GF-UNet: A Cropland Extraction Method Based on Attention Gates and Adaptive Feature Fusion

Chuanhu Li*, Yunyan Wang
School of Electrical and Electronic Engineering, Hubei University of Technology, China
E-mail: nccvsc@126.com
*Corresponding author

Keywords: attention gate; feature fusion; high-resolution imaging; extraction of cropland; semantic segmentation.

Received: February 1, 2024

Cropland plays a critical role in maintaining national food security, but its extraction is often hindered by factors such as type of cropland, crop category, and surrounding vegetation, resulting in low extraction accuracy. This paper proposes a cropland extraction network, called GF-UNet, to address the challenges of accurately extracting cropland from very high-resolution (VHR) remote sensing images. GF-UNet builds on the Attention U-Net network and introduces Attention Gates (AGs) to improve the ability to discriminate between similar features of cropland and non-cropland in complex situations. This helps to improve the accuracy of cropland extraction. In addition, an Adaptive Feature Fusion Module (AFFM) is incorporated to integrate multi-scale cropland features, further enhancing the network's ability to identify cropland. The Spatial Feature Extraction Module (SFEM) is also introduced into the skip connection to improve the extraction of detailed features in the results. The research data used in the study consists of GF-2 satellite images of Xuan 'en County, Hubei Province, from June to September 2019. Comparative experiments were conducted with SOTA models, the results demonstrate that GF-UNet outperforms the other models in terms of accuracy, F1-score, and IoU. The accuracy, F1-score, and IoU of GF-UNet were reported as 91.25%, 92.41%, and 84.56%, respectively. The study also explores the impact of SFEM and AFFM on the experimental results. Compared to existing SOTA methods, GF-UNet proves to be more suitable for cropland extraction in complex scenes, providing a practical approach to addressing the challenges of cropland extraction in such scenarios.

Povzetek: Članek predstavlja omrežje GF-UNet za natančno izločanje kmetijskih površin iz slik visoke ločljivosti s pomočjo daljinskega zaznavanja. GF-UNet uporablja modul za adaptivno združevanje značilnosti in prostorski modul za izboljšanje delovanja.

1 Introduction

Cropland plays a crucial role in modern agricultural development and is vital to the survival of human society [1]. Timely and accurate access to agricultural information is of great importance for ensuring national food security and promoting sustainable development of the national economy [2]. Remote sensing technology, with its wide coverage and timely imaging capabilities, enables rapid updates of agricultural information [3,4].

As the spatial resolution of remote sensing imagery continues to improve, ground objects can be represented with greater accuracy. In particular, very high resolution (VHR) remote sensing imagery, with a resolution of less than 5.0 meters [5], can effectively capture the shape and type of land objects, providing accurate data for precise crop monitoring [2,6].

Traditional cropland extraction methods, including K-Means [7], Support Vector Machine (SVM) [8], Decision Tree (DT) [1], and Random Forest (RF) [9], mainly rely on the intrinsic characteristics of the image spectrum, texture, and geometric information to derive cropland information [10]. However, the results of these extraction methods are susceptible to the pepper and salt phenomenon, resulting in reduced accuracy [11]. In addition, these techniques underutilize high-level image features such as image morphology and context information, resulting in cropland extraction results that may not meet practical requirements [1,10].

Convolutional Neural Networks (CNNs) have emerged as a highly effective deep learning architecture for semantic segmentation tasks, including the extraction of ground object information from remote sensing images [12]. Many researchers have used CNNs to extract various features such as buildings, roads, and cropland due to their ability to independently learn abstract features and capture contextual associations in images without relying on hand-crafted features [13,14,15]. For example, Liu et al. [16] used U-Net to identify cropland, effectively mitigating the salt and pepper noise phenomenon associated with traditional methods. Similarly, Du et al. [17] applied DeepLabV3+ to segment irregular small cropland plots.

However, encoder-decoder networks such as U-Net [18], PSPNet [19], and DeepLab_V3+ [17] have a potential disadvantage in that they may introduce irrelevant information that has been filtered out in deep networks. This problem arises from the optimization of the decoder's upsampling results using low-level features from the encoder via skip connections, which may affect the model's segmentation performance [20]. To solve this problem, Oktay et al. [21] proposed the Attention U-Net, which incorporates attention gates (AGs) into the U-Net architecture. The Attention U-Net assigns different weights to the connection features, allowing the suppression of irrelevant areas and the highlighting of significant features that are particularly useful for the specific segmentation task. This method helps to improve segmentation accuracy by focusing on relevant information. However, the Attention U-Net does not fully take into account the local spatial details in the shallow features and the relationship between the overall and local contextual features. It also overlooks the importance of spatial detail and location information [22].

These limitations highlight the need for further improvements in the modeling of spatial detail and the integration of local and global contextual features. By addressing these challenges, it is possible to improve the accuracy and performance of cropland extraction models, enabling more accurate identification and delineation of cropland areas in remotely sensed imagery.

Furthermore, the inclusion of multi-scale features is crucial for improving the accuracy of semantic segmentation. Researchers have explored different approaches to incorporate multi-scale information into the models.

Yang et al. [23] used parallel and cascaded architectures of dilated convolutions to design the DenseASPP module, allowing the model to learn more global features. Liu et al. [24] constructed a new residual ASPP to obtain essential multiscale semantic information while avoiding the problem of gradient disappearance.

However, a common challenge with these methods is the use of all channel information from the input features for the feature scale transformation. While this method enables multi-scale feature fusion, it can increase the computational burden of the model and introduce redundant information. To address this issue, researchers have explored techniques to optimize multi-scale feature fusion. These include methods such as channel attention and feature recalibration mechanisms that selectively emphasize relevant information and suppress redundant or less informative features. In this way, models can achieve a more efficient and effective fusion of multi-scale features, leading to improved segmentation accuracy.

Based on the research mentioned above, we have developed a novel deep-learning approach for the precise extraction of cropland from very high-resolution (VHR) images. Our approach employs a CNN model that integrates attention gates and multi-scale feature fusion. The following are the main innovative aspects and contributions of our approach:

(1) A proposed method for extracting cultivated land from VHR images involves using the GF-UNet model. The GF-UNet model includes an adaptive feature fusion module (AFFM) and a spatial feature extraction module (SFEM) to improve feature recognition and detail extraction capabilities.

(2) The purpose of this study was to collect and process GF-2 satellite data. Data enhancement techniques were used to expand the number of samples to ensure an adequate dataset for the experiment.

(3) To evaluate the effectiveness of our proposed model, we conducted a comparative analysis with several popular semantic segmentation models. The aim was to quantitatively and qualitatively analyze the experimental results, in order to verify the superiority of our method.

The remainder of this paper is organized as follows: Section I gives the introduction, Section II describes the related work, Section III describes the data processing process and the proposed methods, Section IV analyzes the experimental results, Section V discusses the reasons for the performance differences of the models. Finally, Section VI summarizes the thesis.

2 Related work

Land cover information plays a crucial role in the advancement of agricultural remote sensing. Many scientists have made significant contributions to the research of land cover information extraction and land cover mapping.

Hong et al. [25] introduced a farmland boundary extraction technique that systematically incorporates several computational and mathematical methods, including the Suzuki85 algorithm, Canny edge detection, and the Hough transform, to extract farmland distribution information in six South Korean regions. This algorithm extracts boundaries with 80.7% accuracy, 79.7% completeness, and 67.0% quality, allowing for the automatic creation of farm maps.

