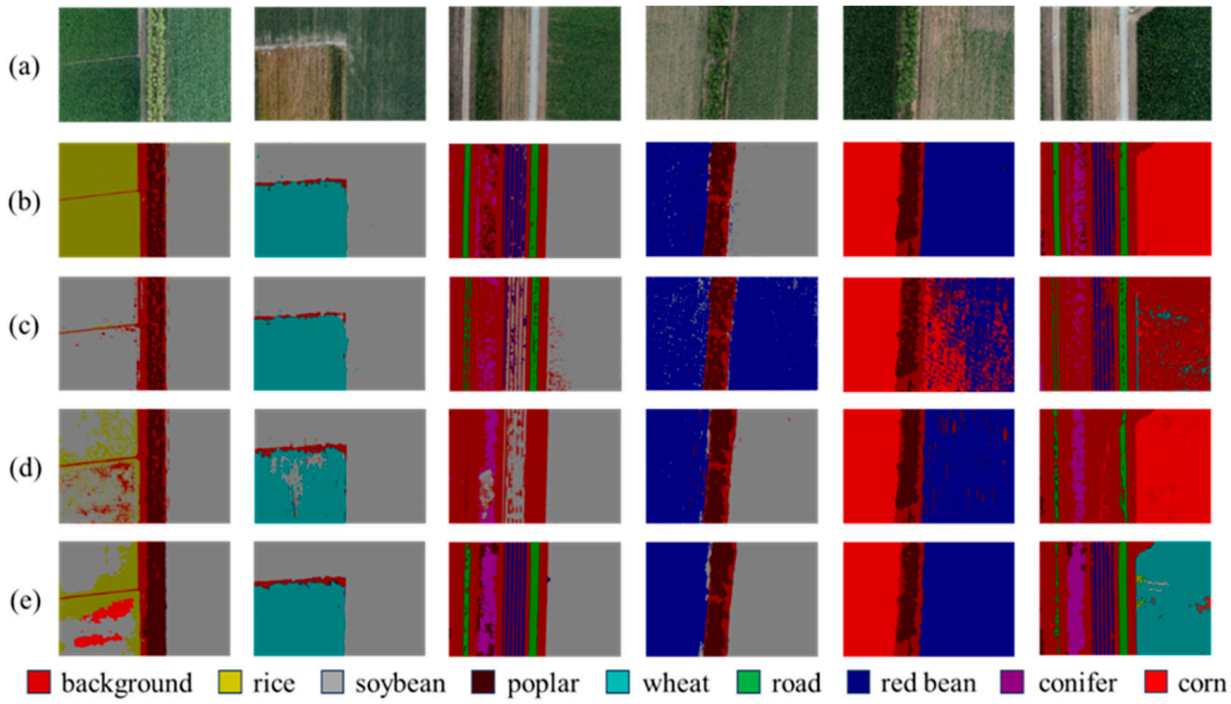


Figure 17. Comparison of the effectiveness of multi-model predictions. (**a**) Original image. (**b**) Ours. (**c**) PP-Liteseg. (**d**) SegNeXt. (**e**) RtFormer.

To assess the generalization ability of the model, we made predictions for the datasets acquired in 2022 listed in Table 3. As shown in Figure 18a–c, the data acquired by low-altitude (30–40 m) UAV remote sensing show clear images and rich details. However, the model incorrectly identified some weeds as rice due to the strikingly high similarity between the weeds growing on the paddy field ridges and the rice plants themselves. In addition, the limited training samples of wheat resulted in insufficient feature learning and some pretzel noise. As shown in Figure 18d, the data obtained through high-altitude (450 m) UAV remote sensing appeared relatively blurred. Nevertheless, our model still achieves satisfactory classification performance.

