remote sensing

MDPI

Article

GLCANet: Global–Local Context Aggregation Network for Cropland Segmentation from Multi-Source Remote Sensing Images

Jinglin Zhang [1], Yuxia Li [1,*], Zhonggui Tong [1], Lei He [2], Mingheng Zhang [2], Zhenye Niu [1] and Haiping He [1]

[1] School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China; 202221060613@std.uestc.edu.cn (J.Z.); 202121060324@std.uestc.edu.cn (Z.T.); 202321060615@std.uestc.edu.cn (Z.N.); 202322060616@std.uestc.edu.cn (H.H.)

[2] School of Software Engineering, Chengdu University of Information Technology, Chengdu 610225, China; helei1978@cuit.edu.cn (L.H.); 3220707029@stu.cuit.edu.cn (M.Z.)

* Correspondence: liyuxia@uestc.edu.cn

Abstract: Cropland is a fundamental basis for agricultural development and a prerequisite for ensuring food security. The segmentation and extraction of croplands using remote sensing images are important measures and prerequisites for detecting and protecting farmland. This study addresses the challenges of diverse image sources, multi-scale representations of cropland, and the confusion of features between croplands and other land types in large-area remote sensing image information extraction. To this end, a multi-source self-annotated dataset was developed using satellite images from GaoFen-2, GaoFen-7, and WorldView, which was integrated with public datasets GID and LoveDA to create the CRMS dataset. A novel semantic segmentation network, the Global–Local Context Aggregation Network (GLCANet), was proposed. This method integrates the Bilateral Feature Encoder (BFE) of CNNs and Transformers with a global–local information mining module (GLM) to enhance global context extraction and improve cropland separability. It also employs a multi-scale progressive upsampling structure (MPUS) to refine the accuracy of diverse arable land representations from multi-source imagery. To tackle the issue of inconsistent features within the cropland class, a loss function based on hard sample mining and multi-scale features was constructed. The experimental results demonstrate that GLCANet improves OA and mIoU by 3.2% and 2.6%, respectively, compared to the existing advanced networks on the CRMS dataset. Additionally, the proposed method also demonstrated high precision and practicality in segmenting large-area croplands in Chongzhou City, Sichuan Province, China.

Keywords: semantic segmentation; remote sensing; multi-source image; cropland; deep learning

check for updates

Citation: Zhang, J.; Li, Y.; Tong, Z.; He, L.; Zhang, M.; Niu, Z.; He, H. GLCANet: Global–Local Context Aggregation Network for Cropland Segmentation from Multi-Source Remote Sensing Images. *Remote Sens.* **2024**, *16*, 4627. https://doi.org/10.3390/rs16244627

Academic Editor: Jocelyn Chanussot

Received: 1 November 2024
Revised: 4 December 2024
Accepted: 5 December 2024
Published: 10 December 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

As the global population grows, protecting and utilizing cropland resources faces unprecedented challenges [1,2]. Strengthening the acquisition and protection of cropland information, enhancing agricultural efficiency, and promoting scientific land use are vital strategies for food security and addressing the global food crisis. For example, in China, with its vast but unevenly distributed cropland, traditional ground-based statistical methods for data collection are often time-consuming and costly [2,3]. Monitoring the conversion of cropland to non-agricultural and non-cereal land is crucial. Using remote sensing images for multi-class land cover segmentation helps enable the future change detection and monitoring of land use transformations. The GaoFen and WorldView satellites offer high-resolution Earth observation imagery, supporting diverse applications like urban planning and environmental monitoring. For example, GaoFen focuses on large-scale optical and radar data, while WorldView provides ultra-high-resolution imagery, reaching up to 30 cm, ideal for detailed mapping and precise analysis. However, the extensive

coverage, diverse appearances, and complex classifications of croplands lead to significant variations in semantic features, complicating extraction processes [4,5]. To address these challenges related to scale, color, and texture variability in remote sensing imagery, a deep learning-based semantic segmentation method is needed to integrate local features with the global context of the remote sensing images.

In recent years, deep learning-based semantic segmentation methods leveraging convolutional neural networks have seen substantial advancements. SegNet [6] utilizes a pooling index mechanism to more precisely restore positional information during the upsampling stage, thereby enhancing segmentation accuracy. UNet [7], with its U-shaped architecture and skip connections, preserves image details and has demonstrated exceptional performance, particularly in medical image segmentation. PSPNet [8] and Deeplabv3+ [9] improve the understanding of complex scenes by aggregating multi-scale contextual information through pyramid pooling and atrous spatial pyramid pooling (ASPP). Models such as HRNetV2 [10], D-LinkNet [11], and Eff-UNet++ [12] have further improved segmentation accuracy and feature extraction capabilities through enhanced network architectures and convolutional strategies. Additionally, for very high-resolution (VHR) remote sensing images, studies [13,14] have specifically designed modified U-Net architectures, which, by incorporating refined skip connections and an atrous spatial pyramid pooling (ASPP) module, effectively enhance segmentation accuracy and edge-handling capabilities.

In addition, Transformer architecture has become a key focus in the current research on deep learning for semantic segmentation tasks. The self-attention mechanism in Transformer models [15] enables capturing dependencies between any two points within a sequence while allowing for efficient parallel computation. In 2020, Transformers were first applied to computer vision; the Vision Transformer (ViT) demonstrated exceptional performance in image classification tasks by segmenting images into patches and processing them as 1D sequences. Following this, researchers began to explore ViT's application in semantic segmentation. Given that semantic segmentation requires classifying every pixel in an image, it demands strong spatial understanding and sensitivity to details from the model. Segmenter [16] combines Transformer encoder and decoder structures to leverage its global information processing capabilities for precise pixel-level predictions. Swin Transformer [17] incorporates a hierarchical structure and window-based self-attention mechanism to effectively handle multi-scale characteristics, exhibiting impressive performance in semantic segmentation. Moreover, several studies [18–20] have integrated Transformers with U-Net structures, enhancing the model's capability to understand and process complex scenes.

The rapid development of deep learning techniques has also significantly advanced research in the field of cropland information extraction from remote sensing data. Numerous studies [21–26] have demonstrated the effectiveness of deep learning-based models for remote sensing information extraction, showing marked improvements in land cover classification and feature extraction [27–30], thereby enhancing the semantic segmentation of remote sensing images. Wang et al. [31] proposed the UNet-former model, which integrates the UNet structure with a CNN-based encoder and a Transformer-based decoder, showing advantages in remote sensing land cover extraction within urban environments. Linus et al. [32] achieved comparable performance to fully supervised methods by pretraining a Vision Transformer on large-scale, unlabeled remote sensing data in a self-supervised manner, thus requiring only about one-tenth of the labeled dataset for effective training. Ding et al. [33] introduced the Wide Context Network (WiCoNet), which adds an additional Transformer-based contextual branch to traditional CNN local feature extraction, enabling the aggregation of information from image regions broader than those input to the local branch. This context transformation helps overcome the locality limitations of convolutional neural network modules.

The bilateral encoder (BE) framework has gained significant traction in semantic segmentation due to its ability to effectively integrate high-level semantic and low-level spatial features. This approach employs dual pathways, often designed to process different scales

of information, thereby enhancing segmentation accuracy. Recent advancements include the use of Densely Aggregated Bilateral Networks, which excel in real-time applications by maintaining computational efficiency while preserving high accuracy [34]. FBSNet effectively balances segmentation performance by combining a symmetrical encoder–decoder for semantic features with shallow layers for spatial detail preservation, ensuring high accuracy and efficiency [35].

While considerable progress has been made in cropland information extraction from remote sensing images using deep learning methods, there remain numerous challenges in applying these methods to large-area, real-world cropland extraction tasks.

First, the use of single-source imagery and limited sample sizes constrains the performance of deep learning models. The existing remote sensing datasets for cropland information extraction typically consist of images from a single source, with relatively fixed spatial resolutions and texture styles, leading to limited image diversity and the insufficient reflection of land cover distribution in real-world applications. Second, the inherent attributes of croplands contribute to interclass confusion and unclear boundaries. In China, for instance, cropland usage is highly intensive while dispersed, with many parcels characterized by small, fragmented plots.

Additionally, certain surface elements surrounding croplands have structural and spectral characteristics that are highly similar to, or even indistinguishable from, cropland, which can lead to interclass confusion during extraction, as illustrated in Figure 1a. This interclass similarity increases the difficulty of accurately delineating cropland boundaries, requiring deep learning models to integrate both local features and global context to mitigate the resulting errors. Lastly, the variability in cropland feature scales across multi-source images further complicates extraction. Differences in imaging conditions, including sensor type, angle, season, and climate, cause shifts in cropland feature distributions across different images, as shown in Figure 1b. Additionally, variations in the inherent resolution of multi-source remote sensing images result in differing scales of cropland representation for the same area, as illustrated in Figure 1c. The combination of these factors further reduces the separability of croplands in remote sensing imagery.

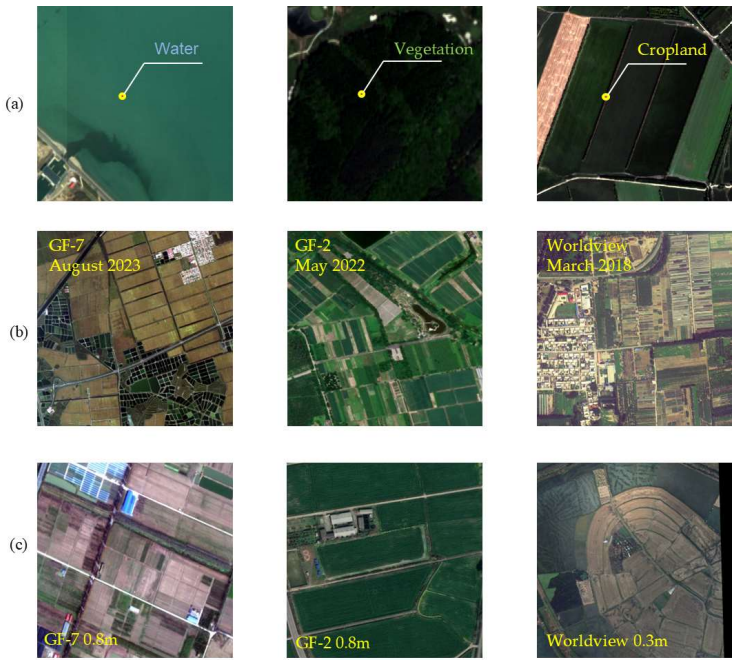


Figure 1. Issues present in multi-source remote sensing imagery. (**a**) Confusion between croplands and similar land cover types. (**b**) Variability in cropland features caused by different imaging conditions. (**c**) Variation in cropland scales due to differences in resolution and geographical conditions.

To address the issues, this study proposes GLCANet (Global–Local Context Aggregation Network), a semantic segmentation network for croplands that combines multi-scale global context and local feature information. By thoroughly exploring and integrating global and local information, GLCANet ensures precise extraction and global connectivity modeling, while introducing multi-scale progressive upsampling to enhance the network's capability to fuse information at various scales. In addition, general semantic segmentation methods are typically tested on datasets with sizes of 1024×1024 or smaller. To evaluate the accuracy of GLCANet on a broader scale, such as at an administrative region level, we conducted data collection and experiments in Chongzhou City, Sichuan Province, China. The main contributions of this paper are as follows:

1. We construct a self-labeled dataset for cropland information extraction, termed TriCities. Compared to the existing datasets, TriCities offers advantages such as diverse scenes, varying resolution spans, high labeling accuracy, and applicability for cropland information extraction and change monitoring. Additionally, based on TriCities, along with the public datasets GID and LoveDA, a multi-source imagery dataset for cropland information extraction, CRMS, has been created, significantly enhancing the data diversity and realism of simulations for real-world scenarios.

2. We present GLCANet, a semantic segmentation network designed to integrate multi-scale global context and local feature information for improved cropland extraction accuracy. This network is optimized to manage the complexities of large-area cropland information extraction, ensuring precise segmentation and enhanced global connectivity.

3. GLCANet features a Bilateral Feature Encoder (BFE) with parallel CNNs and Transformer branches for extracting global and local features. The Global–Local Mining (GLM) module enhances local–global integration during decoding, and the multi-scale progressive upsampling (MPUS) structure merges multi-scale features, progressively restoring resolution for the accurate pixel-level extraction of land cover types, including cropland. Additionally, we developed a novel loss function to assist and enhance the training of the network.

4. For the large-area remote sensing image extraction of cropland, we collect extensive multi-source remote sensing data from the Chongzhou City of Sichuan Province in China. Regional experiments were conducted to validate the feasibility and efficiency of the proposed method.

2. Methods

To address challenges in large-area cropland information extraction from multi-source remote sensing imagery, such as feature distribution shifts, scale variations, and fragmentation, we propose GLCANet, a semantic segmentation network. This network integrates multi-scale global context and local feature information to enhance extraction accuracy, improve multi-scale capabilities, and reduce class confusion.

2.1. The Overall Structure of GLCANet

GLCANet uses a Bilateral Feature Encoder (BFE) combining ResNet and SegFormer to extract multi-scale features from remote sensing images, adopting a UNet-like encoding–decoding structure. It introduces the Global–Local Mining (GLM) Block to enhance decoding with both global self-attention and local information, and the multi-scale progressive upsampling structure (MPUS) to replace UNet's skip connections, focusing on parameterizable upsampling and weighted feature fusion. Additionally, it uses a loss function that integrates hard sample mining and multi-scale prediction to improve the separation of similar land cover types and enhance the overall segmentation accuracy.

The network structure of GLCANet is illustrated in Figure 2. This network features a UNet-style multi-scale encoding–decoding architecture with skip connections. GLCANet receives as input an optical remote sensing image of size $H \times W \times 3$ and outputs an image of size $H \times W \times 1$, which includes pixel-level predictions for croplands and other land cover types.