Graesser et al. [26] developed a method for cropland area extraction that combines multispectral picture edge extraction, multi-scale contrast-limited adaptive histogram equalization, and adaptive threshold segmentation. The study focused on extracting farmland distribution information from portions of South America, and the extracted results had an F1 score of 91%. The approach is very useful for extracting cropland distribution data over huge areas. It allows for accurate monitoring of agricultural developments.

Zhang et al. [27] proposed a general method for high-resolution cropland mapping using deep convolutional neural networks. Their method utilized the Pyramid Scene Resolution Network (PSPNet), which was slightly modified to combine deep remote features with local shadow features. This combination enabled more detailed predictions and improved accuracy in cropland mapping. The MPSPNet, a modified version of PSPNet, was evaluated using high-resolution satellite imagery in four different research areas in China. The method achieved an overall accuracy of 89.99% in the validation process.

In the study [28], the authors introduced a multi-scale fusion network for cropland extraction that incorporates an attention mechanism. This method utilizes an image gradient attention guide module to improve the accuracy of the extracted cropland information. To ensure comprehensive and complete cropland information extraction, the authors also incorporated a multi-scale spatial feature consensus fusion model into the network. The experimental results demonstrate that this method efficiently extracts information on cropland boundaries and enables the extraction of semantic information related to cropland. The attention mechanism and multi-scale fusion contribute to improved accuracy and a more comprehensive understanding of cropland areas.

In the study [29], Xu et al. introduced a multi-task cascade network model called SGENet for the extraction of farmland plot information. This model was designed to automatically learn multi-scale and multi-level features, enabling it to handle complex planting scenarios and different scales of farmland plots.

In the study [30], Huan et al. proposed a multi-attention encoder-decoder network (MAENet) for the segmentation of agricultural scenes. The authors aimed to improve the segmentation performance of the network by incorporating several modules, including the dual-pooling efficient channel attention (DPECA) module, the dual-feature attention (DFA) module, and the global-guidance information upsampling (GIU) module. The authors evaluated the performance of MAENet on three self-generated farmland image datasets representing UAV data. The results showed that MAENet achieved an impressive MIoU of 93.74% and Kappa score of 96.74%, outperforming other existing methods. The research discussed in this section is summarized in Table 1.

The GF_UNet model is a cropland extraction network that improves the performance of the standard U-Net network architecture by incorporating AGs to distinguish between different categories with similar features [31]. The GF_UNet model is a cropland extraction network that improves the performance of the standard U-Net network architecture by incorporating AGs to distinguish between different categories with similar features. This study proposes modifications to the Attention U-Net architecture to create the GF_UNet model. The purpose of these modifications is to enhance the accuracy and efficiency of extracting croplands. The GF_UNet model underwent evaluation using a self-generated dataset that included a variety of complex cultivated land scenarios. The results of the evaluation demonstrated the model's strong performance, achieving an accuracy rate of 91.25% and an F1-score value of 92.41%. These metrics indicate the model's ability to accurately extract cropland regions from input imagery.

Table 1: Summary of related works

Research works	Summary of methods	Limitation
Development of a Parcel-Level Land Boundary Extraction Algorithm for Aerial Imagery of Regularly Arranged Agricultural Areas [25]	An algorithm for extracting farmland boundaries is presented that uses a combination of computational and mathematical techniques, including the Suzuki85 algorithm, Canny edge detection, and Hough transform.	This algorithm may not be suitable for the case where the cropland deviates significantly from the shape rule.
Detection of cropland field parcels from Landsat imagery [26]	A cropland area extraction method that involves the combination of multispectral image edge extraction, multiscale contrast-limited adaptive histogram equalization, and adaptive threshold segmentation.	It is important to note that this method relies on prior knowledge of the scene, which may limit its applicability for extracting cropland information over large areas.
A generalized approach based on convolutional neural networks for large area cropland mapping at very high resolution [27]	A method for high-resolution cropland mapping using deep convolutional neural networks	Applying the method to scenes with complex land cover patterns or a mix of different surface types may pose challenges.
Study of Multiscale Fused Extraction of Cropland Plots in Remote Sensing Images Based on Attention Mechanism [28]	A multiscale fusion cropland extraction network that incorporates an attention mechanism.	This method may result in cropland boundaries with some degree of fuzziness, and there may be instances where parcels are connected.
Extraction of cropland field parcels with high resolution remote sensing using multi-task learning [29]	A multi-task cascade network model called SGENet	It is important to note that this approach may face challenges when dealing with regions that have similar characteristics of different species.
MAENet: Multiple Attention Encoder–Decoder Network for Farmland Segmentation of Remote Sensing Images [30]	A multi-attention encoder-decoder network (MAENet) for agricultural scene segmentation.	The dataset scenario is straightforward, and the MAENet model's generalization ability is inadequate.

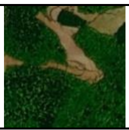
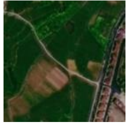
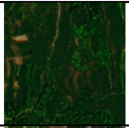
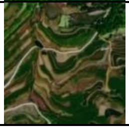
3 Materials and proposed method

3.1 Data preprocessing

Experimental data from high-resolution Earth observation system data and application center of Hubei Province (http://datasearch.hbeos.org.cn), including 16 of micro-cloud cover (<5%), the high quality of GF-2 scene satellite images, it covers Xuen County, Hubei Province in central China (29°01'-33°06'N, 108°21'-116°07'E). First, we used ENVI 5.3 to pre-process the collected GF-2 remote sensing data with orthometric correction, atmospheric correction, radiometric correction, and multispectral and panchromatic band fusion [32]. The atmospheric correction was performed using the FLAASH atmospheric correction [33]. The full-color image was sharpened using a sharpening filter before image fusion. The acquired multispectral and panchromatic images were then fused at a 4:1 ratio, and the fused GF-2 image had a spatial resolution of 1 meter.

Cropland samples of GF-2 images were mapped based on the pre-processed GF-2 images and the actual collected cropland images. The GF-2 images and mapped cropland samples were cropped to 256×256 size and divided into training data, validation data, and test data. Data enhancement operations such as rotation and noise addition were performed on the training and validation data to increase the number of samples, which could avoid overfitting the training model due to insufficient training data during the training process. The dataset for this study consisted of 9863 training sets, 1932 validation sets, and 1860 test sets, with an approximate ratio of 8:1:1. It included four types of croplands, and their characteristics are summarized in Table 2.

Table 2: Cropland type characteristics in dataset

Type	Image	Feature description
Type 1		The grain within the cropland appears uniform, and the boundary of the cropland is clearly defined.
Type 2		The boundary between large croplands is easily distinguishable, but the features of small croplands vary.
Type 3		The characteristics of cropland closely resemble those of the surrounding background.
Type 4		The cropland has an irregular shape, and its boundaries are interconnected.

3.2 Network structure

To accurately extract the cropland distribution information, this paper proposes the GF-UNet network model, and its structure is shown in Figure 4.