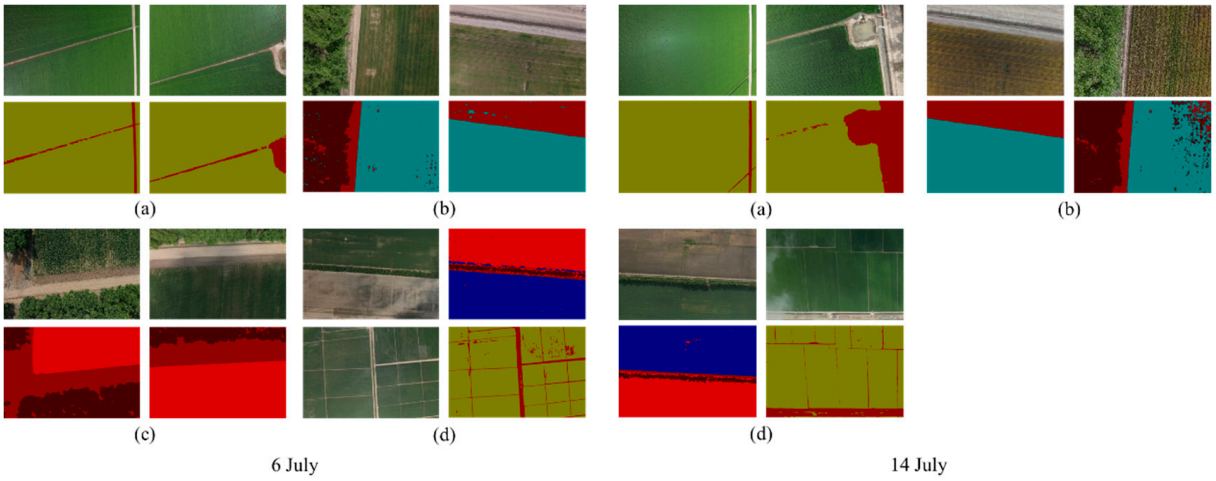


Figure 18. Predictive performance map of the field experimental area. Note: Corn data for 14 July were not collected. (**a**–**c**) the data acquired by low-altitude (30–40 m) UAV remote sensing; (**d**) data obtained through high-altitude (450 m) UAV remote sensing.

3.3. Prediction Results of the Rice Science and Technology Park Dataset

The objects in the rice science and technology park are relatively clear to recognize, mainly including rice, poplar trees, and the background, and the terrain is simple, with fewer ridges and fewer weeds, so the simple terrain environment brings great convenience to semantic segmentation. As shown in Figure 19, the edges of the rice field, the edges of the poplar trees, and the land exposed in the middle of the poplar trees are all accurately recognized and segmented.

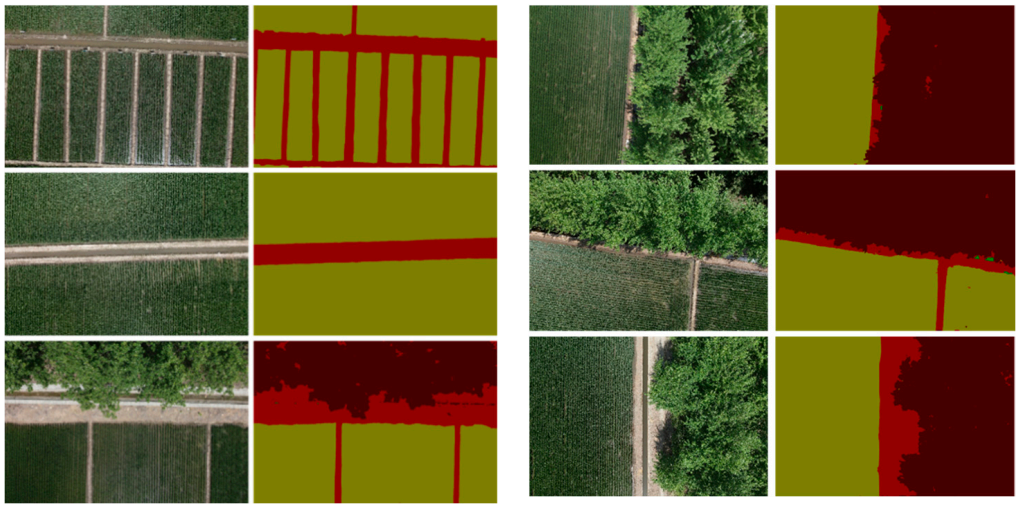


Figure 19. Projected effectiveness of rice science and technology parks.

3.4. Ablation Studies

In order to explore the effects of spatial pyramid pooling and self-attention mechanisms on the model, we conducted ablation experiments. To explore the extent to which the pooling module affects the model accuracy, we use different pyramid pooling modules instead of the SPPM of the original model, including Atrous Spatial Pyramid Pooling (ASPP) in the DeepLabV3+ model, SPPF in the YOLOV5 model, a new model proposed by Liu, S et al. [35] that incorporates the void convolution module RFB, SPPCSPC in the YOLOV7 model, SPPFCSPC obtained by combining SPPF with CSPC, the latest module SPPELAN in YOLOV9 [36], and SPPF-CSPC-A proposed in this paper.