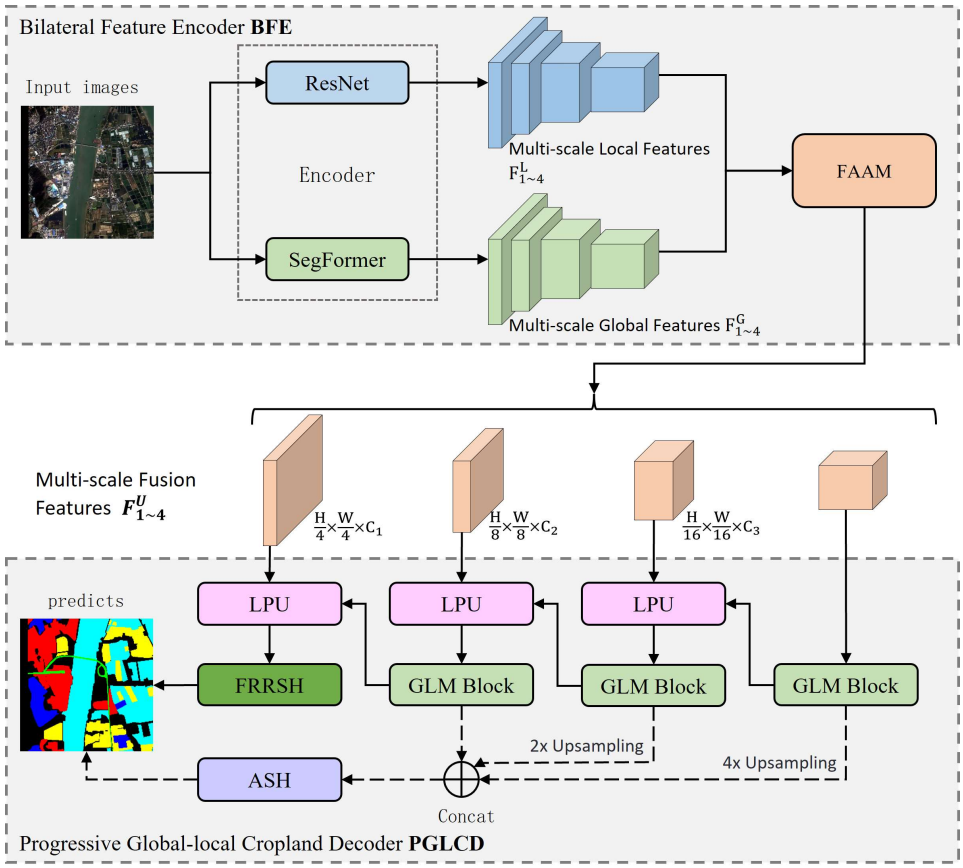


Figure 2. The overall structure of GLCANet.

Specifically, the optical remote sensing image is first processed through a dual-branch encoder composed of ResNet and SegFormer, which are responsible for extracting multi-scale local encoded features $F_i^L(i = 1, 2, 3, 4)$ and multi-scale global encoded features $F_i^G(i = 1, 2, 3, 4)$. Subsequently, these multi-scale local and global encoded features are fed into a feature aggregation attention module. This module employs an attention mechanism to balance and complement global and local features from the encoder, facilitating effective feature selection for cropland information extraction, and ultimately outputs multi-scale fusion features $F_i^U \in \mathbb{R}^{H_i \times W_i \times C_i}, i = 1, 2, 3, 4$.

In the decoding phase, the Global–Local Mining Block (GLM Block) and the multi-scale progressive upsampling structure (MPUS) comprise the decoder portion. The multi-scale fusion features F_i^U serve as the input to the decoder. The GLM Block further mines global–local fusion information beneficial for cropland information extraction. MPUS consists of a learnable progressive upsampling (LPU) and a feature refinement residual segmentation head (FRRSH).

The GLM Block and multi-scale fusion features are used to progressively restore feature map sizes to the original image dimensions. The FRRSH then converts these into pixel-level segmentation results for croplands and other categories. During training, an auxiliary prediction module (ASH) generates additional segmentation results from lower-scale feature maps. These auxiliary predictions are combined with the main predictions under the supervision of the loss function, enhancing the overall accuracy.

2.2. Bilateral Feature Encoder

Cropland extraction from remote sensing imagery requires capturing both global consistency among dispersed areas and local pixel relationships. While networks using attention and Transformers [36–38] aim to extract global context, they face the following challenges: (1) attention-based methods relying on CNNs struggle to fully explore

global context; (2) Transformers excel in global relationships but may neglect local feature extraction, creating an imbalance.

To ensure that the network effectively encodes and extracts both local and global context information related to croplands in remote sensing imagery, we combine ResNet and SegFormer models to propose a Bilateral Feature Encoder (BFE) that integrates both CNN and Transformer architectures. The design of the BFE features a ResNet branch for extracting local feature information and a SegFormer branch for capturing global context, creating a parallel encoder structure. The feature aggregation attention module then fuses these branches to obtain multi-scale fusion features.

The use of skip connections to transfer encoder features to the decoding phase has been validated as an effective strategy in UNet architecture, preserving shallow features and preventing their loss during the encoding–decoding process, which is crucial for the precise extraction of croplands from remote sensing imagery. However, the two branches of the proposed BFE focus on diverse types of features, and directly inputting bilateral features increases the computational burden on the decoder. If these features are propagated directly through skip connections to the decoder, it could interfere with the final predictions for cropland information extraction.

To effectively fuse the multi-scale encoded feature information from both the ResNet and SegFormer branches, a feature aggregation attention module (FAAM) is designed within the BFE, as illustrated in Figure 3. This module facilitates the integration of features, enhancing the decoder's ability to leverage both local and global information for improved cropland segmentation outcomes.

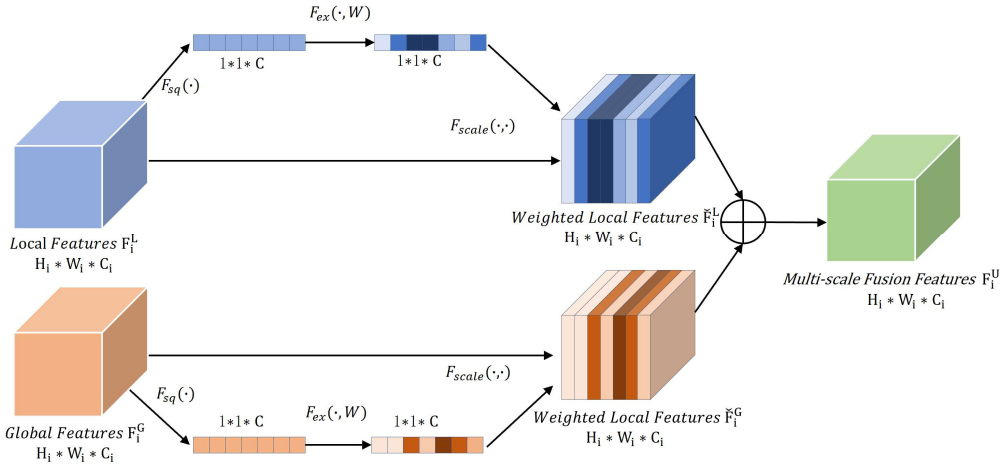


Figure 3. Feature aggregation attention module (FAAM).

In the feature aggregation attention module, the initial calibration of feature maps is performed, followed by the fusion of global and local encoded features. An attention mechanism calibrates the two feature maps, generating feature vectors that maintain the same channel dimensions as the inputs. These vectors learn channel dependencies and adjust the feature maps, with parameters updated during training to filter noise and highlight key information. This fusion applies to four scales of features from the Bilateral Feature Encoder (BFE), providing multi-resolution encoded features. The feature vector generation involves Squeeze and Excitation processes. The Squeeze operation compresses the original feature map to a dimension of $1 \times C \times C$, as shown in Equation (1):

$$Z_c = F_{sq}(F_c) = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} F_c(i, j) \quad (1)$$

where Z_c is the compressed vector, with c representing the $c - th$ element of Z, and W and H denoting the width and height of the original feature map, respectively. The Excitation step consists of two fully connected layers with nonlinear activation functions. The feature

vector learns the necessary weight information from the Excitation process, and Equation (2) provides the formula for this activation. The final fused joint feature information F_i^U is the result of adding the recalibrated global and local feature maps, which is shown in Equation (3).

$$S = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2, \delta(W_1, z)) \quad (2)$$

$$F_i^U = S_L \cdot F_i^L + S_G \cdot F_i^G \quad (3)$$

2.3. Global–Local Mining Block

To achieve comprehensive control of cropland features in remote sensing imagery, it is crucial to extract global context during the decoding process, as it helps to understand the relationships between cropland areas and the overall image. At the same time, retaining local information is equally important, as it contains rich spatial detail. To address both of these needs, this study proposes an innovative Global–Local Mining Block (GLM Block), as illustrated in Figure 4.

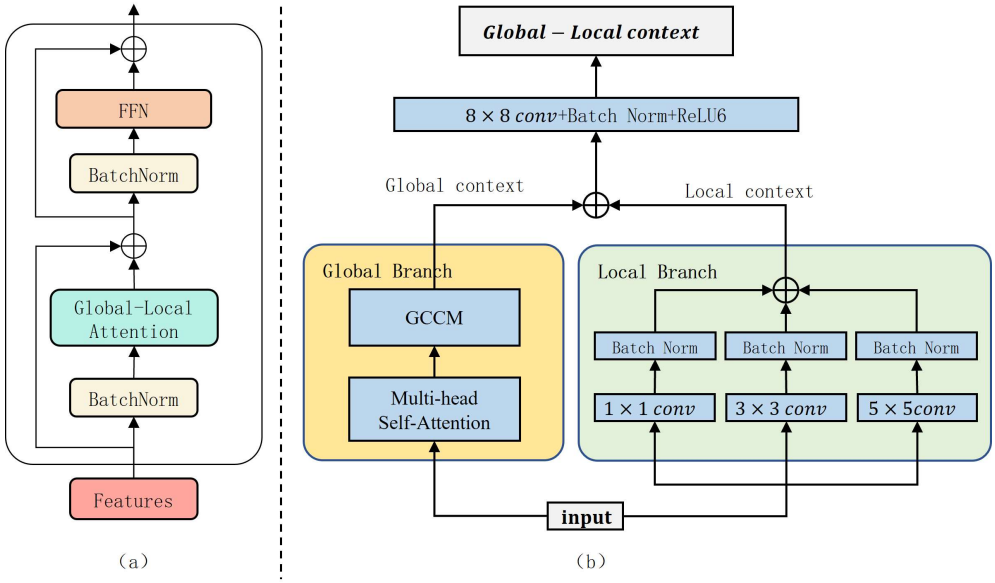


Figure 4. (**a**) Structure of the Global–Local Mining Block. (**b**) Structure of the Global–Local Attention.

The GLM Block, based on a Transformer structure, consists of two branches: a local information branch for extracting detailed features and a global information branch for capturing global context. This dual design enhances spatial connectivity and precise cropland information extraction while maintaining contextual richness.

In the local information branch, this study employs a design of multiple parallel convolutional layers that utilize convolutional layers with different kernel sizes to capture detailed information about objects at various scales. Specifically, three parallel convolutional layers are designed with kernel sizes of 1×1, 3×3, and 5×5. This multi-scale convolution strategy enables the network to learn features at different scales, aiding in the capture of detailed information about multi-scale cropland objects.

For the global information branch, a compressed efficient multi-head attention layer is designed as a global information miner, focusing on learning global information from the decoder feature maps. This global information branch employs a window-based multi-head self-attention mechanism to capture global context.

The window-based multi-head self-attention (W-MSA) mechanism enhances the model's understanding of image content by reducing computational costs through local window operations while preserving global context awareness. In this mechanism, the input feature map is divided into non-overlapping windows where self-attention is computed independently. For remote sensing images, a window size of 8×8 is used, with

each of the 8 attention heads having a channel dimension of 64, capturing both local and global information effectively.

Channel Attention and window-based multi-head self-attention (W-MSA) excel in different areas. Channel Attention is lightweight and focuses on enhancing inter-channel dependencies, making it effective for tasks requiring feature discrimination but less suited for capturing spatial relationships. In contrast, W-MSA is designed to capture both local and global spatial dependencies efficiently by operating within fixed windows, making it ideal for high-resolution tasks. However, it is computationally heavier and less effective at modeling inter-channel dependencies. These mechanisms complement each other in different segmentation scenarios.

Although the window-based self-attention mechanism effectively reduces computational load, it still faces the challenge of insufficient information interaction across windows, which is particularly critical when dealing with remote sensing images that possess rich spatial relationships. To overcome this limitation, we designed a Global Connection Compression Module (GCCM) that compresses and fuses feature information between windows from both horizontal and vertical directions, as illustrated in Figure 5.

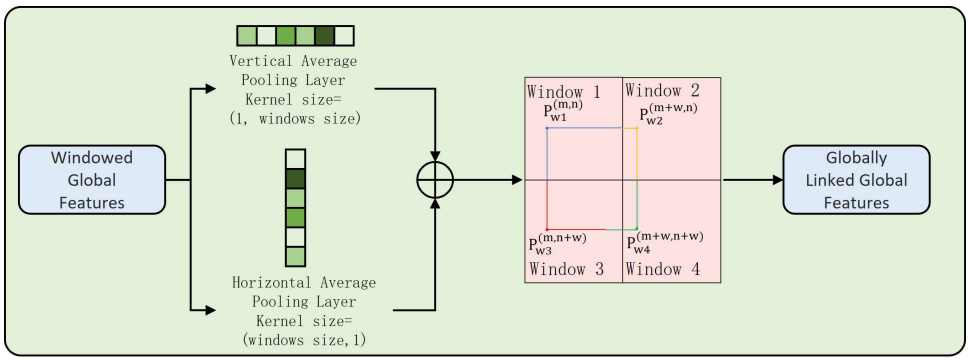


Figure 5. Global Connection Compression Module.