GF-UNet adds the AFFM and SFEM based on Attention U-Net, which mainly consists of four parts: the encoder, the decoder, the SFEM, and the AGs. Similar to Attention U-Net, the first three layers of the GF-UNet encoder consist of two convolutional layers (Conv) with a batch normalization layer (BN) and a linear rectification function (ReLU) in series. The fourth layer consists of AFFM. AFFM extracts the global semantic information of the farmland by fusing the multi-scale farmland features and combines the squeeze and excitation (SE) channel attention mechanism [34] to obtain the channel weight distribution values of the fused farmland features, which enhances the ability of the network to recognize the farmland attributes. To effectively extract the spatial detail information of the low-level features and preserve the location information of the spatial details, GF-UNet adds SFEM to the skip connection. Then, the result of SFEM is used as the input of AGs to increase the responsiveness of the network to the cropland features. Finally, the final cropland distribution information is output through the convolution module in the decoder.

3.3 Adaptive feature fusion module

In the task of semantic segmentation, it is crucial to integrate multiscale information due to variations in segmented objects. Relying solely on a single scale of features often leads to inadequate extraction outcomes [35]. This paper proposes an Adaptive Feature Fusion Module (AFFM), which comprises a parallel multi-branch network consisting of a multi-scale feature fusion module and an attention enhancement module.

AFFM effectively captures both the global contextual features of the field and the primary semantic information of the cropland. It accomplishes this through its multiscale feature fusion module that comprehensively learns field features, along with an attention enhancement module that learns channel weight distribution for cropland features while reducing redundant information during network training. Figure 2 illustrates the structure of AFFM.

AFFM divides the input feature X into four sub-features: X_1, X_2, X_3 and X_4, in channel order. These sub-features capture different aspects of the input data. To capture feature information at different scales, X_1, X_2, and X_3 are sequentially pooled. Specifically, X_1 is subsampled 8 times, X_2 is subsampled 4 times, and X_3 is subsampled 2 times. The process of pooling reduces the spatial resolution of the features while preserving their essential information. However, X_4 does not undergo any pooling operation. Therefore, X_4 retains its original spatial resolution and is not downsampled like X_1, X_2, and X_3. By keeping the spatial details intact in X_4, the network can capture fine-grained information and maintain the location information of the features. To capture global contextual information and achieve a broader receptive field, we utilized depth-separable convolution with a 3×3 kernel size to extract four sub-features. After extracting the features, X_1, X_2, and X_3 are upsampled to match the spatial resolution of the input feature X, resulting in four different scales of feature maps: Y_1, Y_2, Y_3, and Y_4. To combine information from the four scales, concatenate the feature maps Y₁, Y₂, Y₃, and Y₄ sequentially. This creates a new feature map that encapsulates information from different scales. Apply a 1×1 convolution operation to the concatenated feature map to allow for the interaction of channel information across the different scales.

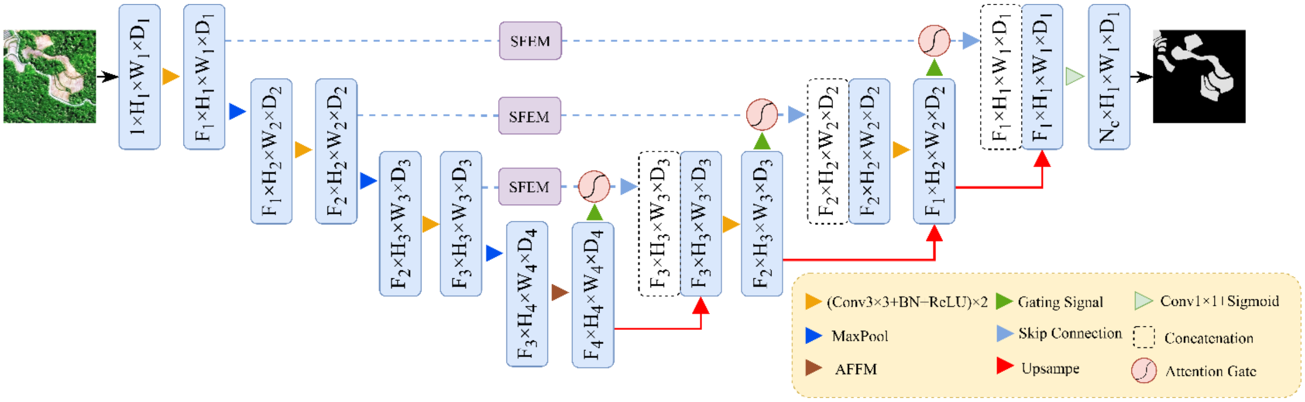

$1\times H_1\times W_1\times D_1$
$F_1\times H_1\times W_1\times D_1$
$F_1\times H_2\times W_2\times D_2$
$F_2\times H_2\times W_2\times D_2$
$F_2\times H_3\times W_3\times D_3$
$F_3\times H_3\times W_3\times D_3$
$F_3\times H_4\times W_4\times D_4$
$F_4\times H_4\times W_4\times D_4$
SFEM
SFEM
SFEM
$F_3\times H_3\times W_3\times D_3$
$F_3\times H_3\times W_3\times D_3$
$F_2\times H_3\times W_3\times D_3$
$F_2\times H_2\times W_2\times D_2$
$F_2\times H_2\times W_2\times D_2$
$F_1\times H_2\times W_2\times D_2$
$F_1\times H_1\times W_1\times D_1$
$F_1\times H_1\times W_1\times D_1$
$N_c\times H_1\times W_1\times D_1$
(Conv3×3+BN–ReLU)×2
Gating Signal
Conv1×1+Sigmoid
MaxPool
Skip Connection
Concatenation
AFFM
Upsampe
Attention Gate

Figure 1: Architecture of GF-UNet

To adaptively weigh the spatial information, the paper uses the sigmoid activation function to obtain the spatial attention weight. This weight is then multiplied element-wise with the input feature X, resulting in a spatially adaptive feature map denoted as Yₛ. The multiplication process ensures that the features deemed important by the attention weight receive more emphasis, while less important features are downweighted.

$$Y_i = UP(DWConv3\times3(Pool(X_i))) \quad (1)$$

$$Y_s = X \otimes \sigma[Conv1\times1(concat(Y_i))] \quad (2)$$

Where *UP* is upsampling, *DWConv* denotes depth-separable convolution, *Pool* is pooling operation, *concat* is stacking according to channel, σ denotes *Sigmoid* , ⊗ denotes pixel multiplication.

The channel weights for the spatial adaptive feature map Ys are determined using global average pooling (GAP), a fully connected layer (FC), and a sigmoid activation function. These operations generate the channel weight values for Yₛ. By multiplying these channel weights with Yₛ, we can redistribute the weights of Yₛ and obtain the adaptive feature Y. This adaptive feature captures the refined and tuned contributions of each channel, ultimately enhancing the discrimination of features.

$$Y = \sigma_2[FC(\sigma_1(FC(GAP(Y_s))))] \otimes Y_s \quad (3)$$

Where *FC* represents the fully connected layer, *GAP* represents the global average pooling, σ_1 represents the *ReLU* activation function, and σ_2 represents the *Sigmoid* activation function.