According to the results in Table 10, SPPF, SPPFCSPC and SPPF-CSPC-A are all able to improve the performance of the model, and the module proposed in this paper is more competitive in all the evaluation metrics. This is due to the incorporation of global average pooling, a pooling operation that increases the model's ability to perceive the entire image semantics without introducing additional parameters. This helps the model to better understand the semantic information of the image, thus improving the accuracy of semantic segmentation.

Table 10. Impact of the inclusion of different pooling modules on model accuracy.

Model	MIoU	Acc	Kappa
ASPP	90.39%	98.37%	97.20%
SPPF	92.06%	98.39%	97.90%
RFB	92.68%	98.55%	98.11%
SPPCSPC	80.17%	95.55%	94.18%
SPPFCSPC	92.75%	98.69%	98.28%
SPPELAN	91.72%	98.38%	97.88%
SPPF-CSPC-A	**93.54%**	**98.81%**	**98.44%**

In order to verify the influence of complex self-attention (SA) versus lightweight SA on model accuracy, we chose the more complex SA, multi-head attention (MHA), scaled

dot-product attention (SDPA), lightweight local self-attention (LSA), block self-attention (BSA), and sparse self-attention (SSA).

The results in Table 11 demonstrate that employing complex modules does not necessarily improve model performance. On the contrary, integrating lightweight modules does not negatively affect the model performance while largely reduces the overall parameter count. The lightweight self-attention mechanism achieves this by introducing certain techniques or improvements that reduce computational overhead and parameter numbers while preserving performance. For instance, by utilizing sparse self-attention mechanisms, only a small subset of neighboring pixels is considered for each pixel, thus reducing computational complexity.

Table 11. The effect of the inclusion of different attention mechanisms on model accuracy.

Model	MIoU	Acc	Kappa
SA	88.17%	98.12%	96.78%
MHA	84.71%	95.48%	94.11%
SDPA	91.32%	98.68%	98.28%
LSA	91.81%	98.41%	97.93%
BSA	92.39%	98.41%	97.92%
SSA	**92.74%**	**98.74%**	**98.35%**

4. Discussion

4.1. Impact of Different Weights

CrossEntropyLoss (CEL) is a simple, effective and commonly used loss function in multi-category classification problems that is mathematically rigorous and easy to interpret. Weights play an important role in CEL, and they can be used to deal with the case of category imbalance or to adjust the model's focus on different categories. According to the data in Table 5, the rice sample points account for more than 60% of the data, 76 times more than the wheat sample. To balance the impact caused by too high proportion of rice on the accuracy, we designed three methods for calculating the weights, as follows:

$$W_1: w_i = \frac{N}{N_i} \tag{12}$$

$$W_2: w_i = \frac{N_{max}}{N_i} \tag{13}$$

$$W_3: w_i = \frac{N}{N_i * n} \tag{14}$$

where w_i is the weight of each category, N is the total number of pixels, N_i is the number of pixels in each category, N_{max} is the maximum number of pixels in the category, and n is the number of categories. The specific testing results are shown in Table 12.

Table 12. Weighting results.

Serial Number	Classification	Number of Pixels	Percentage	W_1	W_2	W_3
0	background	636,663,905	10.25%	9.75	6.11	1.08
1	road	62,928,541	1.01%	98.6	61.8	11
2	rice	3,889,723,226	62.64%	1.6	1	0.18
3	red bean	371,221,364	5.98%	16.7	10.5	1.86
4	coniferous tree	61,888,531	1.00%	100	62.9	11.1
5	wheat	51,259,191	0.83%	121	75.9	13.5
6	soybean	447,241,100	7.20%	13.9	8.7	1.54
7	poplar tree	177,057,377	2.85%	35.1	22	3.9
8	corn	511,421,695	8.24%	12.1	7.61	1.35

These three weight calculation methods are the most commonly used methods at present. According to the results in Table 13, none of the three weights showed overwhelmingly better results. Ohem achieves the best results and therefore is selected in this paper. Ohem utilizes Online Hard Example Mining technology to dynamically select difficult examples for training, instead of manually setting the category weights beforehand. This means that the model can adaptively adjust its focus on difficult-to- categorize samples during training without human intervention. This adaptivity can better cope with the problem of category imbalance in different datasets and scenarios.

Table 13. Comparison of accuracy with different weights.

Weights	MIoU	Acc	Kappa
W_1	91.72%	98.44%	97.33%
W_2	89.40%	97.55%	95.80%
W_3	91.25%	98.33%	97.14%
Ohem	**94.79%**	**99.02%**	**98.72%**

4.2. Impact of Image Quality on Prediction Results

During the pre-growth period of rice, the leaves are small and have difficulty in covering the water surface. The data used in this paper were taken in early July, when the rice was in the early stage of nodulation and tasseling, and the weather was sunny with sufficient light. The time of UAV aerial photography is from 10 a.m. to 1 p.m., so the acquired images will have obvious light spots (Figure 20) due to the reflection of the water surface in the rice field. The center of these light spots is incorrectly identified as the background or the road.

