remote sensing

MDPI

Article

GCN-Embedding Swin–Unet for Forest Remote Sensing Image Semantic Segmentation

Pingbo Liu [1,2], Gui Zhang [1,*] and Jianzhong Li [3]

[1] College of Forestry, Central South University of Forestry and Technology, Changsha 410004, China; liupingbo@csuft.edu.cn
[2] College of Computer Science and Mathematics, Central South University of Forestry and Technology, Changsha 410004, China
[3] Changsha Shuobo Electronic Technology Co., Ltd., Changsha 410205, China; lijianzhong@csshuobo.com
* Correspondence: zgui@csuft.edu.cn

Highlights

What are the main findings?

- A hybrid deep learning model based on Swin Transformer and GCN can improve the accuracy of land cover segmentation in forested areas by capturing multi-scale features of forest targets in large-scale remote sensing images, as well as the geometric features and edge details of irregular targets (e.g., tree crowns, rivers). Additionally, the model-fusion approach based on a channel attention mechanism further enhances the segmentation accuracy of objects with channel-wise similarity features (e.g., forest, agricultural land, water bodies).
- The integration of a Zero-DCE module effectively mitigates the issue of low contrast between different land cover categories caused by low illumination in dense forest areas by enhancing image brightness. This enhancement plays a crucial role in facilitating accurate remote sensing image segmentation of forested regions.

What is the implication of the main finding?

- The enhancement of semantic segmentation performance in forest remote sensing imagery holds significant implications for forest resource management applications: in forest fire scenarios, high-precision segmentation outputs can provide critical decision-making support for firefighting operations, particularly in optimizing evacuation routes and deploying firebreaks; furthermore, more accurate segmentation results can offer more reliable canopy coverage for forest biomass estimation, thereby improving the spatial resolution of biomass modeling and laying the foundation for carbon sink monitoring and ecological assessment.

Check for updates

Academic Editor: Zhenwei Shi

Received: 12 November 2025
Revised: 25 December 2025
Accepted: 8 January 2026
Published: 12 January 2026

Correction Statement: This article has been republished with a minor change. The change does not affect the scientific content of the article and further details are available within the backmatter of the website version of this article.

Copyright: © 2026 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.

Abstract

Forest resources are among the most important ecosystems on the earth. The semantic segmentation and accurate positioning of ground objects in forest remote sensing (RS) imagery are crucial to the emergency treatment of forest natural disasters, especially forest fires. Currently, most existing methods for image semantic segmentation are built upon convolutional neural networks (CNNs). Nevertheless, these techniques face difficulties in directly accessing global contextual information and accurately detecting geometric transformations within the image's target regions. This limitation stems from the inherent locality of convolution operations, which are restricted to processing data structured in Euclidean space and confined to square-shaped regions. Inspired by the graph convolution network (GCN) with robust capabilities in processing irregular and complex targets, as

well as Swin Transformers renowned for exceptional global context modeling, we present a hybrid semantic segmentation framework for forest RS imagery termed GSwin–Unet. This framework embeds the GCN model into Swin–Unet architecture to address the issue of low semantic segmentation accuracy of RS imagery in forest scenarios, which is caused by the complex texture features, diverse shapes, and unclear boundaries of land objects. GSwin–Unet features a parallel dual-encoder architecture of GCN and Swin Transformer. First, we integrate the Zero-DCE (Zero-Reference Deep Curve Estimation) algorithm into GSwin–Unet to enhance forest RS image feature representation. Second, a feature aggregation module (FAM) is proposed to bridge the dual encoders by fusing GCN-derived local aggregated features with Swin Transformer-extracted features. Our study demonstrates that, compared with the baseline models TransUnet, Swin–Unet, Unet, and DeepLab V3+, the GSwin–Unet achieves improvements of 7.07%, 5.12%, 8.94%, and 2.69% in the mean Intersection over Union (MIoU) and 3.19%, 1.72%, 4.3%, and 3.69% in the average F1 score (Ave.F1), respectively, on the RGB forest RS dataset. On the NIRGB forest RS dataset, the improvements in MIoU are 5.75%, 3.38%, 6.79%, and 2.44%, and the improvements in Ave.F1 are 4.02%, 2.38%, 4.72%, and 1.67%, respectively. Meanwhile, GSwin–Unet shows excellent adaptability on the selected GID dataset with high forest coverage, where the MIoU and Ave.F1 reach 72.92% and 84.3%, respectively.

Keywords: forest remote sensing; graph convolutional network (GCN); semantic segmentation; feature aggregation

1. Introduction

Forest remote sensing segmentation, as a core technology for precision ecological monitoring, faces multiple natural challenges: dense canopy occlusion blurs tree boundaries, irregular and variable object shapes, shadow regions cause brightness distortion, and spectral confusion occurs between forests, agricultural land, and water bodies. These complexities are particularly prominent in key applications such as fire early warning, resource dynamic tracking, disaster assessment, and smart forestry. For instance, precise segmentation can identify high-risk fire zones to optimize emergency response or support carbon sink assessment and ecological restoration planning through 3D modeling. However, existing general-purpose models (e.g., CNN/ViT) struggle to meet practical demands due to their lack of targeted design for occlusion, small objects, and low brightness, underscoring the necessity of developing specialized architectures (e.g., GSwin–Unet) to enhance the reliability and application value of forest mapping.