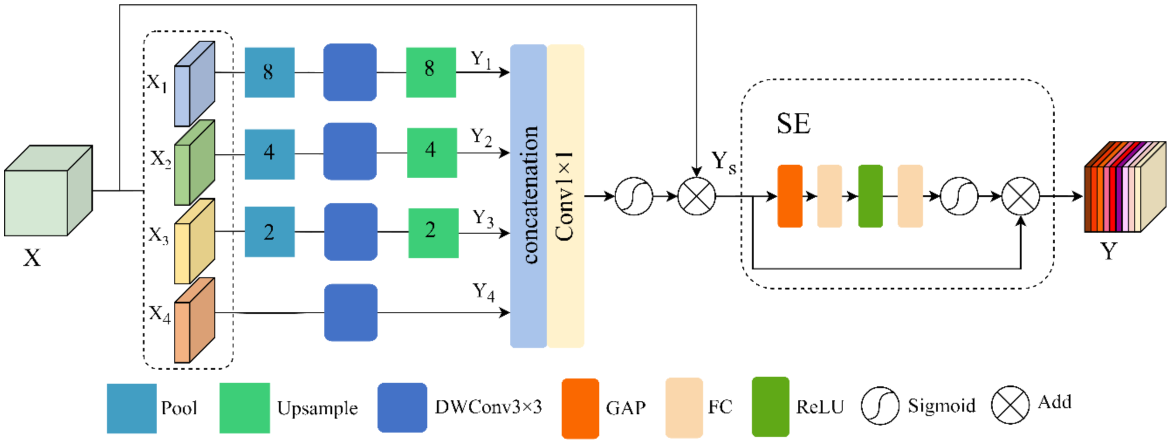


Figure 2: Architecture of AFFM

3.4 Spatial feature extraction module

The attentional mechanism, inspired by human vision, is effective in focusing on important detailed features during network training, thus improving network performance [36]. The CBAM attention mechanism, which includes both channel and spatial attention, enhances the network's learning capability [37].

To obtain spatial attention, CBAM calculates spatial attention weights by performing average pooling and maximum pooling operations on the channel dimensions. These operations extract the maximum and average values within each channel at the same spatial location. However, it is important to note that the pooling operation can result in the loss of channel information, which can be detrimental to the effective transfer of information during network training.

Retaining channel information is crucial for preserving the discriminative power of features. Without channel information, the network may struggle to capture the fine-grained details necessary for accurate segmentation. Therefore, it is important to address the potential loss of channel information when using spatial attention mechanisms like CBAM.

To tackle the problem of potential loss of channel information due to pooling operations in the CBAM attention mechanism, SFEM removes the pooling layer of channel dimension in CBAM. Instead, it uses two layers of 7×7 depth-separable convolutions to increase the receptive field and capture more global spatial feature information. The use of depth-separable convolutions reduces computational complexity while maintaining effectiveness in capturing spatial features.

To facilitate the comprehensive understanding of input features, a 1×1 convolution is used to enable the interaction of channel information, performing upscaling and downscaling of feature channels.

SFEM improves the contextualization of spatial features by incorporating these modifications. SFEM contributes to the detailed restoration of plowing results by preserving the overall structure and details of the image. Figure 3 depicts the structure of SFEM, showcasing the arrangement of the components involved in capturing global spatial feature information and promoting the interaction of channel information.

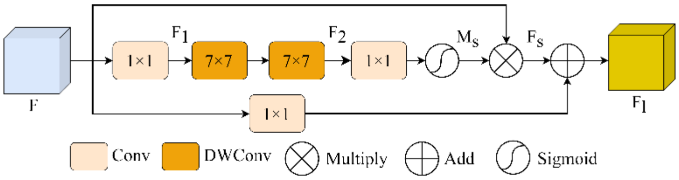


Figure 3: Architecture of SFEM

3.5 Attention gates

Attention Gates (AGs) identify salient feature regions using the attention coefficient $\alpha \in [0,1]$ and suppress the responses of irrelevant features, maximizing the retention and activation of neurons associated only with salient features [38]. The structure of AGs can be seen in Figure 4.

In AGs, relevant feature representations from the input are captured by extracting the input feature x_l and the gate signal x_g as two types of feature information using 1×1 convolutions W_l and W_g, respectively. The two types of feature information are then fused together to obtain the feature image x_t, which combines the salient information from both x_l and x_g. To obtain feature information x_q, activate the feature of x_t using the ReLU function and reduce the feature dimensionality by 1 ×1 convolution ψ. The feature image x_t is passed through the

ReLU activation function to enhance its activation. Then, it undergoes dimensionality reduction using a 1×1 convolution ψ. This step reduces the number of feature channels while preserving important information x_q.

Next, the sigmoid activation function is applied to x_q to obtain the attention coefficients α. These coefficients are then resampled and multiplied with the input features x_l to obtain $x_l^{'}$. This process enhances the representation of salient features while suppressing irrelevant feature regions.

$$x_t = W_g(x_g) \oplus W_l(x_l) \tag{4}$$

$$x_q = \psi(\sigma_1(x_t)) \tag{5}$$

$$\alpha = Resampler(\sigma_2(x_q)) \tag{6}$$

$$x_l^{'} = \alpha \otimes x_l \tag{7}$$

Where W_g, W_l, and ψ are 1×1 convolution, σ_1 is $ReLU$, σ_2 is $Sigmoid$, $Resampler$ is upsampling, $\oplus$ and $\otimes$ are pixel addition and pixel multiplication, respectively.

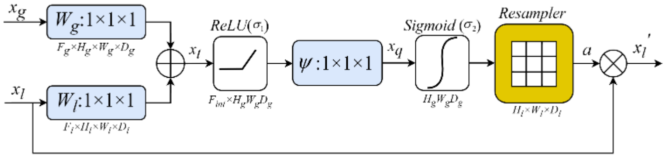

x_g
W_g:1×1×1
$F_g \times H_g \times W_g \times D_g$
x_l
W_l:1×1×1
$F_l \times H_l \times W_l \times D_l$
x_t
$ReLU(\sigma_1)$
$F_{int} \times H_g W_g D_g$
ψ:1×1×1
x_q
$Sigmoid\ (\sigma_2)$
$H_g W_g D_g$
$Resampler$
$H_l \times W_l \times D_l$
a
x_l'

Figure 4: Architecture of ags

3.6 Loss function

In remote sensing images, the imbalance between cultivated and non-cultivated land is particularly noticeable in mountainous areas. When using binary cross entropy loss (BCE Loss), equal weight is given to each class, which can cause the model to learn in the wrong direction [39]. On the other hand, Dice Loss [40] is better at extracting the foreground and is more suitable for unbalanced samples. However, the loss function presents a gradient instability problem, which can result in suboptimal convergence of training results [41]. Therefore, the BCE-Dice Loss function, which combines the Dice Loss and BCE Loss, is more suitable for measuring the fitness of predicted and actual values in cultivated land extraction results. The calculation formula is as follows:

$$L_{Dice} = 1 + \frac{1}{N}\sum_{i=1}^{N}\frac{2y_i y_i^{'}}{y_i + y_i^{'}} \tag{8}$$

$$L_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log y_i^{'} + (1-y_i)\log(1-y_i^{'})] \tag{9}$$

$$L_{BCE\_Dice} = L_{BCE} + L_{Dice} \tag{10}$$

Where y_i represents the true value of the i th pixel, y_i' represents the predicted value of the i th pixel, and N is the number of pixels.