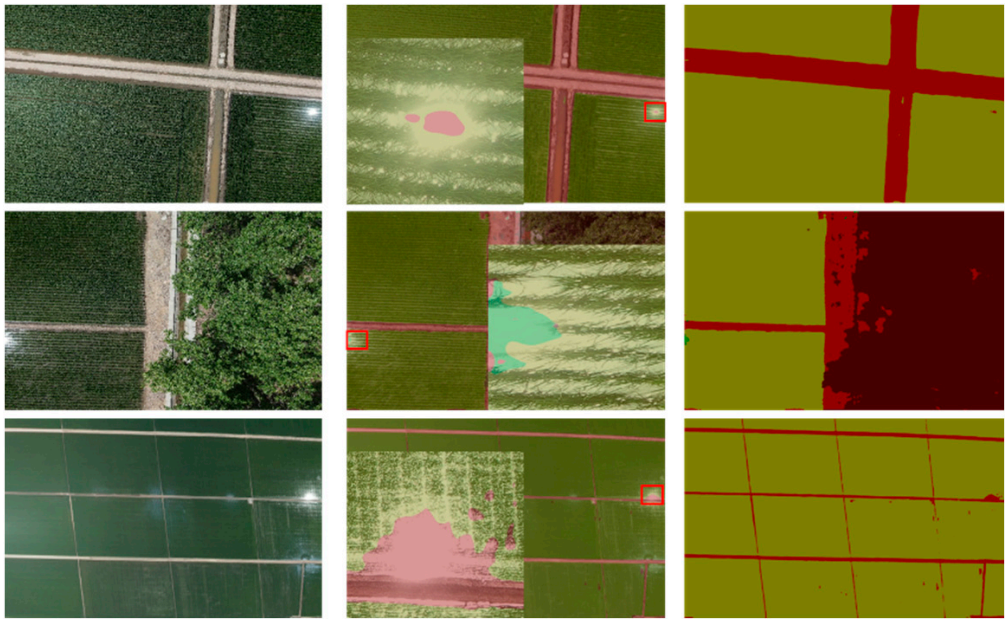


Figure 20. Light effects.

When the UAV conducts aerial photography between 6:00 a.m. and 8:00 a.m., due to the large diurnal temperature difference and low wind speed at the location of the experimental farmland, fog easily forms on the following morning (Figure 21). The dense fog significantly hinders object recognition, where foggy areas notably disrupt model predictions.

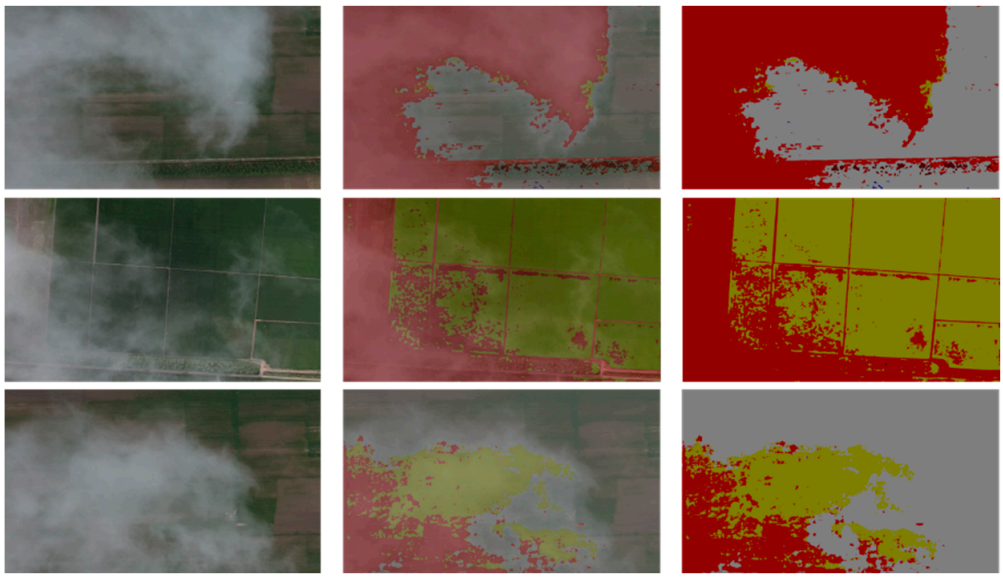


Figure 21. The effects of fog.