In recent years, the rapid advancement of deep learning has catalyzed significant progress in semantic image segmentation. In particular, convolutional neural networks (CNNs) with an encoder–decoder architecture have played a vital role. For instance, the U-shaped network (Unet) [1] employs a decoder to capture spatial dependencies from corresponding encoding stages via skip connections. In this context, Transformers exhibit unique advantages and have emerged as a predominant backbone architectures in semantic image segmentation, particularly in medical imaging applications [2–8]. The self-attention mechanism confers upon the model an enhanced capacity to capture long-range contextual dependencies, thereby demonstrating the distinct advantages in feature extraction and capturing global image features. For example, SegFormer [9] introduces a hierarchical Transformer architecture that facilitates the concurrent extraction of high-resolution shallow features and low-resolution semantically rich features. Similarly, Swin–Unet employs a shifted-window Swin Transformer as its contextual feature-extraction encoder, paired

with a symmetric decoder with a Swin Transformer backbone to progressively upsample feature maps.

Forest RS images typically have rich texture features and complex interlaced scenes, with various categories embedded within each other, lacking fixed geometric shapes and clear boundaries. The particularities of forest ground objects—such as low brightness, small scale, mutual occlusion, and high similarity—pose challenges to semantic segmentation for forest RS images, as illustrated in Figure 1. The low brightness in forest RS images, caused by shadow effects, results in lower contrast of ground objects, often leading to semantic ambiguity.

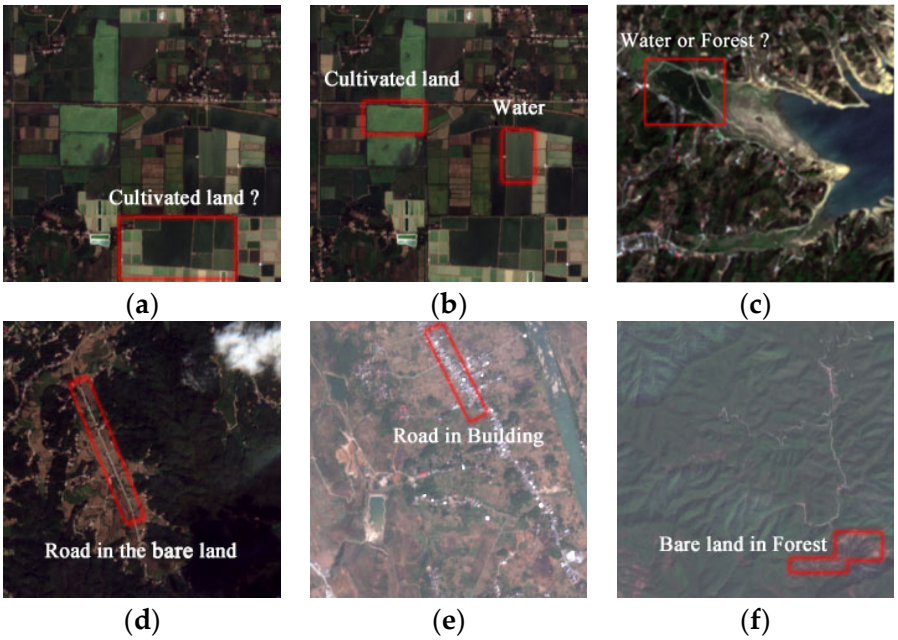

Figure 1. Examples of the characteristics of forest RS images. (**a**,**b**) Water in the farmland has a similar appearance to "Farmland"; (**c**) Forest along the river has a similar appearance to water; (**d**) "Road" and "Bare land" have the same material; (**e**) "Road" is almost invisible in "Building"; and (**f**) Scattered bare land in the forest.

CNNs demonstrate superiority in efficient local perception; however, they struggle to directly model contextual information and global semantic interactions due to the inherent locality of convolution operation [10]. Existing advanced networks [10–15] typically aggregate global information from locally extracted CNN features, rather than encoding global context directly. Consequently, this indirect paradigm hinders the acquisition of clear global scene information in RS images with complex backgrounds [16]. Transformers provide a completely new approach for modeling global relationships [17]. As the first pure Transformer-based encoder for semantic segmentation, SETR [18] abandons convolutional backbones and frames the task as a sequence-to-sequence reconstruction problem. Its strong capability in global context modeling is especially useful for handling long-range dependencies in large-scale images. Mask2Former [19] adopts a mask attention mechanism to dynamically focus on key areas, with strong multitasking adaptability and outstanding performance in complex scene segmentation and video tracking. Swin Transformer [20] achieves a balance between performance and efficiency by introducing a layered architecture and a sliding window mechanism. TransUnet [10] integrates Transformers and CNNs architectures to synergistically capture global semantics and local details, showing superior performance in medical image segmentation tasks including Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) segmentation. Following the Unet architectural paradigm, Cao et al. [4] introduced Swin–Unet, which integrates Swin Transformer's

hierarchical window attention mechanism with skip connections for multi-scale feature fusion. Swin–Unet has shown superior segmentation performance on medical imaging tasks compared to conventional CNN-based models and Transformer–convolution hybrid approaches, particularly in modeling long-range spatial dependencies while maintaining computational efficiency.

However, Transformer-based models rely on global contextual information to make decisions, which may neglect the importance of local details and easily cause small-scale features to be discarded [10]. Additionally, CNNs and Transformers can only be applied to Euclidean spatial data and square regions and cannot capture the geometric variations in the target area of the imagery, resulting in poor capturing ability for the same type of objects with different shapes. In forest remote sensing, ground objects in forests with the same semantic categories may have different sizes and shapes, while different categories may exhibit analogous material and spectral properties, thereby impeding accurate discrimination between them. Therefore, more hierarchical details and effective processing capabilities are needed to fuse global and local contextual features for semantic reasoning.