3.7 Performance assessment

To evaluate the results of cropland extraction, four evaluation metrics based on the confusion matrix were utilized: Precision, Recall, F1-score, and Intersection over Union (IoU) are all metrics used to evaluate the performance of classification models. Precision is the ratio of true positive predictions to all positive predictions, Recall is the ratio of true positive predictions to all true positive values, F1-score is the harmonic mean of Precision and Recall, and IoU is the ratio of the intersection to the union of the predicted and true values. The formulas for calculating these metrics are as follows:

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{13}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{14}$$

Where TP is true positive, TN is true negative, FN is false negative, and FP is false positive.

4 Result analysis

The experiment is based on the TensorFlow 2.6 deep learning framework and uses Python 3.6 to execute the code. The computer hardware uses an Intel Core i7-11700F CPU, an NVIDIA GeForce RTX 3060 graphics card with 12 GB of video memory, and CUDA 11.2 accelerated computing. The computer's operating system is Windows 10.

The Adam optimizer was selected for training the model. Adam is an abbreviation for Adaptive Moment Estimation and is commonly used in deep learning tasks. It combines the advantages of adaptive learning rate methods and momentum-based methods, which help to alleviate the issues caused by gradient oscillation during training.

To examine the effect of batch size and learning rate on training and validation accuracy, we conducted various experiments with different parameter values. The results were analyzed and presented in Figure 5 and Figure 6, which display the training and validation accuracy curves, respectively.

Figure 5 illustrates the training accuracy curve over the course of training, while Figure 6 shows the validation accuracy curve. By analyzing these curves, one can gain insights into how various batch sizes and learning rates impact the model's performance. Upon analyzing the change curves of training accuracy and validation accuracy, it is evident that an increase in the learning rate leads to gradual improvement in training accuracy. However, the validation accuracy initially increases but then starts to decrease, indicating that a higher learning rate may result in faster convergence during training, leading to improved training accuracy but also potentially causing overfitting and a decrease in validation accuracy. When the learning rate is less than 5e-3, the validation accuracy exceeds the training accuracy. This suggests that the network model is overfitting, which means it is overly optimized for the training data and has difficulty generalizing to unseen data. As for the batch size, it is noted that setting it to 12 maximizes both the training and validation accuracy. It is suggested that a batch size of 12 strikes a balance between computational efficiency and model performance for the given task. Based on these observations, it is determined that the initial learning rate should be set to 5e-3 and the batch size should be 12. These values are chosen to optimize the training process and achieve the best possible accuracy on both the training and validation datasets.

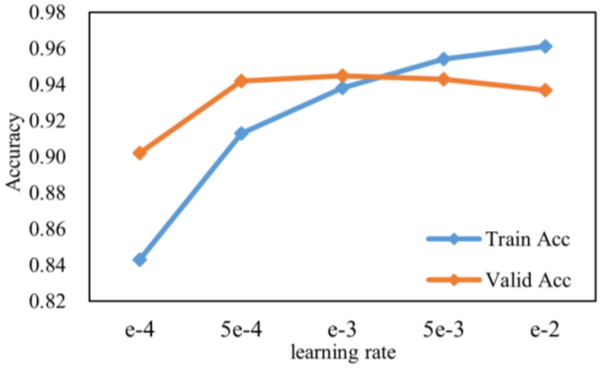


Figure 5: The influence curve of learning rate on training accuracy and verification accuracy

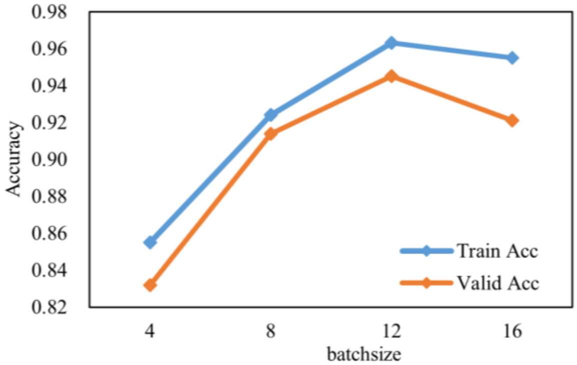


Figure 6: The influence curve of batch size on training accuracy and verification accuracy

We compared the proposed architecture with other state-of-the-art methods on datasets and verified its superiority through quantitative and qualitative results. The quantitative results are presented in Table 3. In remote sensing image segmentation, the network's performance on precision, recall, F1-score, and IoU metrics is of particular interest.

The evaluation results of the proposed architecture, GF-UNet, demonstrate its superior performance compared to classical semantic segmentation models such as U-Net, PSPNet, and DeepLab_v3+. GF-UNet achieved the highest Precision, F1-score, and IoU values among all the compared methods, with values of 0.9125, 0.9241, and 0.8456, respectively.

The effectiveness of the proposed architecture for remote sensing image segmentation tasks is highlighted by the significant improvement of GF-UNet over classical models.
This improvement can be attributed to the unique design choices and architectural modifications made in GF-UNet, which have enhanced its ability to accurately segment remote sensing images.

Additionally, GF-UNet outperforms two recent methods, MAENet and SGENet. This indicates that the proposed architecture outperforms not only traditional models but also more contemporary approaches, showcasing its state-of-the-art performance in remote sensing image segmentation.

The high Precision, F1-score, and IoU values achieved by GF-UNet demonstrate its ability to accurately identify positive samples, achieve a balance between precision and recall, and accurately capture the overlap between predicted and true positive areas. These metrics highlight the robustness and quality of the segmentation results produced by GF-UNet. Figure 7 displays the semantic segmentation outcomes of different methods. The first column presents the cropland image, the second column shows the cropland label and the remaining columns exhibit the segmentation results of various methods.

The performance of different methods is described qualitatively by analyzing the results in Figure 7. GF-UNet's segmentation results show clear edge features, accurate detail features, and the best overall result, making it the best performing network. On the other hand, U-Net, PSPNet, and DeepLab_v3+ tend to miss small area targets, resulting in poor segmentation results. The qualitative analysis of the segmentation results proves the superiority and effectiveness of GF-UNet.