During the data collection process, we deliberately considered various possible weather conditions. Through testing, we found that the ubiquitous shadows and insufficient lighting do not interfere with the model's predictions. The UAV can operate normally in strong winds and obtain clear images. Only reflections and fog significantly affect the predictions, suggesting that we should avoid midday and early morning as much as possible when collecting data.

4.3. Advantages of Lightweight Models

In order to verify whether the accuracy of the models in this paper meets the requirements, we chose several high-precision complex models such as DeepLabV3+ [37], U-NET, FCN, SegNet, and the semantic segmentation model ViT based on Transformer [38]. Based on the dataset used in this paper, the accuracy comparison results are detailed in Table 14. The results indicate that the DeepLabV3+ model achieved an impressive accuracy of 97.62%. However, the model has a high FLOPs, leading to significant time consumption during training. Although the accuracy of the model proposed in this paper is slightly lower than that of DeepLabV3+, FCN, and ViT, it has exceeded the classical complex models such as U-NET, PSPNet, HRNet, and SegNet, and fully meets the requirements of agricultural applications. Moreover, our models have smaller FLOPs, significantly reducing training time. The model parameters are also relatively low, requiring less demanding hardware, making them suitable for deployment on edge computing devices and lower-performance mobile devices.

Table 14. Training results of high precision models.

Models	FOLPs	Params	MIoU	Acc	Kappa
DeepLabV3+	**191.97 G**	**45.84 M**	**97.62%**	**98.56%**	**98.29%**
FCN	93.6 G	65.94 M	97.02%	98.15%	97.80%
U-NET	125.4 G	13.41 M	90.53%	95.02%	94.06%
U^2-NET	152.4 G	44.15 M	92.66%	95.73%	94.91%
U-NET++	119.9 G	8.37 M	91.34%	95.63%	94.80%
PSPNet	343.4 G	86.96 M	81.20%	87.80%	85.65%
ViT	227.9 G	35.48 M	96.29%	97.70%	97.28%
HRNet	161.6 G	70.08 M	80.75%	90.09%	89.68%
SegNet	342.6 G	29.61 M	92.30%	96.06%	95.32%
Ours	**9 G**	**12.08 M**	**94.79%**	**99.02%**	**98.72%**

4.4. Confusion Matrix

We randomly selected 10 labeled and predicted maps for generating the confusion matrix heat map (Figure 22), and the results show that the model's prediction performance is distinctive for different crops and background categories. Rice was categorized well and most samples were accurately identified. However, some samples were misclassified as soybean and background, reflecting the similarity between rice and these categories. In the prediction of soybean, the performance was equally good, but there were still some samples misclassified as rice, wheat and red bean, revealing confusion between soybean and other crops. Classification stability was high for wheat, although individual samples were mislabeled as soybean or background. Predictions for red beans were also satisfactory, although some samples were mislabeled as soybean and background, showing overlap of features with these categories. Corn was particularly well classified, with fewer misclassifications, mainly in the background category. Poplar was also well recognized, but confusion with rice and soybeans still needs attention. Predictive performance in the road category was more favorable; however, many samples were classified as background due to similarities with background features. The background category had a large sample size, but many of these samples were misclassified as other crops, especially soybeans and red beans, further highlighting the problem of overlap between background and crop characteristics. Conifer predictions similarly showed similarities to background.