In recent years, graph convolutional networks (GCNs) have emerged as a powerful paradigm for extracting local features (e.g., edges and textures) from irregular geometric objects by leveraging graph-structured data. As a deep learning framework, GCN performs convolution operations on graph nodes and their neighborhoods, enabling effective feature aggregation for tasks such as classification, clustering, and prediction. In 2019, Chen et al. [21] pioneered a multi-label image-recognition approach that modeled label dependencies through graph structures, demonstrating the potential of GCN in semantic reasoning. Building on this, Jung et al. proposed SGCN [22], which integrated dual graph convolutions—spectral convolution with topology adaptiveness and spatial convolution with weighted node sampling—to capture node semantics in super-pixel graphs. This architectural innovation laid the groundwork for the Vision Graph Neural Network (VIG) [23], which has exhibited remarkable performance in processing complex scenes and irregularly shaped objects. While GCN-based models have achieved significant success in imagery recognition and object detection, their application in RS image segmentation remains underexplored.

In this article, to address the challenges of remote sensing image segmentation caused by the complexity of forest scenarios, we propose the GSwin–Unet architecture. The architecture first introduces a lightweight deep network (DCE-Net) to enhance low-light forest RS images, thereby improving the contrast between different land cover types. Subsequently, it adopts a parallel dual-encoder design: Swin Transformer serves as the principal encoder for efficient global modeling, while graph convolution acts as an auxiliary encoder to capture local features and irregular geometric boundaries. Furthermore, a dedicated feature aggregation module (FAM) is introduced to establish a unidirectional auxiliary-to-main-encoder information propagation mechanism to enable effective feature fusion. This design enhances the discriminability of similar objects across channels by capturing the meaningful content of the image through the channel attention mechanism, while the residual connection synergizes the strengths of both principal and auxiliary modules.

The self-learning semantic segmentation form of spectral, texture, and shape features extracted from forest RS images through intelligent algorithms is used to mine the information of land targets, thus providing a good foundation for accurately obtaining the basic ground features distribution of forest ecological area.

2. Method

2.1. Vision Transformer and Vision GCN

2.1.1. Vision Transformer

Initially proposed for machine translation [24], the standard Transformer architecture comprises three core components: multi-head self-attention (MSA), layer normalization (LN), and multi-layer perceptron (MLP), as schematically shown in Figure 2. MSA is a critical component and enables the establishment of global dependencies between output and input sequences.

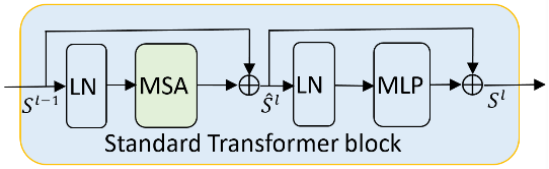

Figure 2. Schematic architecture of the standard Transformer block [17].

In 2020, Dosovitskiy et al. proposed the Vision Transformer model [17], which introduced Transformers into the field of computer vision. This model quickly achieved high performance in object classification, semantic segmentation, and instance segmentation, addressing the limitations of Recurrent Neural Networks (RNNs) in parallel computing and CNNs in effectively learning global feature information. Researchers have since proposed a series of advanced architectures derived from Vision Transformer, including Segmentation Transformer (SETR) [18], Swin Transformer [20], Visual Pyramid Transformer (PVT) [25], TNT [26], and Swin–Unet [4]. These models have demonstrated outstanding performance in the area of vision, particularly in the field of medical imagery segmentation.

SETR [18] pioneered the application of ViT to semantic segmentation. The SETR model uses an encoder with a pure Transformer architecture to replace CNN, altering the existing semantic segmentation model architecture. The multi-scale attention mechanism facilitates robust capture of long-range dependencies between distant structural regions, which is critical for semantic segmentation spanning the images. However, since Transformers did not perform well in capturing low-level details and had computational complexity, many improvements were subsequently made. Wang et al. [25] integrated a pyramid structure into ViT (PVT), mimicking CNN's hierarchical design to generate multi-scale features. By strategically reducing sequence length via patch-embedding layers, their approach dynamically adapts computational complexity while preserving feature granularity. Han et al. [26] proposed the TNT (Transformer-in-Transformer) model, which effectively captures both local and global features through a nested Transformer architecture, thereby generating richer representations. Yuan et al. [27] proposed a recursive token-aggregation mechanism that progressively merges adjacent tokens to capture hierarchical local structures while reducing computational complexity through token count reduction.

Although the models mentioned above improved the feature capture capability, their complexity remains quadratic with respect to image size, a property that directly impairs training and inference speed when processing large-scale datasets. To mitigate this limitation, Liu et al. [20] developed Swin Transformer, a groundbreaking architecture that employs a shifted-window strategy to confine multi-head self-attention (MSA) within local windows while enabling cross-window communication. This approach yields linear computational complexity relative to input size and achieves advanced performance across vision tasks such as image classification, object detection, and semantic segmentation. Swin Transformer replaces standard MSA with Window-based MSA (W-MSA) and Shifted Window-based MSA (SW-MSA) mechanisms, as illustrated in Figure 3. Wang et al. [28] enhanced PVT by introducing a linear-complexity attention mecha-

nism and overlapping patch embedding, which reduced computational complexity from quadratic to linear while maintaining superior performance on fundamental vision tasks. Cao et al. [4] developed specialized U-shaped encoder–decoder architectures based on Swin Transformers, optimized for precise segmentation tasks. Lin et al. [29] implemented a dual-encoder framework using Swin Transformers with dual image scales to extract multi-semantic features.

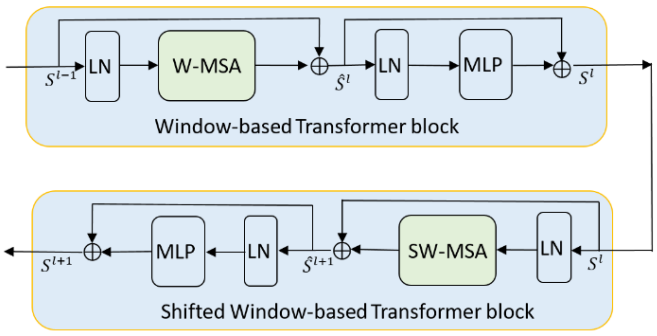

Figure 3. Two consecutive Swin Transformer blocks [20].