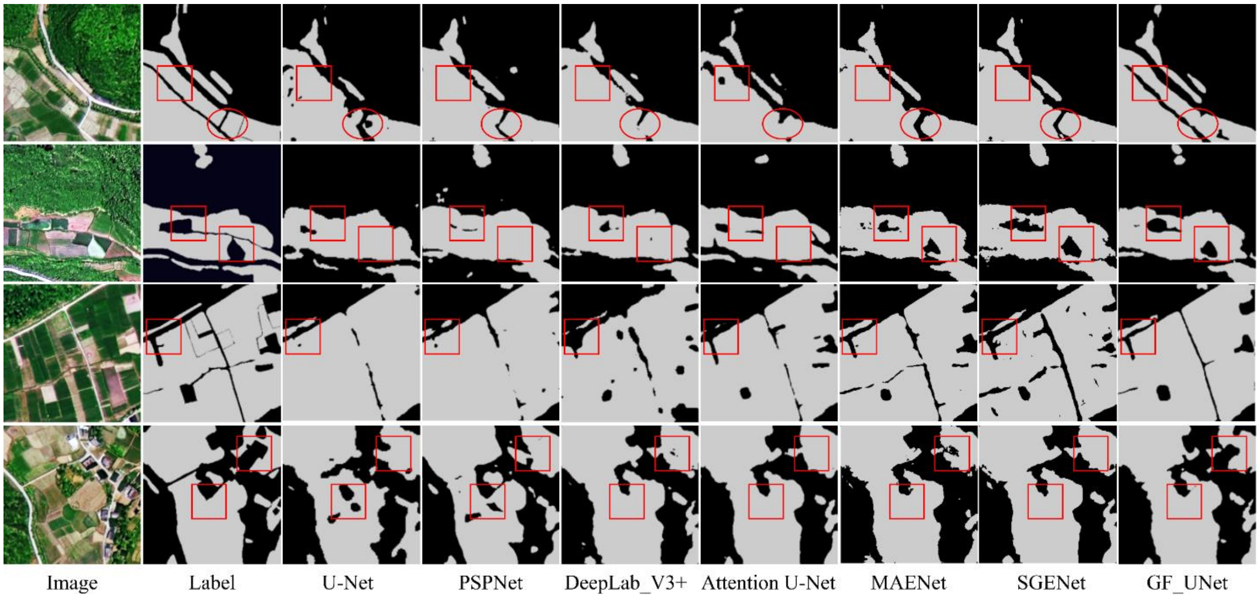


Figure 7: Visualization of cropland extraction results by multiple methods.

Table 3: Results of evaluation indicators of multiple methods.

Models	**Precision**	**Recall**	**F1-score**	**IoU**
U-Net[18]	0.7802±0.0172	0.8886±0.0187	0.8344±0.0179	0.7454±0.0142
PSPNet[19]	0.7435±0.0214	0.8737±0.0194	0.8086±0.0201	0.6716±0.0187
DeepLab_V3+[17]	0.8501±0.0206	0.8929±0.0218	0.8715±0.0197	0.7716±0.0162
Attention U-Net[21]	0.8826±0.0145	0.9434±0.0173	0.9130±0.0158	0.7946±0.0159
MAENet[30]	0.8875±0.0148	0.9007±0.0204	0.8941±0.0171	0.8125±0.0137
SGENet[29]	0.9014±0.0153	**0.9442±0.0185**	0.9228±0.0164	0.8087±0.0141
GF-UNet	**0.9125±0.0142**	0.9357±0.0180	**0.9241±0.0165**	**0.8456±0.0128**

5 Discussion

The spectral and textural characteristics of cropland vary due to complex cropland types, diverse crop varieties, and different phenological characteristics [42]. Additionally, VHR remote sensing commonly exhibits homogeneity and heterogeneity, further complicating the extraction of cropland characteristics [43].

This paper validates the proposed model using self-made cropland datasets. To increase the dataset's sample capacity, data preprocessing and data augmentation methods are used. The dataset includes four different scenarios to capture the variability present in cropland imagery. The performance of each method is analyzed by comparing the proposed model with other state-of-the-art (SOTA) models. This analysis compares the proposed model to existing approaches to determine its effectiveness and superiority.

Table 3 provides a comprehensive comparison of the performance metrics of the different models. GF-UNet stands out with a precision of 0.9125, outperforming MAENet and SGENet by 2.5% and 1.11%, respectively, indicating its superior ability to accurately identify positive samples. SGENet achieves the highest recall of 0.9442, slightly ahead of GF-UNet by 0.85%. While SGENet performs slightly better at capturing all true positive samples, the margin is relatively small. The F1 score, which balances precision and recall, highlights the superiority of GF-UNet over Attention U-Net and SGENet, with an F1 score of 0.9241. GF-UNet strikes a better balance between precision and recall, providing a more holistic assessment of its segmentation performance. In addition, GF-UNet achieves the highest Intersection over Union (IoU) value of 0.8456, outperforming MAENet and SGENet by 3.31% and 3.69%, respectively. The IoU metric indicates the accuracy and quality of the segmentation results, with GF-UNet demonstrating superior precision in capturing true positive areas. Both the F1 score and IoU serve as critical evaluation metrics for network models, with GF-UNet emerging as a top performer in both categories. These results underscore its exceptional overall performance and minimal disparity between predicted and actual results.

To investigate the factors that enhance the performance of the proposed model, ablation experiments were conducted using Attention U-Net as the base network to analyze the impact of AFFM and SFEM on model performance. The ablation effect was evaluated using IoU as the index. The results of the ablation experimental evaluation indexes are presented in Table 4, and the ablation experimental results are shown in Figure 8.

Table 4 shows that the addition of the SFEM module to Attention U-Net increases the model's IoU by 1.78%, resulting in a total of 81.24%. Similarly, the addition of the AFFM module to Attention U-Net increases the model's IoU by 3.4%, resulting in a total of 82.86%. Both modules were added separately to Attention U-Net and have been shown to improve the model's performance.

Table 4: Ablation results

Models	AFFM	SFEM	IoU
baseline	×	×	0.7946±0.0159
	√	×	0.8286±0.0173
	×	√	0.8124±0.0147
Our Model	√	√	**0.8456±0.0152**

In combination with the results of Figure 8, it is evident that the SFEM module enhances the clarity of the cropland edge features extracted by the model, particularly the portion connected to the farmland and buildings.

The SFEM module improves the clarity of cropland edge features extracted by the model, especially in areas where farmland and buildings are connected. It allows for more precise delineation of the boundaries between cropland and other structures, which is particularly noticeable in regions where cropland and buildings are adjacent or overlapping. The SFEM module refines the segmentation results by emphasizing distinctive features associated with cropland edges, resulting in clearer and more accurate boundary delineation.

The AFFM module improves the model's ability to distinguish between farmland and forest land with similar features. It integrates multi-scale features, enlarges the model's receptive field, and extracts richer semantic information. Additionally, the channel attention mechanism SE strengthens the model's learning ability and improves its performance.

Our proposed model utilizes both SFEM and AFFM, resulting in an increased IoU of 84.56%. By combining the advantages of SFEM and AFFM, our model not only enhances its learning ability but also retains more detailed information. Compared to other state-of-the-art models, our model achieves higher segmentation accuracy and more precise edge features.

However, during the experiment, we encountered an issue where our model tended to overlook small, isolated areas of farmland, resulting in segmentation loss. We need to improve the extraction of broken, irregularly shaped farmland with connected edges. Despite this, our model effectively enhances feature discrimination and preserves details. Our model can be applied to semantic segmentation in complex scenarios, which includes but is not limited to the extraction of farmland features.