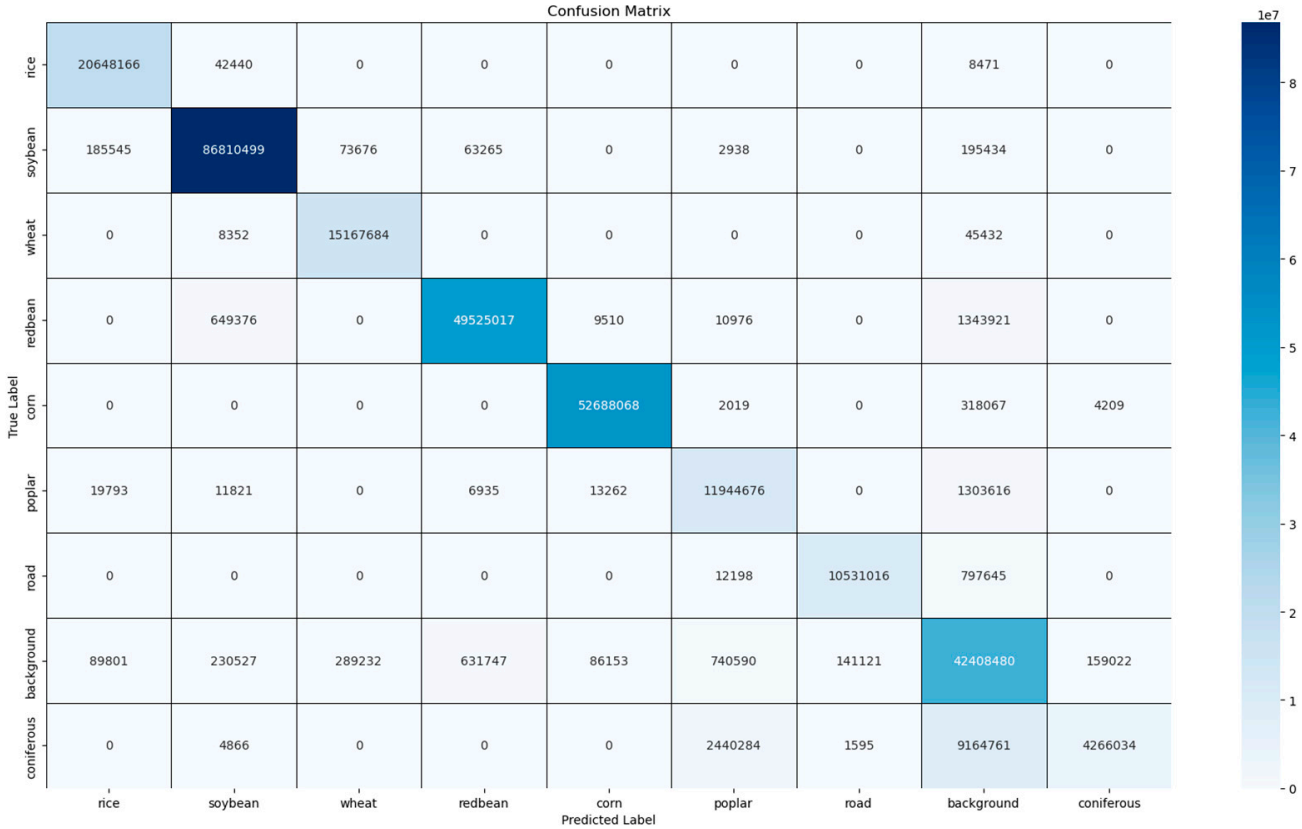


Figure 22. Confusion matrix heat map.

Overall, while the model performed well in classifying rice and soybeans, confusion was still evident among the other categories, especially between soybeans, red beans, and backgrounds. This suggests that the classification accuracy of the method may be degraded in situations with high coverage between crops or less than ideal light conditions. In the future, we plan to explore more robust multimodal data fusion techniques combining multispectral or hyperspectral data to further improve the classification accuracy. Meanwhile, we also plan to investigate more optimization algorithms for different crop types to cope

with crop classification needs at different growth stages, and integrate the model into real agricultural monitoring systems for real-time application testing.

5. Conclusions and Outlook

To achieve real-time crop classification based on UAV imagery, this paper proposes a lightweight semantic segmentation model tailored for crop recognition. The model is based on Baidu's PP-Liteseg model and incorporates a pooling module, SPPF-CSPC-A, and a sparse self-attention mechanism designed in this study. SPPF-CSPC-A can better deal with the features of high-resolution images acquired by UAVs, and the combination of global average pooling and sparse self-attention mechanisms can cope with different scales of images acquired by UAVs. Compared with some of the top models, our model is more competitive in terms of classification accuracy and prediction results. The accuracy of our model reaches 94.79%, which is a 2.79% improvement over the best results among the benchmark models. In future work, we will focus on achieving more accurate classification, which can accurately grasp the climatic period of each crop at this time while identifying different crops. In addition, we will build a comprehensive data management platform to integrate multi-source data such as multi-spectral, infrared, and visible images to improve data processing efficiency [39,40]. The platform will integrate the methods in this paper with machine learning models to realize the automated processing of crop classification, and combine with GIS functions to calculate the growth and planting area of each crop. Meanwhile, the platform supports real-time data transmission and processing to meet the real-time decision-making needs of precision agriculture. In addition, it is crucial to ensure data security and privacy protection, and the platform should adopt encryption and access control to meet relevant regulatory requirements. Finally, the platform also needs to be open and scalable to adapt to future technological advances and changes in demand. Through the platform, the precision and efficiency of agricultural management will be significantly improved, accelerating the process of agricultural digitization and intelligence. At the same time, we hope to obtain more data from geographic regions and different crop types through the data platform to verify the applicability of the model, and at the same time explore how to adjust and optimize the model according to different agricultural practices or environmental conditions. It provides a wider range of application ideas for subsequent research.