In addition, hybrid ViT-based architectures that harness the synergistic strengths of both ViTs and CNNs have demonstrated superior performance across diverse downstream tasks. Chen et al. [10] first combined the global context-modeling capability of ViT with the local-discriminative-feature-preservation mechanism of Unet (TransUnet), and achieved collaborative optimization of long-range dependencies and spatial details in medical imagery segmentation through an encoder–decoder architecture. Zhang et al. [30] introduced a multi-scale cross modal Transformer architecture (TransFuse), which achieves feature-level fusion of camera and LiDAR data through a dynamic-attention mechanism, addressing the challenge of information alignment between heterogeneous sensors in autonomous driving scenarios. Khan et al. [31] proposed MaxViT–Unet, a hybrid Vision Transformer leveraging multi-axis self-attention to enhance medical image segmentation by integrating global context with local inductive biases. The model outperforms pure Transformers on small-object segmentation tasks like Monuseg18, demonstrating the effectiveness of hybrid architectures in medical imaging.

The excellent global modeling capability and hierarchical window attention mechanism of Swin Transformer enable it to effectively capture multi-scale features of forest objects in large-scale remote sensing images. Therefore, this study selected Swin Transformer as the backbone network of the model, whose key module structure is shown in Section 2.1.2.

2.1.2. Swin Transformer Module

As mentioned in Section 2.1, the output s^l of layer l in the standard Transformer block can be described as follows:

$$\hat{s}^l = MSA(LN(s^{l-1}) + s^{l-1}), \quad (1)$$

$$s^l = MLP(LN(\hat{s}^l)) + \hat{s}^l \quad (2)$$

Standard MSA computes global attention across all tokens, resulting in quadratic computational complexity relative to token count. This limitation restricts its applicability in dense prediction tasks and high-resolution image processing. To address this, Swin Transformer introduces Window-based Multi-head Self-Attention (W-MSA) with two partitioning schemes: regular Window partitioning (W-MSA) and Shifted Window partitioning (SW-MSA). Both configurations constrain self-attention to separate local windows of $D \times D$

patches, but neglect cross-window-token interactions. As illustrated in Figure 3, sequential Swin Transformer blocks alternate between W-MSA and SW-MSA operations to establish inter-window connections.

W-MSA is represented as

$$\hat{s}^l = W - MSA\left(LN\left(s^{l-1}\right) + s^{l-1}\right) \tag{3}$$

$$s^l = MLP\left(LN\left(\hat{s}^l\right)\right) + \hat{s}^l \tag{4}$$

SW-MSA is represented as

$$\hat{s}^{l+1} = SW - MSA\left(LN\left(s^l\right) + s^l\right) \tag{5}$$

$$s^{l+1} = MLP\left(LN\left(\hat{s}^{l+1}\right)\right) + \hat{s}^{l+1} \tag{6}$$

Here, s^l and s^{l+1} represent the output feature of the l-th W-Trans block and the subsequent SW-Trans block, respectively.

2.1.3. Vision GCN

Graph convolution is a deep learning method used for graph data capable of handling complex data in irregular geometric spaces and spectrum bands of different quantities and configurations. Spatial-based GCN defines convolution operations on the spatial neighborhood of nodes and aggregates the features of neighboring nodes through convolution operations. Spectral-based GCN, on the other hand, is based on graph signal processing and spectral theory. It treats graph convolution as filtering of signals and defines convolution operations using the eigenvalues and their corresponding eigenvectors of the Laplacian matrix (which contains the topological structure of the nodes). Convolution operation is defined as transforming feature vectors and then inverting them back into the original space. This transformation utilizes the spectral information of feature vectors to capture the global and local structures of nodes in the graph.

In traditional convolutional neural networks, convolution operations are defined in Euclidean space and can only be applied to square regions, unable to capture geometric changes in the target area of the image. Therefore, GCNs are introduced into image processing to adapt to these changes and process complex data in irregular geometric spaces. Kipf et al. [32] first proposed GCN, applying convolutional networks commonly used in the field of deep learning for images to graph data and enabling convolution operations to process complex data in irregular geometric spaces. Subsequently, GCN has achieved good results in pixel-level image processing tasks [33–43]. Chen et al. [42] combined the traditional convolutional neural network SegNet, an attention mechanism, and a graph convolution structure to construct the SE-SegGCN semantic segmentation network, which improved the neural network's receptive field and avoided the loss of local position information. Han et al. proposed ViG [23], a novel backbone network that represents images as graph-structured data, enabling Graph Neural Networks (GNNs) to tackle three classic computer vision tasks: semantic segmentation, image classification, and object detection. Validated through extensive experiments, ViG demonstrates performance on par with or superior to CNN, Transformer, and MLP architectures. GCN has also demonstrated superior performance within remote sensing image processing, with some outstanding models emerging such as AGCN [43], CNN-Enhanced-GCN [44], SCG-Net [45], OBIC-GCN [46], MSSG-Unet [47], KGGCN [48], MSCG-Net [49], KSPGAT [50], etc.

The VIG's outstanding local-feature aggregation capability and its applicability to non-Euclidean graphs enable it to effectively capture the geometric features of irregular objects

(e.g., tree crowns, shadows) and the fine details of forest object edges. This capability can enhance the segmentation accuracy of irregular targets affected by occlusion and illumination variations. Therefore, this study incorporated VIG as an auxiliary encoder into Swin Transformer for model improvement. The key modules of VIG are described in Section 2.1.4.

2.1.4. VIG Module

The VIG block consists of two components: a graph convolution network (Grapher) and a feedforward neural network (FFN) (see Figure 4).