As shown in Figure 5, the core idea of the horizontal and vertical global connection modules is to establish connections between windows in the horizontal and vertical directions through average pooling layers. First, the global feature map is processed through a horizontal average pooling layer, aiming to establish horizontal relationships between local windows. For any point $P(m, n)$ within a window and its corresponding point $P(m + w, n)$ in an adjacent horizontal window, the dependency modeling process achieved through the horizontal average pooling layer can be described by the following Equations (4)–(7):

$$P_{w1}^{(m,n)} = \frac{\sum_{i=0}^{w-m-1} P_{w1}^{(m+1,n)} + \sum_{j=0}^{m} P_{w2}^{(m+w-j,n)}}{w} \tag{4}$$

$$P_{w1}^{(m+i,n)} = D_i\left(P_{w1}^{(m,n)}\right) \tag{5}$$

$$P_{w2}^{(m+w-j,n)} = D_j\left(P_{w2}^{(m+w,n)}\right) \tag{6}$$

$$P_{w1}^{(m,n)} = \frac{\sum_{i=0}^{w-m-1} D_i(P_{w1}^{(m,n)}) + \sum_{j=0}^{m} D_j(P_{w2}^{(m+w,n)})}{w} \tag{7}$$

In the above equations, w represents the size of the window, while D_i and D_j denote the operations that model the dependencies between pixels within a single local window during multi-head self-attention calculations. By this means, this study not only establishes dependencies between pixels within the window but also builds horizontal dependencies between adjacent windows through the horizontal average pooling layer. The processing method for the vertical direction is similar, using a vertical average pooling layer to establish

dependencies between windows in the vertical direction, thus effectively capturing global contextual information.

Cross-window dependency modeling enables the model to maintain local attention while integrating information between distant areas, enhancing performance in cropland information extraction. After computing features from both global and local branches, their outputs are summed to aggregate comprehensive global–local contextual information. This study uses deep convolution, batch normalization, and 1×1 convolutions to represent this detailed information accurately.

2.4. Multi-Scale Progressive Upsampling Structure

To address the multi-scale presentation of croplands due to varying resolutions in multi-source imagery and the differences in plot sizes across different terrains, this study constructs a robust multi-scale feature aggregation structure for the decoding process of GLCANet. The goal is to effectively extract croplands across different scales in multi-source imagery by integrating multi-scale features. In the process of aggregating multi-scale feature maps, upsampling is a crucial step aimed at restoring the image resolution that was reduced during downsampling in the encoding phase.

A traditional multi-scale upsampling structure typically concatenates representations of different scales before performing upsampling. However, this approach separates the upsampled features from the subsequent decoding process, which may result in the loss of fine-grained details. To address this issue, we propose the multi-scale progressive upsampling structure (MPUS), as illustrated in Figure 6. MPUS consists of two modules: the learnable progressive upsampling (LPU) and the feature refinement residual segmentation head (FRRSH).

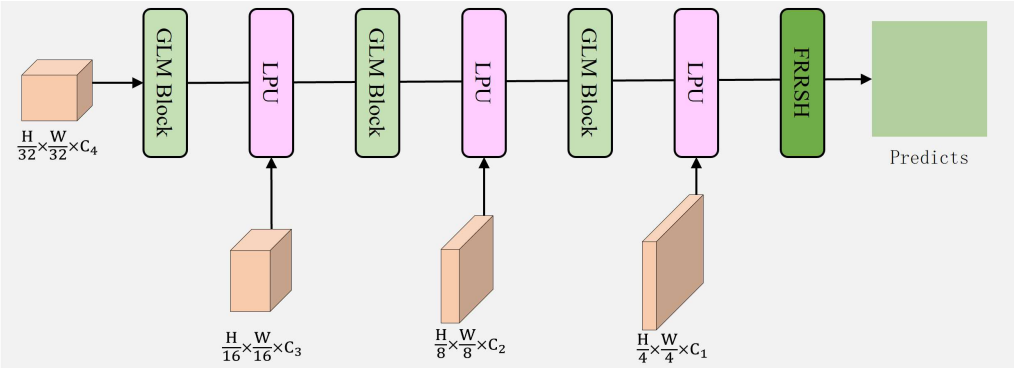


Figure 6. The multi-scale progressive upsampling structure.

The LPU modules interspersed between GLM Blocks in the decoder enable progressive multi-scale fusion. Each LPU upsamples the previous GLM Block's output to the next scale and fuses it with corresponding encoder features. Instead of bilinear interpolation, learnable transpose convolution is used for upsampling, offering greater flexibility and adaptability [39]. This method, combined with the GLM Block, effectively restores both coarse semantic features and fine details. For fusion, a self-learning weighted mechanism computes contribution weights, ensuring a more generalized fused feature. The formula for this fusion mechanism is the following Equation (8):

$$Fusion = \sum_i \frac{W_i}{\varepsilon + \sum_j W_j} \cdot I_i \tag{8}$$

Among them, $\sum_i$ represents the accumulation of all the feature maps. $\frac{W_i}{\varepsilon+\sum_j W_j}$ calculates the weighted ratio for each feature map, where W_i is the weight of the feature map i, and ε is a small quantity used to prevent the denominator from being zero. I_i represents feature map i. The weights W_i of the feature maps are calculated through the

ReLU function, which ensures that the output value is non-negative and further normalizes the weight values between 0 and 1, achieving rapid and effective feature fusion. Through this method, the fused feature maps will serve as the input for the next GLM Block layer. By applying learnable progressive sampling strategies in different layers, it realizes the gradual integration and transmission of multi-scale encoder and decoder feature maps.

In the final step of fusion in MPUS, the shallow features produced by the first layer of the bilateral encoder are combined with the features from the deepest GLM Block. The shallow features from the bilateral encoder retain rich spatial details of the remote sensing of croplands but lack semantic information. In contrast, the deepest GLM Block features provide more precise semantic information but have coarse spatial characteristics. If the differences between semantic and spatial information are not addressed, directly using the fused features to generate a pixel-level prediction map for croplands will likely reduce the accuracy of cropland segmentation [40]. To tackle this issue, this study employs the fundamental principles of attention mechanisms to design a feature refinement residual segmentation head (FRRSH). This component aims to narrow the semantic gap between the two feature types and refine the output features from the fusion process, thereby further enhancing the accuracy of cropland information segmentation.

As shown in Figure 7, the feature refinement residual segmentation head (FRRSH) comprises three components: the spatial branch, the channel branch, and the residual branch. Its purpose is to enhance the semantic information fusion and feature representation in both the channel and spatial directions. The spatial branch utilizes a 3×3 depthwise separable convolution to generate a spatial attention map $S \in \mathbb{R}^{h \times w \times 1}$, which strengthens the feature representation capabilities of spatial positions. The channel branch employs a global average pooling layer to produce a channel-level attention map $C \in \mathbb{R}^{1 \times 1 \times c}$. This is followed by two 1×1 convolution layers that perform a compression–expansion operation to enhance feature representation in the channel direction. These two attention maps are used to apply attention weighting to the input fused features through matrix multiplication, and the resulting attention features from both branches are further combined using summation. Finally, a residual skip connection adds the output of the residual branch to the fused features, employing 1×1 convolutions and upsampling operations to map the number of channels in the feature map to the number of segmentation classes N, resulting in the final segmentation output map.

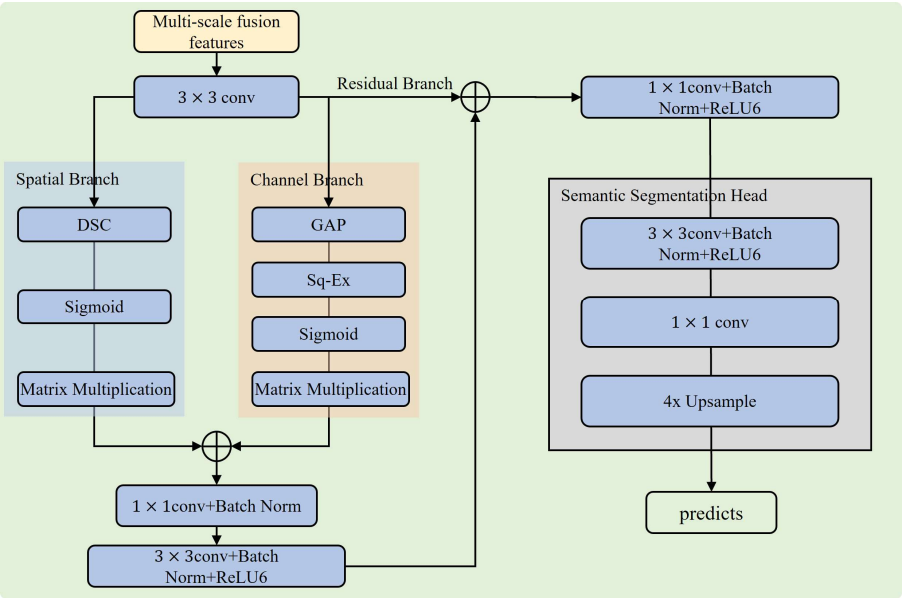


Figure 7. The structure of feature refinement residual segmentation head (FRRSH).

2.5. Multi-Constraint Difficult Sample Mining Loss Function

Given the fragmented and diverse visual representations of croplands in remote sensing images, along with class imbalance, this study observed that certain cropland samples significantly deviate from the overall feature distribution. This deviation hinders accurate cropland extraction during GLCANet training, affecting its generalization. Additionally, disparities exist between easily segmentable and challenging samples, especially in the later training stages where optimization focuses on difficult cases. Thus, developing a strategy to optimize these challenging samples is crucial for enhancing model performance in cropland extraction tasks.

This paper proposes a difficult sample mining mechanism that aims to optimize the network's learning of difficult samples through two steps: difficult sample mining estimation and dynamic confidence adjustment. The difficult sample mining estimation seeks to assign a weight to each pixel instance during the training process to reflect its classification difficulty, thereby adjusting the distribution of the loss function [41]. This method assumes that difficult samples should occupy a higher weight in the loss function. We employ the factor $(1-p)\gamma$ as the weight for difficult samples, where p represents the predicted probability output by the network, and γ is the focal factor, set to $\gamma = 2$ in this study. With the inclusion of this weight factor, the weighted sum of the loss function can be expressed as follows:

$$\sum l(p_i, y_i) = (1-p_i)^{\gamma} \cdot l(p_i, y_i) \tag{9}$$

However, directly applying the aforementioned weight factor may alter the overall sum of the original loss values, potentially leading to the issue of vanishing gradients. To adjust the loss distribution without changing the total loss sum, this study introduces a normalization factor Z to maintain the stability of the total loss value. The calculation formula for the normalization factor Z is as follows:

$$Z = \frac{\sum l(p_i, y_i)}{\sum (1-p_i)^{\gamma} \cdot l(p_i, y_i)} \tag{10}$$

Although the concept of difficult sample mining estimation is ideal, its practical application presents a challenge: the process of difficult sample mining is highly dependent on the model's current predictive capability. This means that in the early stages of training, when the model's predictive ability is still unstable, the weight factors obtained from difficult sample mining may misguide the model's training direction. To overcome this challenge, this study introduces a dynamic confidence strategy based on a cosine annealing function, expressed as follows, where the variable *annealing_step* represents the predetermined total number of training steps:

$$l'(p_i, y_i) = \left[\frac{1}{Z}(1-p_i)^{\gamma} + \zeta(t)\left(1 - \frac{1}{Z}(1-p_i)^{\gamma}\right)\right] \cdot l(p_i, y_i) \tag{11}$$

$$\zeta(t) = 0.5 \times (1 + \cos\left(\frac{t}{annealing_step}\pi\right)) \tag{12}$$

Besides calculating loss gradients between predictions and ground truth, the training process uses auxiliary segmentation branches from three GLM Blocks' global–local decoding features, combined with an Auxiliary Segmentation Head (ASH), to introduce multi-scale feature information. This kind of method has been proven effective in segmenting cropland objects [42,43]. The multi-branch constraint enhances segmentation performance. As shown in Figure 8, the ASH uses decoding feature maps from the GLM Blocks to generate auxiliary predictions via convolutional layers and upsampling, including a 3×3 convolution with batch normalization and ReLU activation, followed by a 1×1 convolution and upsampling. This module provides additional gradient information to optimize GLCANet's performance.

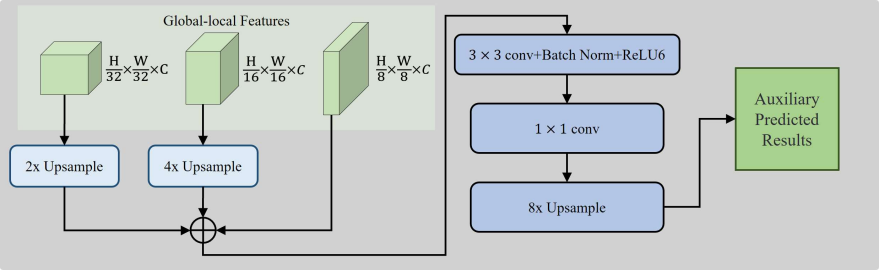


Figure 8. The structure of Auxiliary Segmentation Head (ASH).

For the primary segmentation result, the primary loss function $L_{primary}$ employs a combination of Dice Loss and Soft Cross Entropy Loss. Building upon the Dice Loss and Soft Cross Entropy Loss, the primary loss function further integrates a hard example mining mechanism. The expression for the primary loss function is as follows:

$$L_{basic} = L_{dice} + L_{soft_ce} \tag{13}$$

$$L_{primary} = \left[\frac{1}{Z}(1-p_i)^{\gamma} + \zeta(t)\left(1 - \frac{1}{Z}(1-p_i)^{\gamma}\right)\right] \cdot L_{basic} \tag{14}$$

The auxiliary loss function is calculated using Soft Cross Entropy Loss. Considering that the auxiliary segmentation results are not the primary output of the network, this study introduces an auxiliary factor α to adjust the weight of the auxiliary loss within the total loss. Empirically, α is generally set to 0.4:

$$L_{total} = L_{primary} + 0.4 \times L_{\text{auxiliary}} \tag{15}$$

2.6. Datasets and Experimental Parameter Settings

2.6.1. Datasets

In this section, we will first introduce the datasets utilized in our experiments. We conducted training and testing on three datasets, including two publicly available datasets and one self-labeled dataset. Furthermore, we integrated these three datasets to construct a multi-source imagery dataset named CRMS.

- TriCities Dataset

The original data for the self-labeled dataset primarily comes from high-resolution optical ortho-images acquired by the Gaofen-2, Gaofen-7, and WorldView satellites. The self-labeled dataset, named TriCities (Triplet Cities), was selected based on characteristics from three different regions: Chongzhou and Meishan in Sichuan Province, and Wuzhen in Zhejiang Province. The spatial resolutions of the three satellites are 1 m, 0.8 m, and 0.3 m, respectively, providing precise surface detail information. The TC dataset consists of 1445 small-sized images of 1024 × 1024 pixels, with 1157 images designated for the training set and 288 images reserved for testing and validation.