Author Contributions: Project administration, L.G. and H.Y.; writing—original draft preparation, Z.Z.; funding acquisition, H.Y.; review and editing, L.G. and Q.L.; resources, H.Y.; data curation, J.Y. and Z.Z.; conceptualization and methodology, L.G.; supervision, W.Y.; All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the National Key R&D Program of China (No. 2021ZD0110901) and the Science and Technology Innovation Program of AII-CAAS (No. CAAS-ASTIP-2024-AII).

Data Availability Statement: The data generated in this study are not publicly available due to its use in an ongoing study by the authors but can be made available from the corresponding author upon reasonable request.

Acknowledgments: We would like to thank all the authors for their administrative and technical support during the course of this study. We also thank the Department of Computing, Harbin Institute of Technology, and the Institute of Agricultural Information, Chinese Academy of Agricultural Sciences, for providing equipment for the experiments, and this support played an important role in the successful completion of this study.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Liu, Y.; Ma, X.; Shu, L.; Hancke, G.P.; Abu-Mahfouz, A.M. From Industry 4.0 to Agriculture 4.0: Current Status, Enabling Technologies, and Research Challenges. *IEEE Trans. Ind. Inform.* **2020**, *17*, 4322–4334. [CrossRef]
2. Lu, H.; Liu, Q.; Liu, X.; Zhang, Y. A Survey of Semantic Construction and Application of Satellite Remote Sensing Images and Data. *J. Organ. End User Comput.* **2021**, *33*, 1–20. [CrossRef]