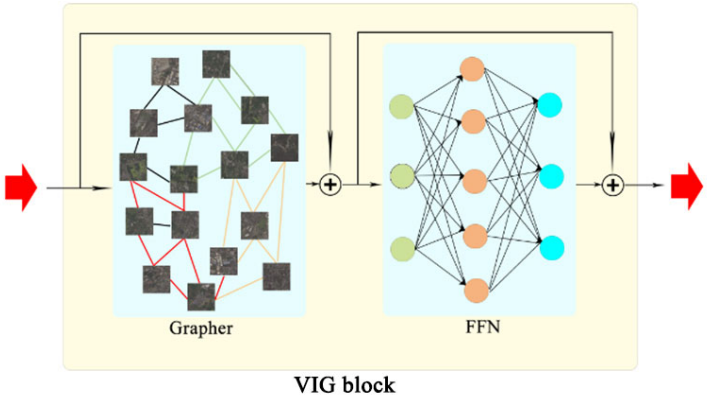

Figure 4. VIG block [23]. The output of the Grapher layer and its input undergo residual connection, followed by layer normalization, and are then fed into the FFN layer.

The Grapher first employs the max-relative graph convolution to aggregate the feature information of the direct neighbors of nodes, as shown in Equation (7). The adjacency matrix of nodes is generated by the self-attention mechanism.

$$x_i'' = \max\left(\left\{x_i - x_j \middle| j \in N(x_i)\right\}\right) \tag{7}$$

Then, the aggregated features are decomposed into h heads via a multi-head structure: $head^1$, $head^2$, …, $head^h$. Each head updates in parallel with its own distinct weights, and they are ultimately concatenated together, as shown in Equation (8).

$$x_i' = x_i'' W_{update} = \left[head^1 W^1_{update}, head^2 W^2_{update}, \cdots, head^h W^h_{update}\right] \tag{8}$$

In the equation, x_i' denotes the updated feature vector, $head^h$ the h-th independent graph convolution head, and W_{update} the update weight matrix. The multi-head concatenation refers to concatenating the output tensors of all heads along the feature dimension (dim = 1) to form a high-dimensional feature vector.

The above graph-level processing can be expressed $GraphConv()$. Before $GraphConv()$, a linear layer projects features into a unified domain to facilitate cross-head fusion, followed by nonlinear activation to prevent layer collapse. The Grapher processing can be expressed in Equation (9).

$$Y = \sigma(GraphConv(XW_{in}))W_{out} + X \tag{9}$$

In the equation, $Y \in R^{N \times D}$, which represents the new node feature matrix after graph convolution, linear transformation, activation function, and residual connection; W_{in} and W_{out} denote the weights matrices of fully connected layers; σ represents the activation functions (such as ReLU and GeLU), and ignores biases; X represents the node feature matrix input to the Grapher module.

The FFN, composed of multi-layer perceptron (MLP) layers, further transforms features while mitigating over-smoothing. Batch normalization is applied after each layer for stable training.

2.2. Improved Hybrid ViT Architecture Integrating GCN

2.2.1. Network Structure

The GSwin–Unet network adopts a dual-encoder structure for feature encoding of input patches. The main encoder is based on a pure Transformer model. It utilizes a layered Swin Transformer with offset windows to extract contextual features. The auxiliary encoder consists of the Grapher and the FFN module of the VIG model. The Grapher module uses graph convolution to aggregate neighbor-node information and update feature representations, while the FFN module uses two linear layers to transform node features. The Grapher module embeds the extracted local features into the main encoder through the feature aggregation module (FAM) to improve the precision of segmenting small-scale and dense objects in forest RS imagery.

Given a forest RS image $X \in R^{H\times W\times 3}$ (with $H \times W$ as spatial resolution and 3 for RGB bands), the Zero-DCE module first adjusts image brightness and enhances ground object contrast by performing convolutional filtering without changing the thematic content. Then, ViG and ViT divide the image into non-overlapping patches, which serve as analogs to the "nodes" of graph and the "tokens" of sequence data. Next, the linear embedding layer flattens the spatial structure of these patches and then projects them into a unified feature space of dimension C through a learnable transformation. After this step, the patch tokens and nodes are fed into the main encoder (composed of Swin Transformer blocks) and auxiliary encoder stacked by ViG blocks. Both the main and auxiliary encoders have three feature-extraction layers, and the output of each layer is, respectively, represented as S_n and G_n (n = 1, 2, 3). The Swin Transformer block comprises two processing steps: shifted W-Trans (SW-Trans) and window-based Transformer. The ViG block consists of two parts: GCN (graph convolution network) and FFN (feedforward neural network). The output resolution of the n-th layer in both the main and auxiliary encoder is $\left(H / \left(2^{n+1}\right)\right) \times \left(W / \left(2^{n+1}\right)\right)$ and the dimensions are $2^{n-1}C$. At each stage, S_n and G_n are fed into the FAM, and the fusion result is returned to the main encoder. The FAM establishes the connection between the main and auxiliary encoders through a channel attention mechanism and deformable convolution.

After the above three coding stages, we obtain the feature $F \epsilon R^{(H/32)\times(W/32)\times 8C}$, which is fed into the bottleneck layer to solve the convergence. In general, achieving convergence is difficult because the Transformer is too deep. Herein the bottleneck layer uses two consecutive Swin Transformer blocks to further process the high-level semantic features from the encoder, while preserving the spatial resolution and channel dimensionality of the bottleneck layer.

The decoder uses a symmetric Transformer-based structure integrating Swin blocks and patch extension layers. Compared to the patch merge layer, the patch extension layer is specifically designed for upsampling to restore the spatial resolution of feature maps and reduce feature dimensions. It reshapes adjacent dimension feature maps into larger feature maps, achieving a resolution of 2× upsampling. In the last patch extension layer, 4× upsampling is performed to recover the spatial dimensions of the feature map to match those of the input image ($W \times H$). In addition, contextual features derived from decoder-based upsampling are concatenated with multi-scale encoder features via skip connections, mitigating spatial-information loss incurred during downsampling. Finally, a linear projection layer is employed to generate pixel-level segmentation predictions from these upscaled features.