- GID Dataset

Gaofen Image Dataset [44] is a large-scale, high-spatial-resolution dataset with extensive coverage. The GID dataset comprises 150 scenes of Gaofen-2 imagery with a spatial resolution of 4 m, each scene measuring 6800 × 7200 pixels, covering approximately 560 square kilometers of ground area. By clipping the images into windows of 1024 × 1024 pixels with a 25% overlap rate, we obtained 6720 images for training and 1680 images for testing and validation.

- LoveDA Dataset

The LoveDA [45] (Land cover dataset for Domain Adaptation) dataset contains 5987 high-resolution remote sensing images from the cities of Nanjing, Changzhou, and

Wuhan. Each image measures 1024 × 1024 pixels with a resolution of 0.3 m. Following the data division strategy proposed by the authors for non-transfer learning scenarios, we obtained a total of 2522 images for training, 1669 for validation and 1796 for testing.

- CRMS Dataset

In this study, we selected multiple datasets to support remote sensing cropland information extraction. However, a key challenge is the discrepancy between the training data and real-world data in terms of satellite resolution and sensor characteristics. To overcome this, we analyzed and adjusted our self-labeled and two public datasets, integrating them into a unified cross-resolution multi-source dataset named CRMS (Cross-Resolution Multi-Source Dataset). The CRMS dataset includes six categories: buildings, roads, water bodies, croplands, vegetation, and background. It comprises 15,832 tiles of 1024 × 1024 pixel images, with 10,399 for training, 2653 for validation, and 2780 for testing.

2.6.2. Evaluation Metrics

We have chosen the following accuracy evaluation metrics: overall accuracy (OA), class accuracy (Acc), and mean intersection over union (mIoU).

1. Mean Intersection over Union (mIoU)

The mIoU is a commonly used performance metric for semantic segmentation tasks. It calculates the average of the intersection over union (IoU) for each class. The IoU measures the overlap between the predicted segmentation and the ground truth segmentation for a given class. The mIoU is the mean of these IoU values across all classes. The formula for calculating the IoU for a given class i is as follows:

$$\mathrm{IoU}_i = \frac{TP_i}{TP_i + FP_i + FN_i} \tag{16}$$

The mIoU is then calculated as the mean of the IoU values across all classes:

$$m\mathrm{IoU} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{IoU}_i \tag{17}$$

where

TP_i is the number of True Positives for class i; FP_i is the number of False Positives for class i; FN_i is the number of False Negatives for class i; n is the total number of classes.

2. Overall Accuracy (OA)

This metric measures the overall prediction accuracy of the model. It is defined as the ratio of the number of correctly classified pixels to the total number of pixels. The formula for calculating OA is as follows:

$$\mathrm{OA} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n}(TP_i + FN_i)} \tag{18}$$

where

TP_i is the number of True Positives for class i; FN_i is the number of False Negatives for class; n is the total number of classes.

3. Class Accuracy (Acc)

Class accuracy is a metric that measures the model's prediction accuracy for individual classes. For each class, the class accuracy can be calculated using the following formula:

$$ACC_i = \frac{TP_i}{TP_i + FN_i} \tag{19}$$

3. Experiments and Discussion

To ensure the accuracy and reproducibility of the experimental results, this study utilized a high-performance hardware configuration and a standardized software environment. Specifically, the experimental platform is based on a deep learning workstation equipped with an NVIDIA GeForce RTX 3090 Ti 24GB GPU, an Intel Core i9-10900 CPU, and 32 GB of RAM. The workstation runs Arch Linux, providing a stable and efficient environment for deep learning experiments.

The network model proposed in this study was developed using the PyTorch 2.0 deep learning framework, leveraging the PyTorch Lightning framework to optimize the training process, thereby improving training efficiency and code maintainability. The version of PyTorch used in the experiments is 2.1.2, paired with Python version 3.9.12. To ensure fair comparisons with other benchmark algorithms, official PyTorch implementations of these algorithms were directly adopted for experimentation, avoiding performance biases caused by suboptimal code implementations.

During the training process, the initial learning rate for the proposed algorithm was set to 6×10^{-4}, with AdamW [46] serving as the optimizer. A Poly learning rate decay policy was employed to optimize the training outcomes. The batch size for training each network was uniformly set to 4. The learning rate decay policy follows Equations (3)–(18), where the initial learning rate (*initial_lr*) is set to 0.001, and the iteration count (*iter*) and maximum iteration count (*max_iter*) are calculated based on the current epoch and the number of images in the training set. The hyperparameter *power* is set to 0.9. Each network was trained for 100 epochs to ensure thorough learning.

3.1. Comparative Experiment Results

3.1.1. Experimental Results and Analysis on the CRMS Dataset

We conducted experiments on the CRMS dataset to validate the effectiveness and reliability of the GLCANet network. The experimental results are shown in Table 1. In Table 1, BG represents the background, BD represents buildings, RD represents roads, WT represents water bodies, CL represents croplands, and VG represents vegetation. We conducted a detailed performance analysis of the proposed GLCANet and other novel comparative network methods on the CRMS dataset.

Table 1. Experimental results of different methods on the CRMS dataset.

Methods	Class-Wise Accuracy (Acc%)						OA	mIoU
	BG	BD	RD	WT	CL	VG	(%)	(%)
Unet [7]	65.2	73.2	75.6	87.2	84.5	80.1	79.1	64.2
PSPNet [8]	63.2	74.1	74.2	85.7	81.0	73.1	73.2	57.2
DeepLabv3+ [9]	71.3	83.6	78.6	90.1	88.6	78.1	79.7	66.5
BiSeNet [40]	63.7	69.1	70.9	86.3	80.3	76.2	75.2	56.8
Segformer [47]	63.2	75.8	74.9	85.4	83.2	79.7	76.4	61.9
MANet [48]	73.0	80.5	80.8	84.1	89.3	75.6	79.6	65.9
A2FPN [49]	71.8	83.7	79.5	87.6	89.3	76.3	79.7	65.9
ABCNet [50]	69.4	83.9	**82.9**	92.4	88.6	78.0	79.2	65.3
BANet [51]	72.8	83.1	79.1	87.5	87.6	81.1	80.3	67.1
DC-Swin [52]	70.5	87.2	75.5	87.1	86.9	80.5	81.6	67.2
Unetformer [31]	70.2	84.4	82.4	**94.1**	88.7	77.5	79.4	66.6
GLCANet (ours)	**74.7**	**84.7**	79.2	92.8	**90.3**	**83.5**	**82.6**	**69.2**

As shown in Table 1, GLCANet demonstrates outstanding performance in remote sensing semantic segmentation, achieving top metrics such as 82.6% OA and 69.2% mIoU,

outperforming the other methods. It excels in integrating multi-source data, achieving the highest classification accuracies for croplands (90.3%) and built-up areas (84.7%), showcasing robust capability across categories. While UNetformer performed best for water bodies (94.1%), GLCANet maintained balanced, superior accuracy across all the classes. Compared to BiSeNet, which focuses on lightweight real-time segmentation with limited global modeling capacity, Transformer+CNN architectures like BANet and GLCANet better capture long-range dependencies and local details. Their dual-path design, integrating Transformers for global context and CNNs for multi-scale detail, achieves exceptional performance in complex scenarios and ultra-high-resolution imagery.

In Figure 9, the visualization results of different methods on the multi-source image dataset CRMS are presented, highlighting the test results of the three methods with the highest accuracy from Table 1. The first column of remote sensing images reveals significant differences in features among the images in the multi-source dataset, with varying scales and morphological forms of croplands observed across the images. For instance, in rows (c) and (d) of Figure 9, the cropland patches are relatively small and fragmented, while in row (e), the cropland patches are considerably larger than those in the other rows.

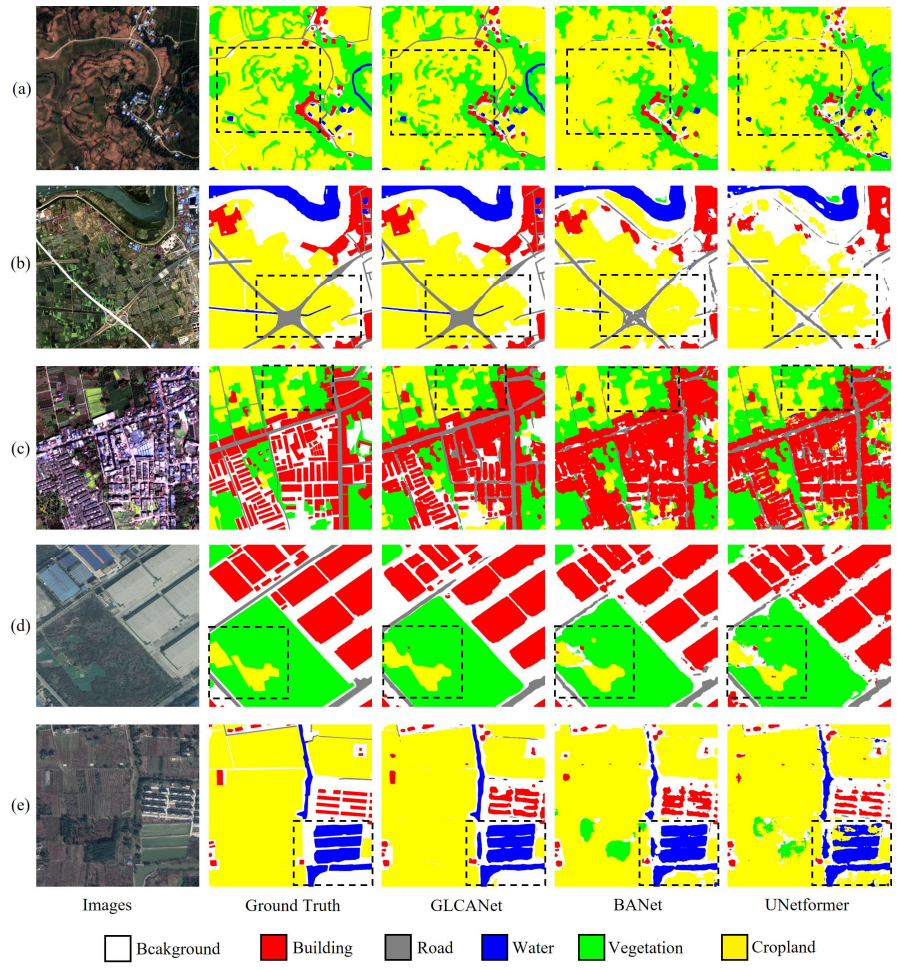


Figure 9. The visualization results of different methods on the multi-source image dataset CRMS.

Compared to the other networks, GLCANet demonstrates superior performance in extracting croplands from multi-source imagery, effectively adapting to the diverse scales and features present in cropland scenes. In rows (a) and (d) of Figure 9, the other networks misclassify vegetation classes that share similar features with croplands; however, GLCANet achieves precise differentiation through its efficient extraction of both the global context and local features. Furthermore, in row (b), which contains both actively planted

croplands and harvested fields, GLCANet accurately extracts all the instances of croplands. Finally, for the croplands presented at varying scales in Figure 9, GLCANet's use of a multi-scale progressive upsampling structure allows for the integration of multi-scale feature maps, enhancing the network's ability to express diverse scale features. As a result, GLCANet achieves commendable extraction results for croplands exhibiting various scale variations.

3.1.2. Experimental Results and Analysis on Three Different Datasets

We conducted a detailed analysis of the performance of the proposed GLCANet network and other comparative methods across different datasets. The comparative methods were categorized into four main types based on their design background and application domains:

- Convolutional neural network (CNN) methods designed for natural scene images, including UNet [7], PSPNet [8], DeepLabV3+ [9], and BiSeNet [40];
- Transformer methods designed for natural scene images, such as Seg-former [47];
- CNN methods specifically designed for remote sensing imagery, including MANet [48], A2FPN [49], and ABCNet [50];
- Transformer methods specifically designed for remote sensing imagery, such as BANet [51], DC-Swin [52], and UNetformer [31].

1. Experimental Results and Analysis on the TriCities Dataset.

Table 2 presents the performance of the selected comparative methods on the self-labeled TriCities dataset, including the accuracy (Acc) for each class, overall accuracy (OA), and mean intersection over union (mIoU) comparison results.

Table 2. Experimental results of different methods on the TriCities dataset.

Methods	Class-Wise Accuracy (Acc%)						OA	mIoU
	BG	BD	RD	WT	CL	VG	(%)	(%)
Unet [7]	70.2	79.7	78.4	90.3	89.9	83.7	83.2	68.7
PSPNet [8]	65.5	78.2	78.4	89.9	85.2	78.9	79.4	61.1
DeepLabv3+ [9]	74.1	81.7	80.8	89.3	90.1	84.7	84.7	70.9
BiSeNet [40]	66.1	73.0	74.8	88.4	83.3	79.0	78.2	60.1
Segformer [47]	67.5	79.5	78.7	89.5	87.6	82.8	81.6	65.6
MANet [48]	67.8	79.0	78.1	88.0	86.5	81.7	80.9	65.1
A2FPN [49]	73.7	81.2	80.6	90.7	89.8	84.4	84.5	70.4
ABCNet [50]	70.5	81.2	79.0	90.5	89.4	83.6	83.3	68.3
BANet [51]	73.9	82.6	**85.0**	91.2	90.0	85.6	85.1	71.5
DC-Swin [52]	73.4	80.9	78.7	89.2	90.8	83.9	84.4	70.6
Unetformer [31]	73.3	83.2	82.7	91.2	90.8	**86.4**	85.3	72.1
GLCANet (ours)	**74.6**	**85.1**	80.2	**92.4**	**91.1**	86.3	**85.7**	**73.6**

On the self-labeled dataset TriCities, compared to the baseline model UNet, GLCANet achieved improvements of 2.5% in OA and 4.9% in mIoU. In terms of class-wise Acc, GLCANet outperformed UNet across all the categories—background, buildings, roads, water bodies, croplands, and vegetation—by 4.4%, 5.4%, 1.8%, 2.1%, 1.2%, and 2.6%, respectively. When compared to the second-best performing model, UNetformer, GLCANet improved OA by 0.4% and mIoU by 1.5%. Specifically, for the cropland category, GLCANet showed improvements of 1.3%, 1.9%, 1.2%, and 0.3% in Acc for background, buildings, water bodies, and croplands, respectively, over UNetformer. The performance of GLCANet on the self-labeled dataset is primarily attributed to its effective utilization of mechanisms for integrating global and local information, which provides superior segmentation accuracy for critical categories like background and cropland. Although the performance gains of GLCANet on the self-labeled dataset are not as pronounced as those on the other datasets, it still maintains a high level of accuracy, demonstrating its robustness and adaptability.