3. Shahi, T.B.; Xu, C.-Y.; Neupane, A.; Guo, W. Machine learning methods for precision agriculture with UAV imagery: A review. *Electron. Res. Arch.* **2022**, *30*, 4277–4317. [CrossRef]
4. Peña, J.M.; Gutiérrez, P.A.; Hervás-Martínez, C.; Six, J.; Plant, R.E.; López-Granados, F. Object-Based Image Classification of Summer Crops with Machine Learning Methods. *Remote Sens.* **2014**, *6*, 5019–5041. [CrossRef]
5. Ok, A.O.; Akar, O.; Gungor, O. Evaluation of random forest method for agricultural crop classification. *Eur. J. Remote Sens.* **2012**, *45*, 421–432. [CrossRef]
6. Huang, H.; Lan, Y.; Yang, A.; Zhang, Y.; Wen, S.; Deng, J. Deep learning versus Object-based Image Analysis (OBIA) in weed mapping of UAV imagery. *Int. J. Remote Sens.* **2020**, *41*, 3446–3479. [CrossRef]
7. Kamilaris, A.; Prenafeta-Boldú, F.X. Deep learning in agriculture: A survey. *Comput. Electron. Agric.* **2018**, *147*, 70–90. [CrossRef]
8. Bouguettaya, A.; Zarzour, H.; Kechida, A.; Taberkit, A.M. Deep learning techniques to classify agricultural crops through UAV imagery: A review. *Neural Comput. Appl.* **2022**, *34*, 9511–9536. [CrossRef]
9. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 640–651.
10. Ronneberger, O.F.P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv* **2015**, arXiv:1505.04597.
11. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *39*, 2481–2495. [CrossRef]
12. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 834–848. [CrossRef] [PubMed]
13. Kattenborn, T.; Eichel, J.; Fassnacht, F.E. Convolutional Neural Networks enable efficient, accurate and fine-grained segmentation of plant species and communities from high-resolution UAV imagery. *Sci. Rep.* **2019**, *9*, 17656. [CrossRef]
14. Yang, M.-D.; Tseng, H.-H.; Hsu, Y.-C.; Tsai, H.P. Semantic Segmentation Using Deep Learning with Vegetation Indices for Rice Lodging Identification in Multi-date UAV Visible Images. *Remote Sens.* **2020**, *12*, 633. [CrossRef]
15. Morales, G.; Kemper, G.; Sevillano, G.; Arteaga, D.; Ortega, I.; Telles, J. Automatic Segmentation of Mauritia flexuosa in Unmanned Aerial Vehicle (UAV) Imagery Using Deep Learning. *Forests* **2018**, *9*, 736. [CrossRef]
16. Zhong, Y.; Hu, X.; Luo, C.; Wang, X.; Zhao, J.; Zhang, L. WHU-Hi: UAV-borne hyperspectral with high spatial resolution (H2) benchmark datasets and classifier for precise crop identification based on deep convolutional neural network with CRF. *Remote Sens. Environ.* **2020**, *250*, 112012. [CrossRef]
17. Haq, M.A. CNN Based Automated Weed Detection System Using UAV Imagery. *Comput. Syst. Sci. Eng.* **2022**, *42*, 837–849.
18. Paszke, A.; Chaurasia, A.; Kim, S.; Culurciello, E. ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation. *arXiv* **2016**, arXiv:1606.02147.
19. Peng, J.; Liu, Y.; Tang, S.; Hao, Y.; Chu, L.; Chen, G.; Wu, Z.; Chen, Z.; Yu, Z.; Du, Y.; et al. PP-LiteSeg: A Superior Real-Time Semantic Segmentation Model. *arXiv* **2022**, arXiv:2204.02681.
20. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All you Need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.
21. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *37*, 1904–1916. [CrossRef]
22. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475.
23. Child, R.; Gray, S.; Radford, A.; Sutskever, I. Generating Long Sequences with Sparse Transformers. *arXiv* **2019**, arXiv:1904.10509.
24. Fan, M.; Lai, S.; Huang, J.; Wei, X.; Chai, Z.; Luo, J.; Wei, X. Rethinking BiSeNet For Real-time Semantic Segmentation. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 9711–9720.
25. Tang, W.; Hu, J.; Zhang, H.; Wu, P.; He, H. Kappa coefficient: A popular measure of rater agreement. *Shanghai Arch. Psychiatry* **2015**, *27*, 62–67. [PubMed]
26. Shrivastava, A.; Gupta, A.; Girshick, R. Training Region-Based Object Detectors with Online Hard Example Mining. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 761–769.
27. Shrivastava, A.; Gupta, A.; Girshick, R. BiSeNet: Bilateral Segmentation Network for Real-time Semantic Segmentation. *arXiv* **2018**, arXiv:1808.00897.
28. Yu, C.; Gao, C.; Wang, J.; Yu, G.; Shen, C.; Sang, N. BiSeNet V2: Bilateral Network with Guided Aggregation for Real-Time Semantic Segmentation. *Int. J. Comput. Vis.* **2020**, *129*, 3051–3068. [CrossRef]
29. Mehta, S.; Rastegari, M.; Shapiro, L.; Hajishirzi, H. ESPNetv2: A Light-Weight, Power Efficient, and General Purpose Convolutional Neural Network. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 9182–9192.
30. Poudel, R.P.; Liwicki, S.; Cipolla, R. Fast-SCNN: Fast Semantic Segmentation Network. *arXiv* **2019**, arXiv:1902.04502.

31. Guo, M.H.; Lu, C.Z.; Hou, Q.; Liu, Z.; Cheng, M.M.; Hu, S.M. SegNeXt: Rethinking Convolutional Attention Design for Semantic Segmentation. *arXiv* **2022**, arXiv:2209.08575.
32. Zhang, W.; Huang, Z.; Luo, G.; Chen, T.; Wang, X.; Liu, W.; Yu, G.; Shen, C. TopFormer: Token Pyramid Transformer for Mobile Semantic Segmentation. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 12073–12083.
33. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. *Neural Inf. Process. Syst.* **2021**, *34*, 12077–12090.
34. Wang, J.; Gou, C.; Wu, Q.; Feng, H.; Han, J.; Ding, E.; Wang, J. RTFormer: Efficient Design for Real-Time Semantic Segmentation with Transformer. *arXiv* **2022**, arXiv:2210.07124.
35. Liu, S.; Huang, D.; Wang, Y. Receptive Field Block Net for Accurate and Fast Object Detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
36. Wang, C.Y.; Yeh, I.H.; Liao, H.Y.M. YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information. *arXiv* **2024**, arXiv:2402.13616.
37. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. 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.
38. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16 $\times$ 16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.
39. Zhou, L.; Tu, W.; Li, Q.; Guan, D. A Heterogeneous Streaming Vehicle Data Access Model for Diverse IoT Sensor Monitoring Network Management. *IEEE Internet Things J.* **2024**, *11*, 26929–26943. [CrossRef]
40. Zhou, L.; Tu, W.; Wang, C.; Li, Q. A Heterogeneous Access Metamodel for Efficient IoT Remote Sensing Observation Management: Taking Precision Agriculture as an Example. *IEEE Internet Things J.* **2022**, *9*, 8616–8632. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.