The architecture of GSwin–Unet is shown in Figure 5.

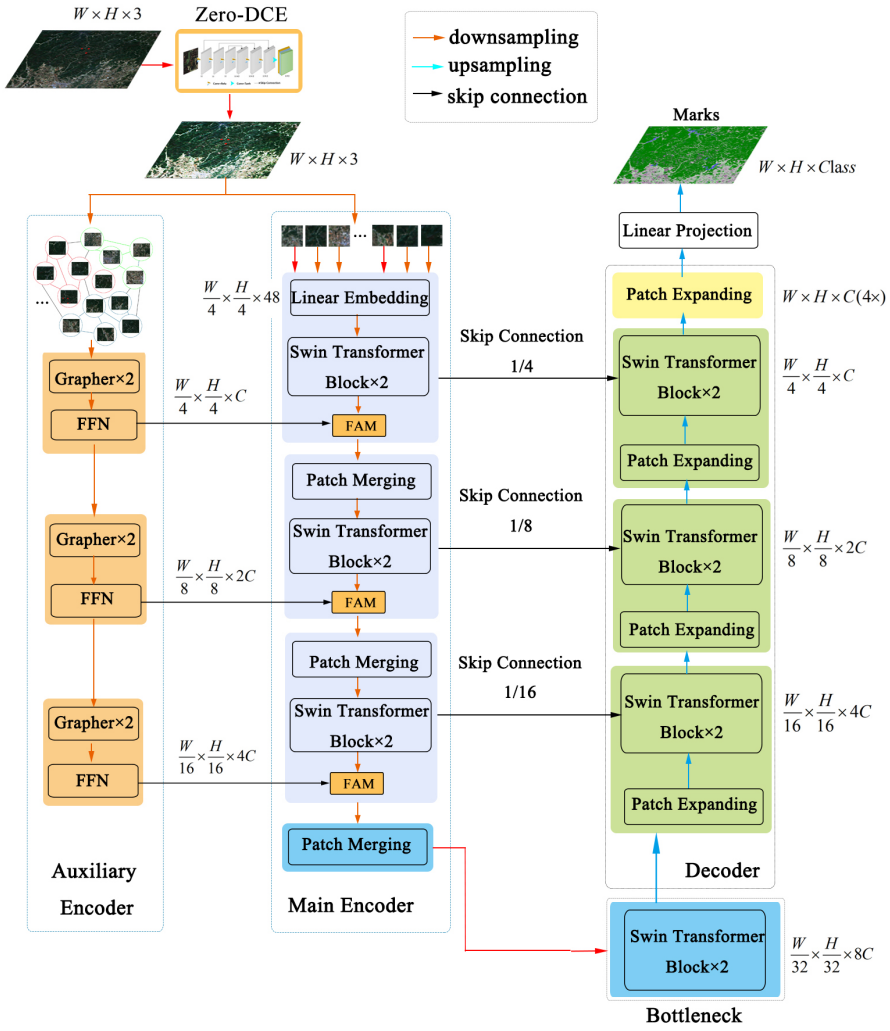

Figure 5. Architecture of GSwin–Unet.

2.2.2. Feature Aggregation Module

In forest RS images, different ground objects are interwoven, and remote sensing features exhibit complex correlations. At the same time, agricultural land, forests, and water have similar distribution patterns in different channels, which can lead to potential problems such as misclassification and unclear boundaries in classification results. In the GSwin–Unet network, the Transformer-based main encoder effectively captures global spatial-feature representation but exhibits limitations in modeling inter-channel relationships. Some methods have shown that encoding channel dimension dependencies can enhance the discriminative capability of features [20,23]. Therefore, we propose the feature aggregation module (FAM), which serves as a bridge between the main and auxiliary encoders. The FAM extracts channel dependencies, emphasizing important and representative channels from the global features of the main encoder through a channel attention mechanism, and integrates into them local features of irregular geometric image regions obtained from the auxiliary encoder.

Through the FAM, more globally discriminative and locally refined features can be extracted to enhance the discrimination accuracy of similar objects and irregular geometric image regions in forest RS imagery. The detailed architecture of FAM is illustrated in Figure 6, where S_n and G_n represent the nth stage outputs of main encoder and auxiliary encoder, respectively. The output G_n from the auxiliary encoder is processed through the

convolutional layer for dimension adjustment, resulting in the derived features G'_n, as shown in Equation (10).

$$G'_n = \varphi(G_n) \tag{10}$$

where φ represents a 1×1 convolutional layer.

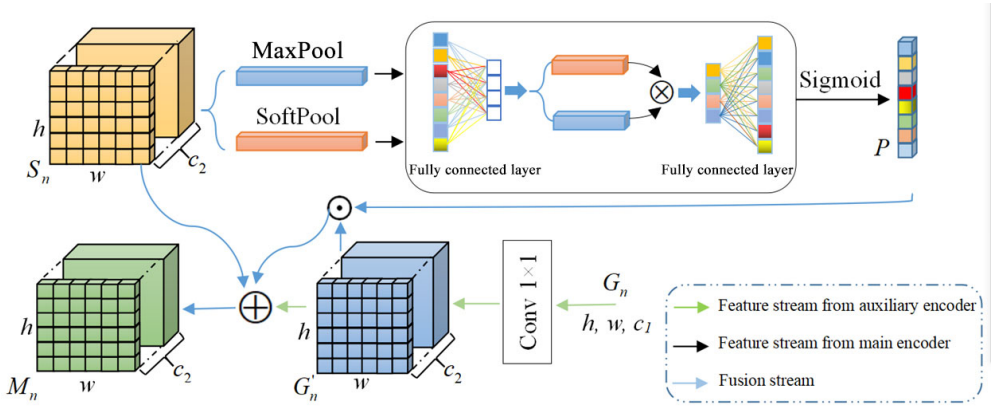

Figure 6. Working mechanism of feature aggregation module. ⊗ stands for matrix multiplication and ⊕ stands for element-level addition.