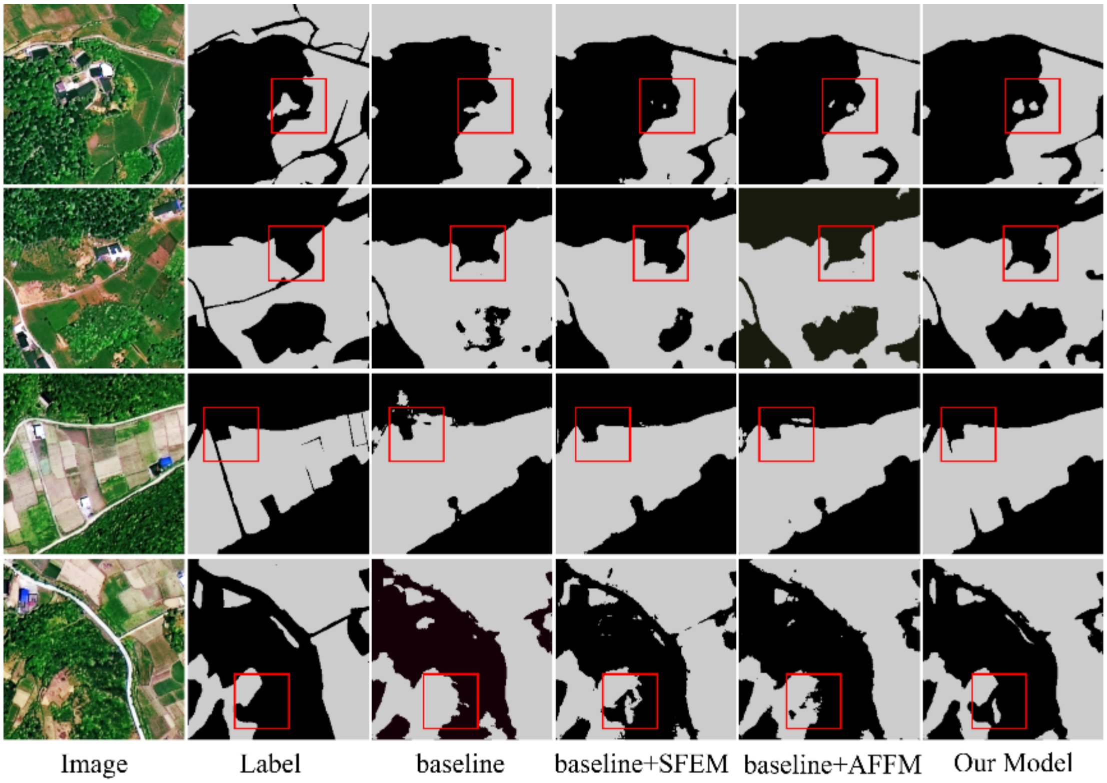


Figure 8: Visualization results of ablation experiment.

6 Conclusion

Due to the distinct spectral and textural features of cropland in high-resolution remote sensing images, the classical semantic segmentation network yielded inaccurate and incomplete results for cropland extraction. To address this issue, the GF-UNet network based on the Attention U-Net architecture is proposed. In GF-UNet, attention gates (AGs) are employed to enhance discriminative ability between partially cropped areas and non-cropland features in complex scenes. In order to improve the network's ability to extract different categories of cropland attributes, we utilize an adaptive feature fusion module (AFFM). Additionally, we introduce a skip connection layer with a spatial feature extraction module (SFEM) to refine detailed features extracted from intermediate layers. Our method is evaluated using GF-2 images captured in Xuan'en County, Hubei Province. The experimental results show that GF-UNet achieves an F1 score of 92.41% and a crossover ratio of 84.56%. Our proposed approach provides more accurate and comprehensive extraction of cropland information compared to SOTA methods. In the future, we will focus on incorporating phenological characteristics specific to different crops to improve categorization accuracy, considering the significant influence of crop types and phenological characteristics on cropland dynamics over time.

References

[1] Z. Y. Dong, J. H. Li, J. Zhang, J. Q. Yu, and S. An, "Cultivated land extraction from high-resolution remote sensing images based on BECU-Net model with edge enhancement," *National Remote Sensing Bulletin*, vol. 27, no. 12, pp. 2847-2859, 2023, https://doi.org/10.11834/jrs.20222268.

[2] W. Liu, Z. F. Wu, and J. C. Luo, "A divided and stratified extraction method of high-resolution remote sensing information for cropland in hilly and mountainous areas based on deep learning," *Acta Geodaetica et Cartographica Sinica*, vol. 50, 2021, https://doi.org/10.11947/j.AGCS.2021.20190448.

[3] B. Petrovska, T. A. Pacemska, N. Stojkovic, A. Stojanova, and M. Kocaleva, "Machine Learning with Remote Sensing Image Data Sets," *Informatica*, vol. 45, no. 3, pp. 347-358, 2021, https://doi.org/10.31449/inf.v45i3.3296.

[4] F. B. Wu, "China Crop Watch System with Remote Sensing," *National Remote Sensing Bulletin*, vol. 6, 2004, https://doi.org/10.11834/jrs.20040601.

[5] S. S. Panda, M. N. Rao, P. Thenkabail, and J.E. Fitzerald, "Remote Sensing Systems—Platforms and Sensors: Aerial, Satellite, UAV, Optical, Radar, and LiDAR," *Remotely Sensed Data Characterization, Classification, and Accuracies*, CRC Press, pp. 37-92, 2015, https://doi.org/10.1201/b19294.

[6] N. Zhou, P. Yang, and C. S. Wei, "Accurate extraction method for cropland in mountainous areas based on field parcel," *Transactions of the Chinese Society of Agricultural Engineering*, vol. 37, pp. 260-266, 2021, https://doi.org/10.11975/j.issn.1002-6819.2021.19.030.

[7] L. Samaniego, and K. Schulz, "Supervised Classification of Agricultural Land Cover Using a Modified k-NN Technique (MNN) and Landsat Remote Sensing Imagery," *Remote Sensing*, vol. 1, 2009, https://doi.org/10.3390/rs1040875.

[8] F. Waldner, et al, "Automated annual cropland mapping using knowledge-based temporal features," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 110, pp. 1-13, 2015, https://doi.org/10.1016/j.isprsjprs.2015.09.013.

[9] P. Teluguntla, et al, "A 30-m landsat-derived cropland extent product of Australia and China using random forest machine learning algorithm on Google Earth Engine cloud computing platform," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 144, pp. 325-340, 2018, https://doi.org/10.1016/j.isprsjprs.2018.07.017.

[10] X. Zhang, Q. Wang, G. Chen, et al, "An object-based supervised classification framework for very-high-resolution remote sensing images using convolutional neural networks," *Remote Sensing Letters*, vol. 9, no. 4, pp. 373-382, 2018, https://doi.org/10.1080/2150704X.2017.1422873.

[11] Y. F. Huang, C. Y. Lu, M. M. Jia, Z. L. Wang, Y. Su, Y. L. Su, "Plant species classification of coastal wetlands based on UAV images and object-oriented deep learning," *Biodiversity Science*, vol. 31, no. 3, 2023, https://doi.org/10.17520/biods.2022411.

[12] R. S. Khudeyer, and N. M. Almoosawi, "Combination of machine learning algorithms and Resnet50 for Arabic Handwritten Classification," *Informatica*, vol. 46, no. 9, pp. 39–44, 2023, https://doi.org/0.31449/inf.v46i9.4375.

[13] X. C. Zhang, J. F. Huang, and T. Ning, "Progress and prospect of cropland extraction from high-resolution remote sensing images," *Geomatics and Information Science of Wuhan University*, vol. 48, 2023, https://doi.org/10.13203/j.whugis20230114.

[14] E. M. Aminoff, S. Baror, E. W. Roginek, et al, "Contextual Associations Represented Both in Neural Networks and Human Behavior," *Scientific Reports*, vol. 12, no. 5570, 2022, https://doi.org/10.1038/s41598-022-09451-y.

[15] Y. Qing, and W. Liu, "Hyperspectral Image Classification Based on Multi-Scale Residual Network with Attention Mechanism," *Remote Sensing*, vol. 13, no. 3, pp. 335, 2021, https://doi.org/10.3390/rs13030335.