To further demonstrate the significant improvement in the cropland information extraction accuracy of the GLCANet model on TriCities, Figure 10 provides the visual results of several high-performing methods on the GID test set, including PSPNet, Segformer, MANet, UNetformer, and GLCANet. The results clearly show the significant capability of GLCANet in making precise predictions by effectively utilizing both global context information and local features.

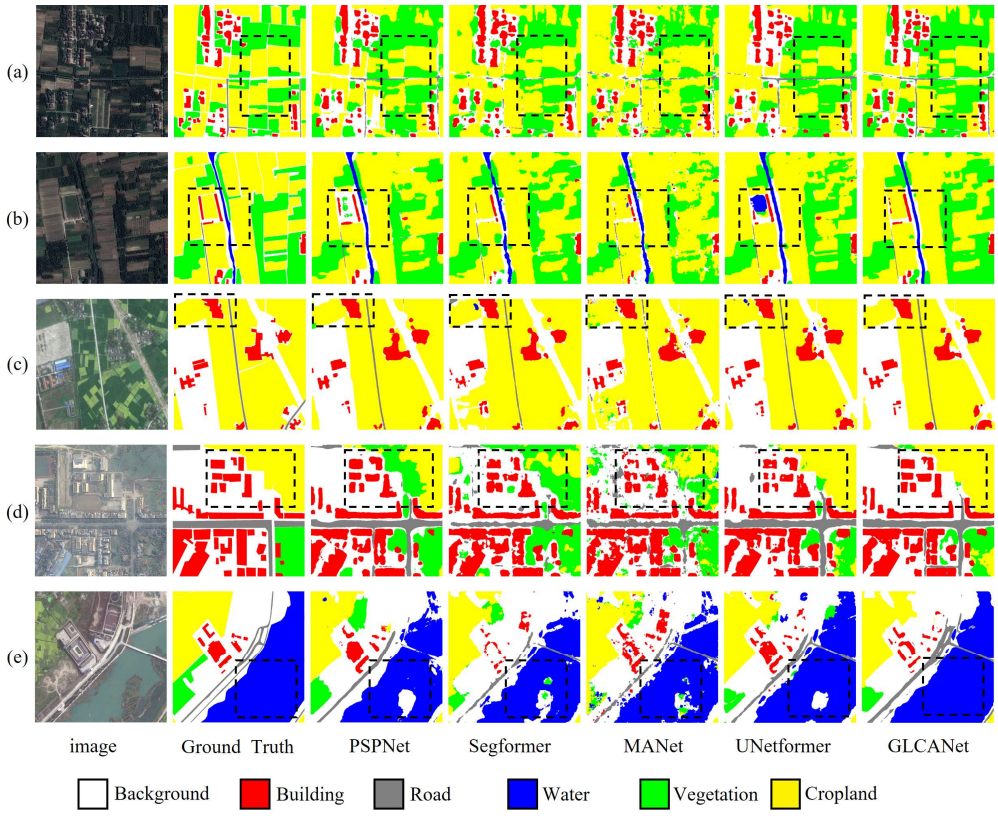


Figure 10. Visualization comparison of different methods on the TriCities dataset (**a**–**e**).

From a visualization perspective, GLCANet is more effective in distinguishing tightly arranged croplands from other land cover types and provides precise extraction boundaries. In rows (a) and (c) of Figure 10, where cropland adjoins vegetation and buildings, GLCANet offers more accurate boundaries and reduces errors. The integration of multi-scale features by GLCANet improves the extraction of smaller land cover types. As shown in row (b), for elongated buildings, GLCANet precisely extracts buildings while maintaining cropland accuracy, unlike PSPNet and UNetformer, which produce erroneous cropland extractions. In scenarios with high inter-class similarities, GLCANet combines global and local features effectively. In the upper right corner of row (d), where croplands and vegetation are similar, and buildings resemble the background, GLCANet achieves precise extraction. In the lower right corner of row (e), the other networks misclassify the riverbed–background distinction, causing voids, whereas GLCANet maintains stable riverbed extraction. These visualizations demonstrate GLCANet's superior performance in handling complex land cover types and maintaining high accuracy in challenging conditions.

2. Experimental Results and Analysis on the GID Dataset.

For the GID dataset, this study also analyzed the performance of GLCANet by comparing it with other trained models. The results of different methods on the GID dataset are summarized in Table 3. The GID dataset includes labels for three distinct forms of croplands: paddy fields (PFs), dry fields (DFs), and irrigated lands (ILs). This characteristic facilitates a deeper analysis of GLCANet's advantages in cropland information

extraction. Quantitatively, on the GID dataset, compared to the baseline model UNet, GLCANet improved OA by 4.9% and mIoU by 11.4%. In terms of class-wise Acc, GLCANet showed significant improvements over UNet for the background, vegetation, and the three cropland categories (paddy fields, dry fields, and irrigated lands), with increases of 6.7%, 3.1%, 17.3%, 6.9%, and 10.4%, respectively. Compared to the second-best performing model UNetformer, GLCANet improved OA by 3.1% and mIoU by 4.3%. Specifically, for the cropland categories, GLCANet improved Acc for paddy fields, dry fields, and irrigated lands by 4.6%, 3.6%, and 2.7%, respectively, over UNetformer.

Table 3. Results of different methods on the GID dataset.

Methods	Class-Wise Accuracy (Acc%)								OA	mIoU
	BG	BD	RD	WT	VG	PF	DF	IL	(%)	(%)
Unet [7]	78.0	87.4	76.0	96.1	87.9	69.9	86.5	82.4	84.5	69.4
PSPNet [8]	73.8	83.8	72.6	94.5	88.2	72.7	84.1	82.0	81.7	64.4
DeepLabv3+ [9]	80.6	85.4	78.6	95.9	85.2	73.7	87.1	84.2	85.4	71.5
BiSeNet [40]	75.9	83.3	71.1	94.1	83.9	65.1	80.1	62.7	80.7	61.9
Segformer [47]	79.2	86.4	70.6	94.2	87.6	81.3	84.4	80.2	84.1	67.3
MANet [48]	78.2	84.6	77.6	95.5	87.7	74.1	88.0	88.8	84.8	70.5
A2FPN [49]	79.8	84.8	72.9	94.5	83.9	73.6	86.2	85.1	84.4	70.3
ABCNet [50]	77.6	85.8	75.9	95.6	88.2	78.6	86.2	87.3	84.4	68.8
BANet [51]	78.4	85.4	77.0	96.8	85.7	56.9	89.7	84.1	84.7	70.5
DC-Swin [52]	80.1	86.1	74.7	95.4	86.0	70.6	83.3	85.8	84.3	67.9
Unetformer [31]	82.2	86.6	78.1	94.6	88.5	82.6	89.8	90.1	86.3	76.5
GLCANet (ours)	**84.7**	**88.7**	**79.5**	**97.7**	**91.0**	**87.2**	**93.4**	**92.8**	**89.4**	**80.8**

Most network methods performed well in the water body category, achieving high accuracies. This could be attributed to the resolution of the GID dataset images being 4 m, making water body features more prominent and easier to distinguish. However, the classification accuracy for paddy fields and dry fields showed considerable variation across different network methods. This variability might be due to the low sample proportions of these categories in the dataset, making it difficult for models to effectively capture their features and increasing the classification difficulty. GLCANet leverages a Transformer structure that effectively integrates global context information and local feature information, achieving high-precision feature extraction in the task of remote sensing cropland information extraction on the GID dataset. The complexity of spatial distribution and subtle differences between categories often found in croplands make it challenging for segmentation tasks. GLCANet's strength lies in its ability to integrate global context and local feature information to capture these differences, thereby enhancing the accuracy and reliability of cropland information extraction. Additionally, GLCANet maintained top-tier performance across other categories, demonstrating its ability to balance high segmentation accuracy while maintaining good performance across different land cover types.

The visualization results in Figure 11 highlight GLCANet's unique advantages in extracting land cover features from remote sensing images. Firstly, GLCANet excels in handling intersecting land cover edges, accurately extracting and preserving information without confusion or loss. This is evident in (a) and (c) of Figure 11, where roads and buildings are closely adjacent. Compared to the other methods, GLCANet demonstrates strong capability in managing complex intersections, avoiding issues like road breaks and building voids through its global context understanding and local detail mining. Secondly, GLCANet shows significant advantages in distinguishing similar land cover categories. For water bodies, irrigated lands, and paddy fields, which share similar features, the other methods often exhibit ambiguous boundaries. In contrast, GLCANet effectively identifies and accurately delineates these boundaries using its multi-scale information extraction mechanism.

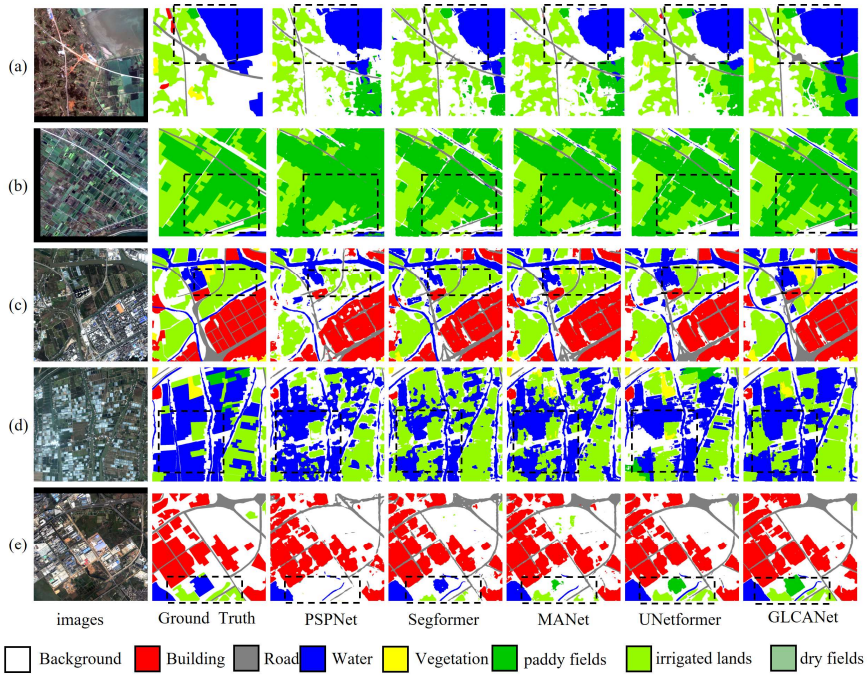


Figure 11. Visualization comparison of different methods on the GID dataset (**a**–**e**).

3. Experimental Results and Analysis on the LoveDA Dataset.

Table 4 shows that Transformer-based methods generally achieve higher mIoU scores than CNN-based methods on the LoveDA dataset. For example, UNet, PSPNet, and MANet have mIoU scores that are 2.8%, 2.3%, and 1.9% lower than DC-Swin, respectively. However, CNN-based methods perform better on the background and barren land categories, possibly due to the complexity of these features and the lack of prominent global information, making it difficult for Transformer-based methods to extract sufficient local details.

Table 4. Results of different methods on the LoveDA dataset.

Methods	Class-Wise IoU (%)							mIoU
	BG	BD	RD	WT	VG	BA	CL	(%)
Unet [7]	43.1	52.7	52.8	73.1	10.3	43.1	59.9	47.8
PSPNet [8]	44.4	52.1	53.5	76.5	9.7	44.1	57.9	48.3
DeepLabv3+ [9]	43.0	50.9	52.0	74.4	10.4	44.2	58.5	47.6
BiSeNet [40]	40.1	51.9	49.8	74.5	11.2	40.8	51.1	45.6
Segformer [47]	42.2	56.4	50.7	78.5	17.2	45.2	53.8	49.1
MANet [48]	43.3	53.7	55.6	76.4	10.4	45.3	58.2	49.1
A2FPN [49]	43.5	55.1	54.3	77.3	16.4	42.3	57.3	49.5
ABCNet [50]	41.5	53.5	53.1	66.6	14.1	40.6	51.7	45.9
BANet [51]	43.7	51.5	51.1	76.9	16.6	44.9	62.5	49.6
DC-Swin [52]	41.3	54.5	56.2	78.1	14.5	47.2	62.4	50.6
Unetformer [31]	44.7	58.8	54.9	79.6	20.1	46.0	62.5	52.4
BAFormer* [53]	44.9	60.6	58.6	80.4	21.3	47.5	61.5	53.5
GLCANet	**48.1**	**61.0**	**59.4**	**81.3**	**23.3**	**48.3**	**64.1**	**55.1**

The LoveDA dataset's test set provides only the corresponding image data without the ground truth labels. Researchers need to submit their prediction results to an official server, which returns only the intersection over union (IoU) scores. Given this setup, the analysis will focus solely on the IoU scores. Specifically, the test accuracy of BAFormer* is sourced from the accuracy reported in its original paper.

The proposed GLCANet method achieved the highest mIoU on the entire LoveDA dataset, surpassing the second-best method by 2.7% in mIoU and by 3.1%, 5.7%, 2.2%, and 0.7% in the IoU scores for background, building, road, and cropland categories, respectively. This demonstrates GLCANet's ability to maintain high segmentation accuracy through its effective integration of local and global information, especially for datasets with complex land cover properties and significant scale variations. GLCANet's superior performance highlights its advantage in extracting cropland information, leveraging its global–local information extraction mechanism to capture both large-scale features and local details.

3.2. Ablation Study

In this section, we conduct ablation experiments on the components of GLCANet to demonstrate their effectiveness and reliability.

3.2.1. Ablation Study of the Decoder of GLCANet

We used BFE-Unet (which means using BFE and Unet's decoder to construct the network) as the baseline and constructed the ablation study schemes. According to the designed experimental schemes, the results of the multi-source imagery dataset CRMS are presented in Table 5.

Table 5. Ablation study of the decoder in GLCANet.