The output S_n of the main encoder is sent to the pooling layer to compute channel-wise statistical features. Considering each feature map channel as an independent feature detector [51], channel dependencies primarily capture the "meaningful content" of the image [52]. To enhance channel relationship modeling, we implement a combined pooling approach. S_n is simultaneously processed through a max-pooling layer and a soft-pooling layer with exponential weights. The statistical features and global weights output by the pooling layer are input into a shared fully connected layer, whose outputs are P_M and P_S, respectively. The calculation process is shown in Equations (11) and (12).

$$P_M = \sigma(S_1(MaxPool(S'_n))) \tag{11}$$

$$P_S = \sigma(S_1(SoftPool(S'_n))) \tag{12}$$

where σ is the ReLU function, S_1 fully connected layers whose size is reduced to mitigate the high computational load, $MaxPool$ the maximum pooling operation, and $SoftPool$ a weighted pooling operation.

The feature representation of each channel can be optimized by multiplying P_M and P_S, as shown in Equation (13):

$$P = \delta\left(\$_2\left(P_M \bigotimes P_S\right)\right) \tag{13}$$

where δ represents the sigmoid function, $\$_2$ the fully connected layer with increased size, and $\bigotimes$ element-level multiplication.

The FAM multiplies the channel dependency calculation result P as a weight with the output result G'_n of the auxiliary-encoder convolution change to obtain refined features, which are connected to the residual structure to form the output features M_n of FAM, as shown in Equation (14).

$$M_n = G'_n \bigoplus S'_n \bigoplus \left(P \bigotimes G'_n\right) \tag{14}$$

2.2.3. Zero-DCE Module

Zero-DCE [53] is a real-time low-light enhancement strategy that does not require reference data. The model automatically determines whether an image has "reasonable brightness" based on certain properties of the image itself, and adjusts the image through

a brightness-mapping curve fitted by deep learning. The entire calculation process is differentiable and can be easily optimized using gradient descent to achieve extremely fast inference speed. The model comprises three important components: an illumination enhancement curve (LE curve), a parameter estimation network (DCE-Net), and a non-reference loss function.

(1) Light-Enhancement Curve (LE Curve)

An LE curve automatically maps a low-light image to a normal light curve and its parameters for enhancement depend on the brightness of input images. It uses an iterative approach to perform high-order curve enhancement and pixel-wise curve enhancement, and can be expressed as

$$LE_n(x) = LE_{n-1}(x) + A_n(x)LE_{n-1}(x)(1 - LE_{n-1}(x)) \quad (15)$$

where n represents iteration count, A_n represents the curve parameters at each pixel position, and the formula of $LE_n(x)$ is as follows:

$$E(I(x), \alpha) = (I(x) + \alpha I(x)(1 - I(x))) \quad (16)$$

where x represents the pixel coordinate and $LE(I(x), \alpha)$ is an enhanced version of the given input $I(x)$. The trainable curve parameters $\alpha \in [-1, 1]$ adjust the size of the LE curve and control the exposure level. Each input pixel is normalized to the range of [0, 1], and all operations are carried out pixel by pixel (each color channel has a separate curve).

(2) Parameter Estimation Network (DCE-Net)

DCE-Net consists of stacked convolutions and activation functions, with a skip connection layer used to reuse shallow features. It ultimately outputs parameter maps with domain values between $[-1, 1]$ through Tanh activation functions.

The network structure of DCE-Net is shown in Figure 7.

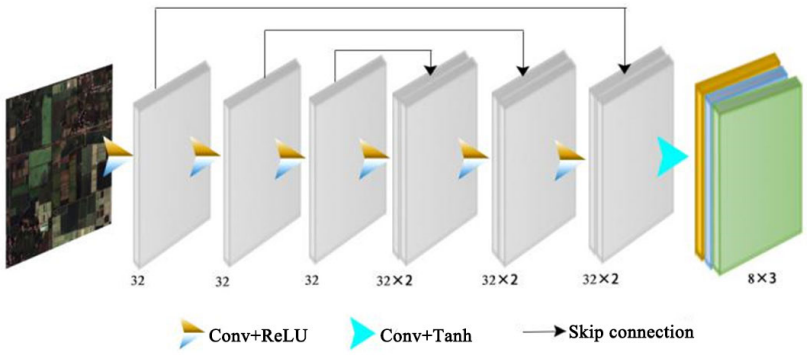

Figure 7. The architecture of DCE-Net [53].

The backbone network of DCE-Net consists of seven convolution layers with symmetric skip connections. The first six convolution layers use ReLU activation functions, while the seventh layer employs a Tanh activation function. All convolutional layers feature a 3×3 kernel size and a stride of 1.

(3) Non-reference Loss Function

The loss function is the key to reference-free learning. Zero-DCE utilizes the optimization process through color consistency loss, spatial consistency loss, illumination smoothing loss, and exposure control loss function. This ensures that the generated images have good brightness and are closer to the original image.

$\boldsymbol{L_{spa}}$—Spatial consistency loss ensures gradient consistency across adjacent regions of the input and its enhanced image (see Equation (17)).

$$L_{spa} = \frac{1}{K}\sum_{i=1}^{K}\sum_{j\epsilon\Omega(i)}\left(\left|Y_i - Y_j\right| - \left|I_i - I_j\right|\right)^2 \tag{17}$$

where K denotes the number of local areas, $\Omega(i)$ the four neighboring areas (left, right, top, bottom) centered around area i, Y the mean value of the corresponding region in the enhanced image, and I the mean intensity value of the same region in the input version.