[16] Z. Liu, et al, "A multi-angle comprehensive solution based on deep learning to extract cropland information from high-resolution remote sensing images," *Ecological Indicators*, vol. 141, 2022, https://doi.org/10.1016/j.ecolind.2022.108961.

[17] Z. Du, J. Yang, and C. Ou, "Smallholder crop area mapped with a semantic segmentation deep learning method," *Remote Sensing*, vol. 11, no.7, 2019, https://doi.org/10.3390/rs11070888.

[18] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” *18th Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp 234-241, 2015, https://doi.org/10.1007/978-3-319-24574-4_28.

[19] J. H. Kim, S. H. Lee, and H. H. Han, “Modified pyramid scene parsing network with deep learning based multi scale attention,” *Journal of the Korea Convergence Society*, vol. 12, no.11, pp. 45-51, 2021, https://doi.org/10.15207/JKCS.2021.12.11.045.

[20] S. S. Li, Q. Cai, and Z. Z. Li, “Attention-aware invertible hashing network with skip connections,” *Pattern Recognition Letters*, vol. 138, pp. 556-562, 2020, https://doi.org/10.1016/j.patrec.2020.09.002.

[21] O. Oktay, et al, “Attention U-Net: Learning where to look for the pancreas,” *ArXiv*:1804.03999, https://doi.org/10.48550/arXiv.1804.03999.

[22] L. Xia, J. Luo, Y. Sun, and H. Yang, “Deep Extraction of Cropland Parcels from Very High-Resolution Remotely Sensed Imagery,” *7th International Conference on Agro-geoinformatics (Agro-geoinformatics)*, pp. 1-5, 2018, https://doi.org/10.1109/AgroGeoinformatics.2018.8476002.

[23] M. Yang, K. Yu, C. Zhang, et al, “Denseaspp for semantic segmentation in street scenes,” *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3684-3692, 2018, https://doi.org/10.1109/CVPR.2018.00388.

[24] R. R. Liu, F. Tao, and X. T. Liu, “RAANet: a residual aspp with attention framework for semantic segmentation of high-resolution remote sensing images,” *Remote Sensing*, vol. 14, pp. 18, 2022, https://doi.org/10.3390/rs14133109.

[25] R. Hong, J. Park, and S. Jang, “Development of a Parcel-Level Land Boundary Extraction Algorithm for Aerial Imagery of Regularly Arranged Agricultural Areas,” *Remote Sensing*, vol. 13, no.6, pp. 1167, 2021, https://doi.org/10.3390/rs13061167.

[26] J. Graesser, et al, “Detection of cropland field parcels from Landsat imagery,” *Remote sensing of environment*, vol. 201, no.8, pp. 165-180, 2021, https://doi.org/10.1016/j.rse.2017.08.027.

[27] D. J. Zhang, et al, “A generalized approach based on convolutional neural networks for large area cropland mapping at very high resolution,” *Remote Sensing of Environment*, vol. 247, pp. 23, 2020, https://doi.org/10.1016/j.rse.2020.111912.

[28] X. Song, H. Zhou, and G. Liu, “Study of Multiscale Fused Extraction of Cropland Plots in Remote Sensing Images Based on Attention Mechanism,” *Computational Intelligence and Neuroscience*, vol. 2022, 2022, https://doi.org/10.1155/2022/2418850.

[29] L. Xu, P. Yang, J. Yu, et al, “Extraction of cropland field parcels with high resolution remote sensing using multi-task learning,” *European Journal of Remote Sensing*, vol. 56, no. 1, pp. 218-230, 2020, https://doi.org/10.1080/22797254.2023.2181874.

[30] H. Huan, Y. Liu, Y. Xie, et al, “MAENet: Multiple Attention Encoder–Decoder Network for Farmland Segmentation of Remote Sensing Images,” *IEEE Geoscience and Remote Sensing Letters*, vol. 19, 2021, https://doi.org/10.1109/LGRS.2021.3137522.

[31] M. Y. Yu, et al, “AGs-Unet: Building Extraction Model for High Resolution Remote Sensing Images Based on Attention Gates U Network,” *Sensors*, vol. 22, pp. 21, 2022, https://doi.org/10.3390/s22082932.

[32] J. J. Liao, B. Zhou, Y. L. Chang, L. Zhang, “A dataset of mangrove forest changes on Hainan Island based on GF-2 data during 2015–2019,” *China Scientific Data*, vol. 7, no. 4, 2022, https://doi.org/10.11922/116035.noda.2021.0016.zh.

[33] D. P. Roy, M. A. Wulder, and T. R. Loveland, “Landsat-8: science and product vision for terrestrial global change research,” *Remote sensing of Environment*, vol. 145, pp. 154-172, 2014, https://doi.org/10.1016/j.rse.2014.02.001.

[34] X. Jin, Y. Xie, and X. S. Wei, “Delving deep into spatial pooling for squeeze-and-excitation networks,” *Pattern Recognition*, vol. 121, pp. 154-172, 2022, https://doi.org/10.1016/j.patcog.2021.108159.

[35] J. Ji, et al, “Semantic Image Segmentation with Propagating Deep Aggregation,” *IEEE Transactions on Instrumentation and Measurement*, vol. 69, 2020, https://doi.org/10.1109/TIM.2020.3004902.

[36] Q. Zhao, J. Liu, Y. Li, and H. Zhang, “Semantic segmentation with attention mechanism for remote sensing images,” *IEEE Transactions on Geoscience Remote Sensing*, vol. 60, pp. 1-13, 2021, https://doi.org/10.1109/TGRS.2021.3085889.

[37] W. Wang, X. Tan, and P. Zhang, “A CBAM based multiscale transformer fusion approach for remote sensing image change detection,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 6817-6825, 2022, https://doi.org/10.1109/JSTARS.2022.3198517.

[38] W. Deng, et al, “Attention-gate-based encoder–decoder network for automatical building extraction,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, 2021, https://doi.org/10.1109/JSTARS.2021.3058097.

[39] D. H. Xie, H. Xu, and X. L. Xiong, “Cropland Extraction in Southern China from Very High-Resolution Images Based on Deep Learning,” *Remote Sensing*, vol. 15, no.5, pp. 123-132, 2022, https://doi.org/10.3390/rs15092231.

[40] F. Milletari, et al, “In V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation,” *2016 Fourth International Conference on 3D Vision (3DV)*, pp. 565-571, 2016, https://doi.org/10.1109/3DV.2016.79.

[41] M. Yeung, E. Sala, CB. Schönlieb, et al, “Unified Focal loss: Generalising Dice and cross entropy-based losses to handle class imbalanced medical image segmentation,” *Computerized Medical Imaging and Graphics*, vol. 95, pp. 102-120, 2022,

https://doi.org/10.1016/j.compmedimag.2021.102026.

[42] L.I. Qiannan, Z. Dujuan, P. Yaozhong, et al, "High-resolution cropland extraction in Shandong province using MPSPNet and UNet network," *National Remote Sensing Bulletin*, vol. 27, no. 2, pp. 471-491, 2023, http://doi.org/10.11834/jrs.20210478.

[43] Z. Shao, "Emerging Issues in Mapping Urban Impervious Surfaces Using High-Resolution Remote Sensing Images," *Remote Sensing*, vol. 15, 2023, http://doi.org/10.3390/rs15102562.