Methods	GLM	MPUS	FRRSH	Class-Wise Accuracy (Acc%)						OA	mIoU
				BG	BD	RD	WT	CL	VG	(%)	(%)
Method 1	×	×	×	69.1	81.2	76.1	90.7	86.1	83.6	77.6	65.7
Method 2	√	×	×	72.6	83.8	76.3	91.2	88.9	84.4	79.2	66.8
Method 3	√	√	×	72.3	84.1	77.6	91.4	89.6	82.7	80.7	67.2
Method 4	√	√	√	**73.6**	**84.7**	**78.2**	**92.8**	**90.3**	**83.5**	**81.4**	**68.9**

While transitioning from Method 1 to Method 2, we observed improvements in accuracy (Acc) for all the categories, as well as in the overall accuracy (OA) and mean intersection over union (mIoU). Specifically, the overall accuracy (OA) improved by 1.6% (from 77.6% to 79.2%), and the mean intersection over union (mIoU) increased by 1.1% (from 65.7% to 66.8%). This change demonstrates the importance of the global–local information mining module in extracting global context and local detail information from complex remote sensing images.

In Method 3, further enhancements were made by adding learnable upsampling to Method 2. The results showed additional improvements in Acc, OA, and mIoU for all the categories. By employing a multi-scale progressive upsampling strategy, the model could more effectively integrate features at different scales, thereby enhancing the segmentation accuracy of objects of varying sizes. This strategy is particularly critical for handling the scale variation issues common in the remote sensing imagery of croplands.

Method 4 combines the global–local information mining module, learnable progressive upsampling, and a feature refinement residual segmentation head. The results indicated that this method achieved the best performance across all the metrics. Notably, there was a significant improvement in the accuracy of cropland and vegetation categories. This suggests that the feature refinement residual segmentation head plays a significant role in improving the model's segmentation accuracy in complex backgrounds and detail-rich areas. By refining feature processing, the model can better capture object boundaries and detailed features, thereby improving the segmentation results at a finer granularity.

3.2.2. Ablation Study of Multi-Constraint Loss

To investigate the impact of multi-constraint loss on network training effectiveness, particularly in handling difficult samples and refining segmentation results, three sets of ablation experiments were designed and conducted.

1. No additional constraints, using only the main loss function (L_{basic}, Equation (13)).
2. Adding difficult sample mining mechanism to the main loss function ($L_{primary}$, Equation (14)).
3. Adding difficult sample mining + ASH to the main loss function (L_{total}, Equation (15)).

Table 6 demonstrates the effectiveness of these designs in enhancing network performance.

Table 6. Ablation experiments of multi-constraint loss.

Methods	Class-Wise Accuracy (Acc%)						OA	mIoU
	BG	BD	RD	WT	CL	VG	(%)	(%)
L_{basic}	74.1	81.4	74.3	90.1	86.3	82.0	78.8	68.1
$L_{primary}$	74.2	83.3	78.8	90.3	87.7	**84.9**	80.1	68.6
L_{total}	**74.7**	**84.7**	**79.2**	**92.8**	**90.3**	83.5	**82.6**	**69.2**

This study found that introducing the difficult sample mining mechanism significantly improved classification accuracy for underrepresented categories like buildings and roads, as well as for the background category with feature confusion. This mechanism effectively enhances the model's recognition of complex features. Additionally, incorporating an auxiliary prediction module further boosted classification accuracy across all the categories and improved the overall accuracy (OA) and mean intersection over union (mIoU). Notably, in the background, building, road, and cropland categories, this approach enhanced detail capture, leading to better segmentation accuracy.

3.3. Discussion

3.3.1. Interpretability Analysis

We conducted a visual analysis of different stages of the GLCANet decoder using SEG-GRAD-CAM [54], with the results presented in Figure 12. SEG-GRAD-CAM generates activation maps for each pixel in the image, allowing for a detailed understanding of the model's basis for classification decisions at the pixel level.

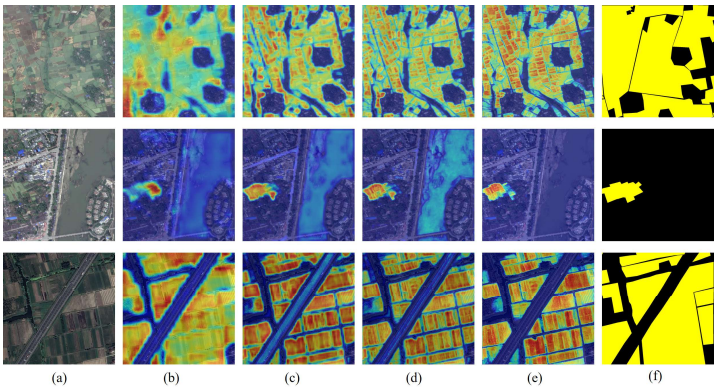


Figure 12. The SEG-GRAD-CAM visualization results for different layers of GLCANet. (**a**) Input image; (**b**) BFE; (**c**) the first layer of the GLM Block; (**d**) the second layer of the GLM Block; (**e**) the third layer of the GLM Block; (**f**) ground truth.

This study shows that the Bilateral Feature Encoder (BFE) can initially identify cropland areas, but the activation boundaries are relatively blurred compared to ground truth labels (Figure 12f), leading to some misactivation in non-cropland regions. While the encoder roughly locates cropland, its understanding of the global context is still coarse, requiring further refinement. As feature maps pass through the GLM Blocks, significant improvements are observed (Figure 12c–e). The model's decision boundaries become clearer, with reduced brightness in non-cropland areas and enhanced definition in cropland boundaries, indicating that the GLM Block effectively improves the model's understanding of both global context and local details, enhancing the accuracy and reliability of

cropland extraction. Additionally, increasing the number of GLM Block layers effectively suppresses misactivation in non-cropland areas, particularly in water regions, where previous misactivations show considerable improvement, reducing the model's misclassification probability. These SEG-GRAD-CAM visualizations highlight the BFE's effectiveness in extracting cropland information and the critical role of the GLM Block in integrating global and local information in remote sensing imagery, thereby enhancing recognition accuracy and decision confidence for cropland extraction.

3.3.2. Analysis of Different Attention Methods in GLCANet

In the GLM module, we introduced the W-MSA attention mechanism to extract global features. To explore the differences in the information mining capabilities of different attention mechanisms, we conducted related experiments, as shown in Table 7.

Table 7. Experiment results of different attention methods.

Attention	Class-Wise Accuracy (Acc%)						OA	mIoU
	BG	BD	RD	WT	CL	VG	(%)	(%)
CBAM [55]	71.3	83.2	77.6	90.3	89.0	82.9	81.2	67.4
Spatial Transformer [56]	72.3	84.3	**81.2**	91.8	89.7	82.7	82.4	68.7
SE Block [57]	73.3	84.1	78.6	92.3	89.4	83.2	82.1	68.4
W-MSA [52]	**74.7**	**84.7**	79.2	**92.8**	**90.3**	**83.5**	**82.6**	**69.2**

The experiment results show that the network using the W-MSA (window-based multi-head self-attention) attention mechanism achieved the best performance in both OA (overall accuracy) and mIoU (mean intersection over union). Compared to CBAM, Spatial Transformer, and SE Block, W-MSA demonstrates significant advantages in handling large-scale image data. First, W-MSA reduces computational complexity by dividing the input feature map into fixed-size windows and performing self-attention within each window, which is especially suitable for resource-constrained environments. Additionally, W-MSA retains the powerful ability of the Transformer model to capture local features and, through window sliding and inter-window connections, effectively maintains the propagation of global information. In contrast, CBAM primarily focuses on channel and spatial features but lacks the benefits of global self-attention mechanisms; Spatial Transformer focuses on spatial transformations, which come with higher computational costs and perform well only in task-specific scenarios; SE Block improves feature selectivity at the channel level but is unable to model spatial features. Therefore, W-MSA excels in computational efficiency and global feature modeling, making it ideal for large-scale vision tasks, especially in the context of visual Transformers (such as Swin Transformer).

The advantage of W-MSA lies in its efficient global feature modeling capability, making it suitable for large-scale data and complex tasks, which improves overall accuracy. On the other hand, Spatial Transformer excels in tasks that require spatial adjustments, particularly in road extraction tasks that involve specific geometric shapes. It can precisely capture local features of roads through adaptive transformations. Therefore, for road extraction tasks, Spatial Transformer provides better accuracy, while W-MSA demonstrates superior performance in overall accuracy.

3.3.3. Analysis of Hyperparameters in GLCANet

In this study, we selected $\alpha = 0.4$ as the optimal weight for the auxiliary loss, as it strikes a good balance between the main and auxiliary tasks. At this value, as shown in Figure 13, we observe a relatively stable or improved overall accuracy (OA) across most datasets, with only a moderate decrease in mean intersection over union (mIoU). This suggests that $\alpha = 0.4$ allows the auxiliary loss to contribute positively without significantly degrading the model's overall performance. Larger values of α tend to overemphasize the auxiliary task, leading to potential overfitting and a loss in performance on the main task. On the other hand, smaller values may not fully leverage the benefits of the auxiliary

loss. Thus, $\alpha = 0.4$ provides an effective compromise, enhancing the model's performance without sacrificing the detailed accuracy captured by mIoU. This finding is supported by the empirical results, where $\alpha = 0.4$ consistently yielded good performance across different datasets, making it a suitable choice for balancing both the overall accuracy and per-class precision.

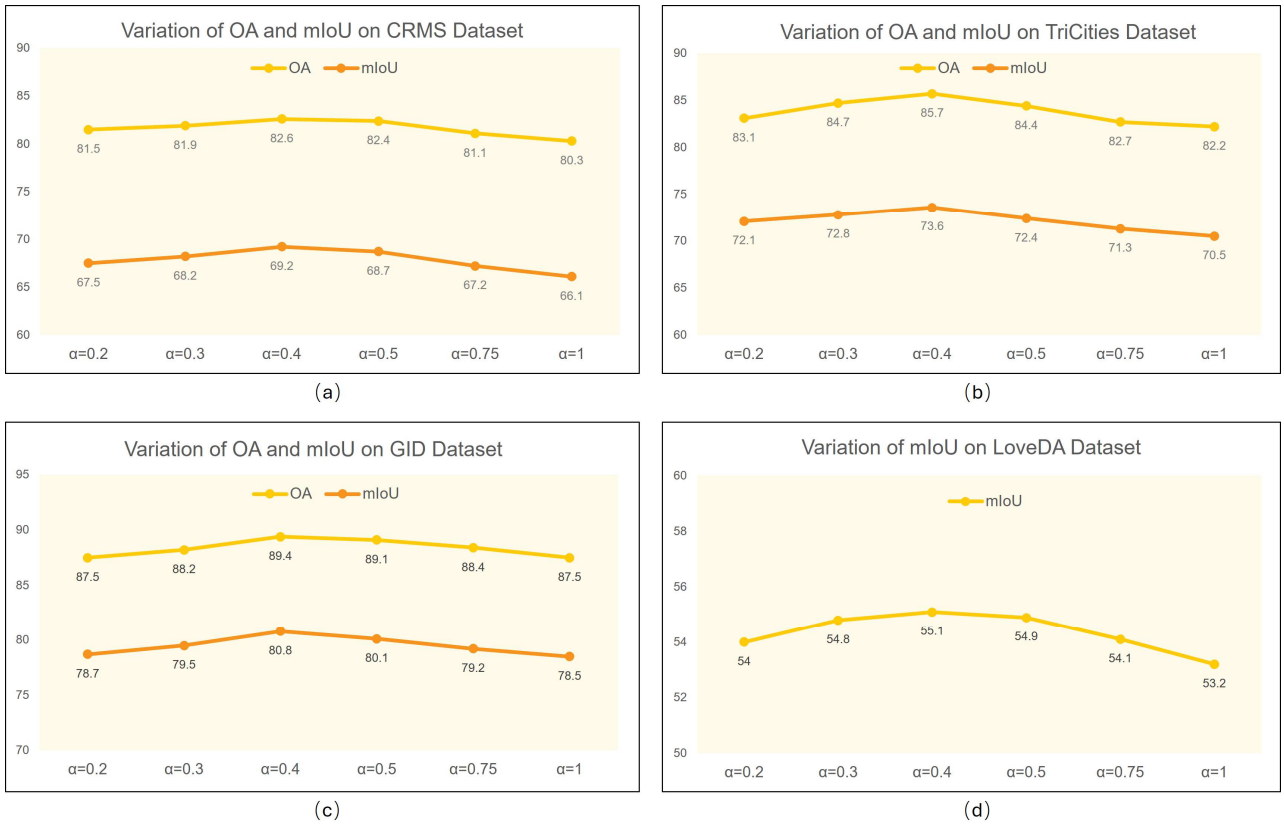


Figure 13. The impact of hyperparameter α on different datasets. (**a**) Accuracy of different hyperparameters on the CRMS dataset; (**b**) Accuracy on the TriCities dataset; (**c**) Accuracy on the GID dataset; (**d**) Accuracy on the LoveDA dataset.

3.3.4. Analysis of the Model's Computational Cost

In this study, to comprehensively evaluate the performance of deep learning methods in the task of cropland information extraction from remote sensing data, this paper not only focuses on the accuracy of the models but also conducts a detailed analysis of their computational efficiency. Specifically, the time complexity of the models is assessed by comparing their parameter counts (Params) and floating-point operations (FLOPs). The relevant results are summarized in Table 8.

As shown in Table 8, convolutional neural network (CNN)-based models typically have lower parameter counts and FLOPs, indicating that these models are more lightweight in design. However, compared to Transformer-based models, CNN methods often achieve lower accuracy in cropland information extraction tasks from remote sensing data. Even when the accuracy of CNN-based models is improved by increasing network depth, parameter counts, and FLOPs, their overall performance often remains inferior to that of Transformer-based models. For instance, the Deeplabv3+ model has parameter counts and FLOPs of 26.68 M and 147.47 G, respectively. Nevertheless, its accuracy in most experimental scenarios is still lower than that of the UNetformer model, which only has 11.72 M parameters and 46.99 G FLOPs.

In the proposed GLCANet model, this study successfully achieves an optimal balance between model performance and computational cost. The GLCANet model has 25.06 M parameters and 76.88 G FLOPs, a configuration that not only ensures the model's lightweight

design but also delivers the best performance in cropland information extraction tasks from remote sensing data. Compared to traditional CNN-based models, GLCANet significantly improves the accuracy of cropland segmentation while consuming more computational resources, demonstrating its strong capability in handling remote sensing data. In summary, through a comparative analysis of the parameter counts and FLOPs of different models, this study concludes that GLCANet effectively enhances the accuracy of cropland information extraction while maintaining relatively low time complexity.

Table 8. Computational complexity of each comparison method.

Methods	Params (M)	FLOPs (G)
Unet	32.52	172.08
PSPNet	24.32	47.73
DeepLabv3+	12.55	37.74
BiSeNet	26.68	147.47
Segformer	27.35	227.24
MANet	35.86	345.27
A2FPN	12.16	91.17
ABCNet	13.67	62.55
BANet	12.73	52.88
DC-Swin	118.39	492.45
Unetformer	11.72	46.99
GLCANet	25.06	76.88

4. The Application of GLCANet in Large-Area Remote Sensing Images

This section primarily discusses the extraction of remote sensing cropland information achieved through the integration of multi-source remote sensing images in Chongzhou City, Sichuan Province, China. The location is shown in Figure 14.

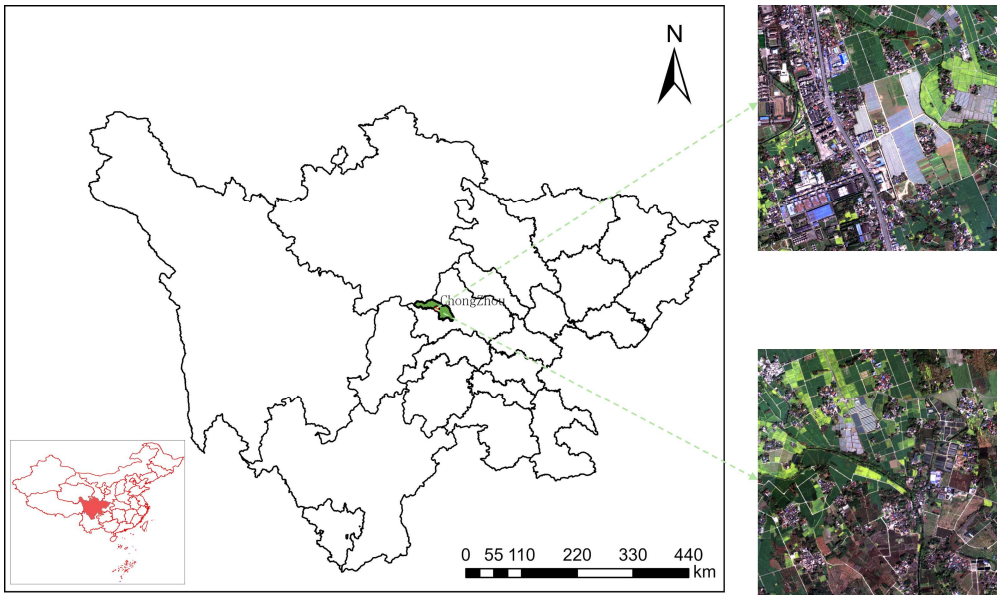


Figure 14. The location of Chongzhou City.

To verify the precision and reliability of our method in large-area cropland extraction, we randomly selected 500 points within the administrative area of Chongzhou and used visual interpretation and Chongzhou's land use data as ground truth for validation. To validate the effectiveness of our proposed method for large-scale applications, we conducted a quantitative analysis of the results using overall accuracy (OA) and Kappa coefficient. Figure 15 shows the full-area remote sensing image of Chongzhou and sampling points.

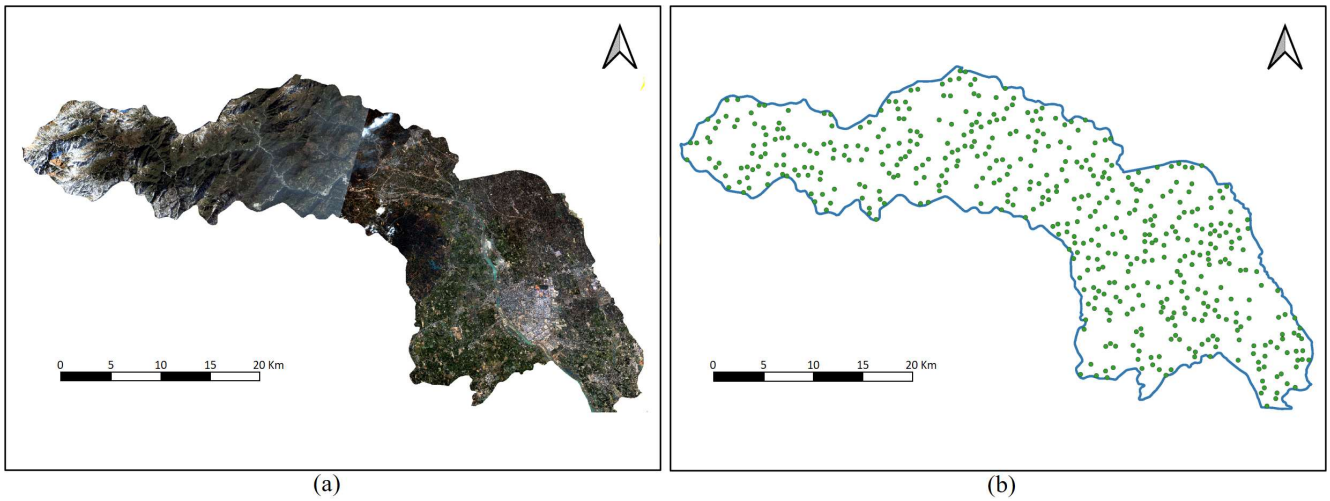


Figure 15. Chongzhou remote sensing imagery and sampling points. (**a**) Full-area remote sensing image of Chongzhou; (**b**) randomly selected sampling points.

This section employs deep learning techniques for the semantic segmentation of three satellite images from Chongzhou, successfully extracting information on buildings, roads, water bodies, cropland, vegetation, and background. Figure 16 presents the results of multi-object information extraction and cropland extraction from the three images.

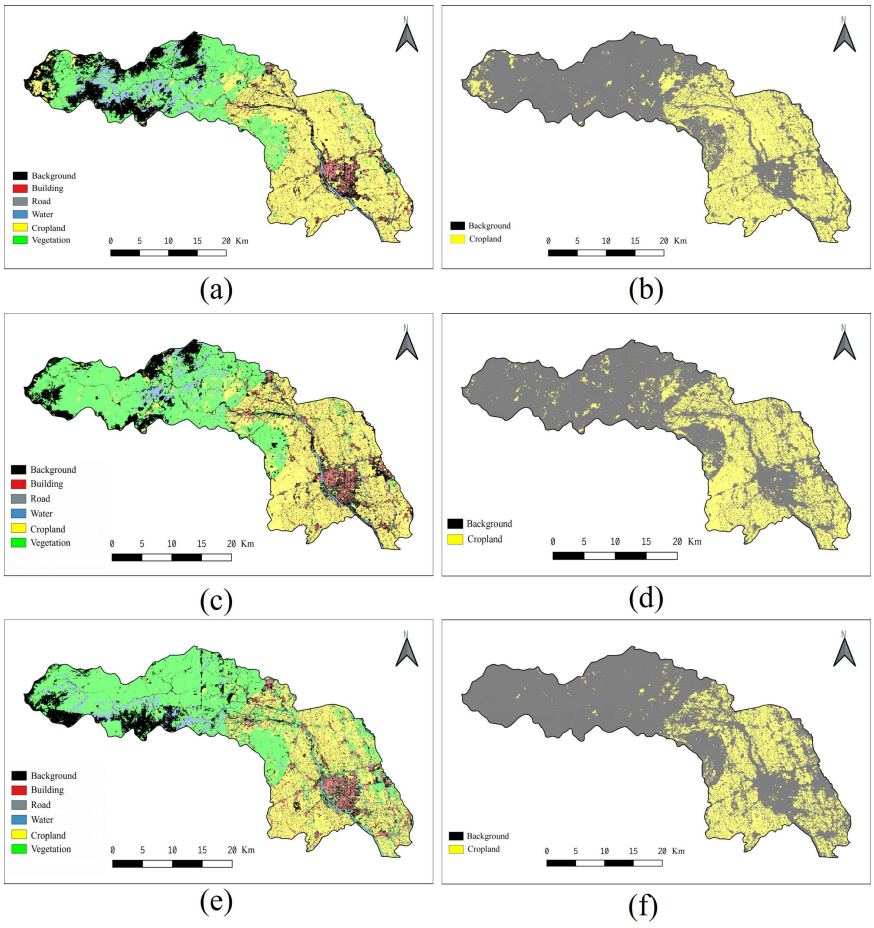


Figure 16. Extraction results for the Chongzhou Area. (**a**,**b**) show the results from the WorldView satellite images from March 2017; (**c**,**d**) display the results from May 2020; (**e**,**f**) depict the results from August 2023 using the GaoFen-1 satellite.

The cropland extraction results indicate that the main land cover type on the western side of Chongzhou is vegetation, while cropland resources are primarily located in the eastern half. More specifically, cropland is predominantly distributed in the outer regions of the eastern half, radiating from the central residential and industrial areas (classified as buildings). This distribution aligns with the topography of Chongzhou, which slopes from northwest to southeast, characterized by higher elevations in the northwest and lower elevations in the southeast.

To validate the accuracy of the cropland data extracted from the four images of Chongzhou, we provide a confusion matrix and accuracy assessment results for the remote sensing image extraction outcomes. Due to external constraints making field verification difficult, visual interpretation methods were employed for accuracy validation, using contemporaneous high-resolution Google images to confirm uncertain land cover types. The specific results are presented in Table 9.

Table 9. Confusion matrix and precision validation for the Chongzhou Region.

March 2017	Class	Cropland (Label)	Non-Cropland (Label)	OA 88.0%
	Cropland (predict)	141	25	
	Non-Cropland (predict)	36	298	Kappa 0.724
	Total	177	323	
May 2020	Class	Cropland (Label)	Non-Cropland (Label)	OA 87.8%
	Cropland (predict)	141	25	
	Non-Cropland (predict)	36	298	Kappa 0.751
	Total	177	323	
August 2023	Class	Cropland (Label)	Non-Cropland (Label)	OA 92.6%
	Cropland (predict)	141	25	
	Non-Cropland (predict)	36	298	Kappa 0.843
	Total	177	323	

In the analysis of the WorldView imagery for Chongzhou, a total of 500 random points were selected for each of the four time periods. For March 2017, there were 158 cropland points, with 130 correctly classified and 28 misclassified, resulting in an overall accuracy of 88.0% and a Kappa coefficient of 0.724, indicating high consistency. In May 2020, among 177 cropland points, 141 were classified correctly, with 36 misclassifications, yielding an overall accuracy of 87.8% and a Kappa coefficient of 0.751, also indicating high consistency. Finally, in August 2023, from 198 cropland points, 171 were correctly classified and 27 misclassified, achieving an overall accuracy of 92.6% and a Kappa coefficient of 0.843, signifying nearly complete consistency.

Figure 17 presents the semantic segmentation results and accuracy validation for two areas of the Chongzhou imagery from the GaoFen-1 satellite after labeling.

In Figure 17, the predicted overall accuracy (OA) is 91.3%, with a mean intersection over union (mIoU) of 77.2%. A closer analysis of the segmentation results reveals that while the region shows commendable extraction accuracy, some errors still persist. For instance, in the lower-left corner of the image, the prediction exhibits a significant void, primarily due to the misclassification of vegetation. This misclassification may be attributed to the similar color characteristics of the surrounding vegetation and the land in the background. Overall, the extracted cropland results for the region are relatively close to the ground truth.

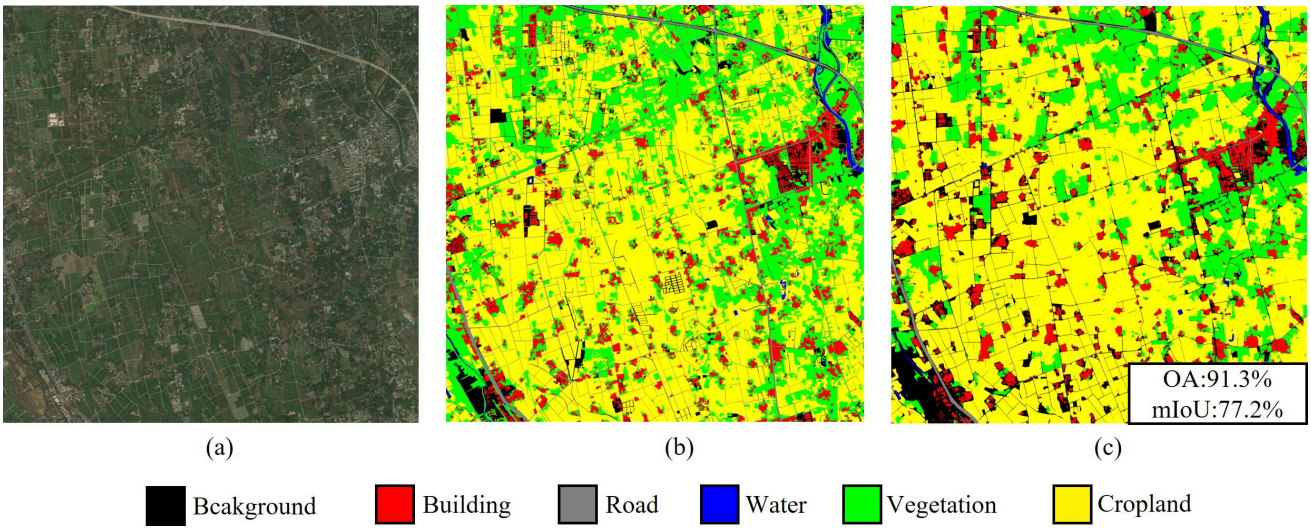


Figure 17. Land cover extraction results for images of Chongzhou City. (**a**) High-resolution images. (**b**) Ground truth. (**c**) Large-area prediction results.

5. Conclusions

This study leverages deep learning semantic segmentation techniques to address the challenges in extracting cropland information from large-scale remote sensing imagery, including the diversity of cropland features, misclassification issues, and low accuracy across various scales. The key innovations are as follows:

Dataset Development: To overcome the limitations of single-source datasets, we constructed a self-annotated dataset called TriCities using multi-spectral images from Gaofen-2, Gaofen-7, and WorldView. We combined TriCities with the public datasets GID and LoveDA to create the multi-source dataset CRMS, which enables effective cropland extraction across different resolutions and scenes with a single deep learning model.

Network Architecture: We proposed GLCANet, a multi-scale semantic segmentation network that integrates global context and local features. GLCANet employs a Bilateral Feature Encoder (BFE) and a Global–Local Information Mining Block (GLM Block) to extract combined features. A multi-scale progressive upsampling structure enhances the model's extraction capabilities. A hard sample mining-based loss function, along with an auxiliary segmentation loss, forms a multi-constraint loss function, significantly improving cropland extraction accuracy. The experimental results show that GLCANet outperforms single-dataset training, with improvements in OA and mIoU on both GID and LoveDA datasets.

Validation of Effectiveness: To validate the method's effectiveness, we focused on Chongzhou in Chengdu, Sichuan Province, using Gaofen-1 imagery and WorldView images. The method achieved over 87.8% precision in cropland extraction and demonstrated feasibility for large-scale monitoring tasks.

However, there is still room for improvement in GLCANet's accuracy. Additionally, the integration of multimodal data is lacking in GLCANet, which will be a focus for our future research.

Author Contributions: Conceptualization, Y.L.; methodology, Z.T. and J.Z.; software, M.Z.; validation, Z.T., J.Z. and Z.N; resources, Z.N., H.H. and Y.L.; writing—original draft preparation, J.Z.; writing—review and editing, Y.L.; visualization, J.Z.; supervision, Y.L.; project administration, Y.L.; funding acquisition, L.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Science and Technology Support Project of Sichuan Province under Grants 2023YFS0366 and 2024YFFK0414, and in part by the Key Projects of Global Change and Response of the Ministry of Science and Technology of China under Grant 2020YFA0608203.

Data Availability Statement: The GID dataset can be acquired at https://x-ytong.github.io/project/GID.html (accessed on 14 January 2024); the LoveDA dataset can be acquired at http://junjuewang.top/ (accessed on 21 January 2024).

Acknowledgments: The author expresses gratitude to the teams responsible for creating and publicly sharing the GID and LoveDA datasets.

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

References

1. Xu, H.; Jiang, L.; Liu, Y. Assessing the Accuracy and Consistency of Cropland Products in the Middle Yangtze Plain. *Land* **2024**, *13*, 301. [CrossRef]
2. Qu, Y.; Zhang, B.; Xu, H.; Qiao, Z.; Liu, L. Interannual Monitoring of Cropland in South China from 1991 to 2020 Based on the Combination of Deep Learning and the LandTrendr Algorithm. *Remote Sens.* **2024**, *16*, 949. [CrossRef]
3. Zhang, L.; Wang, W.; Ma, Q.; Hu, Y.; Ma, H.; Zhao, Y. CCropLand30: High-resolution hybrid cropland maps of China created through the synergy of state-of-the-art remote sensing products and the latest national land survey. *Comput. Electron. Agric.* **2024**, *218*, 108672. [CrossRef]
4. Zhang, Z.; Huang, L.; Tang, B.H.; Le, W.; Wang, M.; Cheng, J.; Wu, Q. MATNet: Multi-attention Transformer network for cropland semantic segmentation in remote sensing images. *Int. J. Digit. Earth* **2024**, *17*, 2392845. [CrossRef]
5. Zhang, W.; Guo, S.; Zhang, P.; Xia, Z.; Zhang, X.; Lin, C.; Tang, P.; Fang, H.; Du, P. A novel knowledge-driven automated solution for high-resolution cropland extraction by cross-scale sample transfer. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 4406816. [CrossRef]
6. Badrinarayanan, V.; Kendall, A.; SegNet, R.C. A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *arXiv* **2015**, arXiv:1511.00561. [CrossRef]
7. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015, Proceedings of the 18th International Conference, Munich, Germany, 5–9 October 2015*; Part III; Springer: Cham, Switzerland, 2015; Volume 18, pp. 234–241.
8. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid Scene Parsing Network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
9. 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; pp. 801–818.
10. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep High-Resolution Representation Learning for Human Pose Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5693–5703.
11. Zhou, L.; Zhang, C.; Wu, M. D-LinkNet: LinkNet with Pretrained Encoder and Dilated Convolution for High Resolution Satellite Imagery Road Extraction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 182–186.
12. Liu, F.; Wang, L. UNet-Based Model for Crack Detection Integrating Visual Explanations. *Constr. Build. Mater.* **2022**, *322*, 126265. [CrossRef]
13. Qiu, W.; Gu, L.; Gao, F.; Jiang, T. Building Extraction from Very High-Resolution Remote Sensing Images Using Refine-UNet. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 6002905. [CrossRef]
14. Jiao, L.; Huo, L.; Hu, C.; Ping, T.; Zheng, Z. Refined UNet V4: End-to-End Patch-Wise Network for Cloud and Shadow Segmentation with Bilateral Grid. *Remote Sens.* **2022**, *14*, 358. [CrossRef]
15. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, pp.5998–6008.
16. Strudel, R.; Garcia, R.; Laptev, I.; Schmid, C. Segmenter: Transformer for Semantic Segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 7262–7272.
17. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-UNet: UNet-Like Pure Transformer for Medical Image Segmentation. In Proceedings of the European Conference on Computer Vision, Online, 23–27 October 2022; pp. 205–218.
18. Zhang, R.; Zhang, Q.; Zhang, G. SDSC-UNet: Dual Skip Connection ViT-Based U-Shaped Model for Building Extraction. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 6005005. [CrossRef]
19. Yan, X.; Tang, H.; Sun, S.; Ma, H.; Kong, D.; Xie, X. After-UNet: Axial Fusion Transformer UNet for Medical Image Segmentation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2022; pp. 3971–3981.
20. Fan, C.M.; Liu, T.J.; Liu, K.H. SUNet: Swin Transformer UNet for Image Denoising. In Proceedings of the 2022 IEEE International Symposium on Circuits and Systems (ISCAS), Austin, TX, USA, 28 May–1 June 2022; pp. 2333–2337.
21. Liu, Y.; Minh Nguyen, D.; Deligiannis, N.; Ding, W.; Munteanu, A. Hourglass-Shapenetwork Based Semantic Segmentation for High Resolution Aerial Imagery. *Remote Sens.* **2017**, *9*, 522. [CrossRef]
22. Zhao, W.; Guo, Z.; Yue, J.; Zhang, X.; Luo, L. On Combining Multiscale Deep Learning Features for the Classification of Hyperspectral Remote Sensing Imagery. *Int. J. Remote Sens.* **2015**, *36*, 3368–3379. [CrossRef]

23. Zhao, W.; Du, S. Learning Multiscale and Deep Representations for Classifying Remotely Sensed Imagery. *ISPRS J. Photogramm. Remote Sens.* **2016**, *113*, 155–165. [CrossRef]
24. Zhang, C.; Pan, X.; Li, H.; Gardiner, A.; Sargent, I.; Hare, J.; Atkinson, P.M. A Hybrid MLP-CNN Classifier for Very Fine Resolution Remotely Sensed Image Classification. *ISPRS J. Photogramm. Remote Sens.* **2018**, *140*, 133–144. [CrossRef]
25. Paisitkriangkrai, S.; Sherrah, J.; Janney, P.; Hengel, A.V.D. Effective Semantic Pixel Labelling with Convolutional Networks and Conditional Random Fields. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Boston, MA, USA, 7–12 June 2015; pp. 36–43.
26. Audebert, N.; Le Saux, B.; Lefevre, S. How Useful Is Region-Based Classification of Remote Sensing Images in a Deep Learning Framework? In Proceedings of the 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 10–15 July 2016; pp. 5091–5094.
27. Paisitkriangkrai, S.; Sherrah, J.; Janney, P.; van den Hengel, A. Semantic Labeling of Aerial and Satellite Imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2016**, *9*, 2868–2881. [CrossRef]
28. Maggiori, E.; Tarabalka, Y.; Charpiat, G.; Alliez, P. Convolutional Neural Networks for Large-Scale Remote-Sensing Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2016**, *55*, 645–657. [CrossRef]
29. Kussul, N.; Lavreniuk, M.; Skakun, S.; Shelestov, A. Deep Learning Classification of Land Cover and Crop Types Using Remote Sensing Data. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 778–782. [CrossRef]
30. Volpi, M.; Tuia, D. Dense Semantic Labeling of Subdecimeter Resolution Images with Convolutional Neural Networks. *IEEE Trans. Geosci. Remote Sens.* **2016**, *55*, 881–893. [CrossRef]
31. Wang, L.; Li, R.; Zhang, C.; Fang, S.; Duan, C.; Meng, X.; Atkinson, P.M. UNetformer: A UNet-Like Transformer for Efficient Semantic Segmentation of Remote Sensing Urban Scene Imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *190*, 196–214. [CrossRef]
32. Scheibenreif, L.; Hanna, J.; Mommert, M.; Borth, D. Self-Supervised Vision Transformers for Land-Cover Segmentation and Classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 1422–1431.
33. Ding, L.; Lin, D.; Lin, S.; Zhang, J.; Cui, X.; Wang, Y.; Tang, H.; Bruzzone, L. Looking Outside the Window: Wide-Context Transformer for the Semantic Segmentation of High-Resolution Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–13. [CrossRef]
34. Yang, S.; Zhang, L.; Liu, S.; Lu, H.; Chen, H. Real-Time Semantic Segmentation via a Densely Aggregated Bilateral Network. *IEEE Trans. Neural Netw. Learn. Syst.* 2023; *in press*. 1–12. [CrossRef]
35. Gao, G.; Xu, G.; Li, J.; Yu, Y.; Lu, H.; Yang, J. FBSNet: A Fast Bilateral Symmetrical Network for Real-Time Semantic Segmentation. *IEEE Trans. Multimed.* **2023**, *25*, 3273–3283. [CrossRef]
36. Gao, Y.; Zhou, M.; Metaxas, D.N. UTNet: A Hybrid Transformer Architecture for Medical Image Segmentation. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2021, Proceedings of the 24th International Conference, Strasbourg, France, 27 September–1 October 2021*; Part III; Springer: Cham, Switzerland, 2021; Volume 24, pp. 61–71.
37. Ma, X.; Ma, M.; Hu, C.; Song, Z.; Zhao, Z.; Feng, T.; Zhang, W. Log-CAN: Local-Global Class-Aware Network for Semantic Segmentation of Remote Sensing Images. In Proceedings of the ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; pp. 1–5.
38. Cui, J.; Liu, J.; Wang, J.; Ni, Y. Global Context Dependencies Aware Network for Efficient Semantic Segmentation of Fine-Resolution Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 2505205. [CrossRef]
39. Dong, X.; Bao, J.; Chen, D.; Zhang, W.; Yu, N.; Yuan, L.; Chen, D.; Guo, B. CSWin Transformer: A General Vision Transformer Backbone with Cross-Shaped Windows. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LO, USA, 18–24 June 2022; pp. 12124–12134.
40. Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; Sang, N. BiSeNet: Bilateral Segmentation Network for Real-Time Semantic Segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 325–341.
41. Zheng, Z.; Zhong, Y.; Wang, J.; Ma, A. Foreground-Aware Relation Network for Geospatial Object Segmentation in High Spatial Resolution Remote Sensing Imagery. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 4096–4105.
42. Poudel, R.P.K.; Bonde, U.; Liwicki, S.; Zach, C. ContextNet: Exploring Context and Detail for Semantic Segmentation in Real-Time. *arXiv* **2018**, arXiv:1805.04554.
43. Poudel, R.P.; Liwicki, S.; Cipolla, R. Fast-SCNN: Fast Semantic Segmentation Network. *arXiv* **2019**, arXiv:1902.04502.
44. Tong, X.-Y.; Xia, G.-S.; Lu, Q.; Shen, H.; Li, S.; You, S.; Zhang, L. Land-Cover Classification with High-Resolution Remote Sensing Images Using Transferable Deep Models. *Remote Sens. Environ.* **2020**, *237*, 111322. [CrossRef]
45. Wang, J.; Zheng, Z.; Ma, A.; Lu, X.; Zhong, Y. LOVEDA: A Remote Sensing Land-Cover Dataset for Domain Adaptive Semantic Segmentation. *arXiv* **2021**, arXiv:2110.08733.
46. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. *arXiv* **2017**, arXiv:1711.05101.
47. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Álvarez, J.M.; Luo, P. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12077–12090.
48. Li, R.; Zheng, S.; Zhang, C.; Duan, C.; Su, J.; Wang, L.; Atkinson, P.M. Multiattention Network for Semantic Segmentation of Fine-Resolution Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5607713. [CrossRef]

49. Li, R.; Wang, L.; Zhang, C.; Duan, C.; Zheng, S. A2-FPN for Semantic Segmentation of Fine-Resolution Remotely Sensed Images. *Int. J. Remote Sens.* **2022**, *43*, 1131–1155. [CrossRef]
50. Li, R.; Zheng, S.; Zhang, C.; Duan, C.; Wang, L.; Atkinson, P.M. ABCNet: Attentive Bilateral Contextual Network for Efficient Semantic Segmentation of Fine-Resolution Remotely Sensed Imagery. *ISPRS J. Photogramm. Remote Sens.* **2021**, *181*, 84–98. [CrossRef]
51. Wang, L.; Li, R.; Wang, D.; Duan, C.; Wang, T.; Meng, X. Transformer Meets Convolution: A Bilateral Awareness Network for Semantic Segmentation of Very Fine Resolution Urban Scene Images. *Remote Sens.* **2021**, *13*, 3065. [CrossRef]
52. Wang, L.; Li, R.; Duan, C.; Zhang, C.; Meng, X.; Fang, S. A Novel Transformer-Based Semantic Segmentation Scheme for Fine-Resolution Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 6506105. [CrossRef]
53. Li, Z.; Wang, Y.; Tian, F.; Zhang, J.; Chen, Y.; Li, K. BAFormer: A Novel Boundary-Aware Compensation UNet-like Transformer for High-Resolution Cropland Extraction. *Remote Sens.* **2024**, *16*, 2526. [CrossRef]
54. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.
55. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
56. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-Resolution Image Synthesis with Latent Diffusion Models. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 10684–10695. [CrossRef]
57. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-Excitation Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2011–2023. [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.