$\boldsymbol{L_{exp}}$—Exposure control loss optimizes exposure by measuring local intensity deviations from the target level (E = 0.6, following [51]) (see Equation (18)).

$$L_{exp} = \frac{1}{M}\sum_{k=1}^{M}\left|Y_k - E\right| \tag{18}$$

where M denotes the count of non-overlapping 16 × 16 local regions and Y_k the mean intensity value of a local region in the enhanced image.

$\boldsymbol{L_{col}}$—Color consistency loss measures deviations between RGB channels to ensure color balance (see Equation (19)).

$$L_{col} = \sum_{\forall(p,q)\in\varepsilon}\left(J^p - J^q\right)^2,\ \varepsilon\epsilon\{(R,G),(R,B),(G,B)\} \tag{19}$$

where $\boldsymbol{J^p}$ and $\boldsymbol{J^q}$ represent the mean intensity value of the p and q channels in the enhanced image, respectively, and (p,q) denotes a pair of color channels.

L_{tvA}—Illumination smoothness loss ensures smoothness by maintaining pixel monotonicity (see Equation (20)).

$$L_{tvA} = \frac{1}{N}\sum_{n=1}^{N}\sum_{c\epsilon\xi}\left(\left|\nabla_x A_n^c\right| + \left|\nabla_y A_n^c\right|\right)^2, \xi = \{R,G,B\} \tag{20}$$

where N represents the number of pixels and A_n^c represents the illumination enhancement coefficient at the n-th pixel for the c-th channel, ∇x and ∇y are the gradient operators in the x-direction (horizontal) and y-direction (vertical) of the image, respectively, $|\nabla_x A_n^c|$ and $\left|\nabla_y A_n^c\right|$ represent the absolute values of the gradients of the illumination enhancement coefficient A_n^c in the x-direction and y-direction, respectively.

With the L_{spa}, L_{exp}, L_{col}, and L_{tvA} determined, the total variation loss can be calculated, as shown in Equation (21).

$$L_{total} = L_{spa} + L_{exp} + W_{col}L_{col} + W_{tvA}L_{tvA} \tag{21}$$

Here, the weights W_{col} and W_{tvA} serve to balance the contributions of the respective loss components.

Under conditions of insufficient sunlight, forest remote sensing imagery—particularly in dense primary forest regions—exhibits low illumination, resulting in obscured boundaries between ground objects and diminished discriminability. The integration of the Zero-DCE module effectively mitigates these issues by enhancing image brightness, thereby improving inter-class contrast among distinct ground object categories. This enhancement plays a pivotal role in facilitating accurate remote sensing image segmentation.

During the model-training phase, Zero-DCE is applied as a data augmentation module to process input images, generating enhanced images for training. In the model-inference phase, Zero-DCE is applied in real time to process input images, thereby improving segmentation accuracy in low-light scenarios.

2.2.4. Loss Function

The loss function is the core of deep learning, quantifying the discrepancy between the model's predictions and the true values, thereby providing the optimization direction and performance evaluation criteria for model training. In deep learning semantic segmentation, the most commonly used loss functions are the cross-entropy loss, the Dice loss, and the Focal loss.

1. Cross-Entropy Loss (CE Loss) [1]

Cross-entropy loss serves as the foundational loss function for classification tasks, quantifying the divergence between predicted probability distributions and ground-truth labels. Its mathematical formulation is expressed below.

$$L_{CE} = -\sum_{i=1}^{C} y_i \log \hat{y}_i \tag{22}$$

where y_i denotes the binary indicator (0 or 1) for class i, $\hat{y}_i$ represents the predicted probability of class i, and C signifies the total number of classes. This loss function optimizes model parameters by minimizing the Kullback–Leibler divergence between distributions, exhibiting computational efficiency and stable gradient properties. However, its performance degrades under class-imbalanced conditions due to overemphasis on dominant classes. Consequently, *CE* loss is predominantly employed in scenarios with balanced class distributions, such as general-purpose image segmentation tasks.

2. Dice Loss (Dice Coefficient Loss) [54]

Dice loss, derived from the Sørensen–Dice coefficient, is specifically designed for semantic segmentation to address class imbalance challenges. The loss function is defined below.

$$L_{Dice} = 1 - \frac{2|X \cap Y|}{|X| + |Y|} \tag{23}$$

where X and Y correspond to predicted and ground-truth segmentation masks, respectively. By directly optimizing region overlap, Dice loss demonstrates robustness to extreme foreground–background imbalances, making it particularly effective for medical image segmentation and fine-grained land cover classification. Its limitation lies in reduced sensitivity to boundary details compared to pixel-wise losses.

3. Focal Loss (FL) [55]

Focal loss is an advanced modification of cross-entropy loss, engineered to mitigate issues of class imbalance and sample difficulty disparity. The formulation incorporates a modulating factor:

$$L_{FL} = -(1 - \hat{y}_i)^{\gamma} y_i \log \hat{y}_i \tag{24}$$

where γ (typically $\geq$2) acts as a focusing parameter that down-weights well-classified samples (those with high $\hat{y}_i$), thereby directing model attention toward challenging examples. This mechanism significantly enhances performance in tasks with long-tailed distributions, such as object detection and instance segmentation in cluttered environments. Despite its higher computational complexity, Focal loss has become instrumental in improving model generalization across diverse domains.

CE Loss is ideal for baseline evaluations in balanced datasets due to its simplicity and gradient stability. Dice loss excels in medical imaging and remote sensing applications where pixel-level imbalance is prevalent. Focal loss addresses both class imbalance and sample difficulty, proving essential for complex scenes with numerous hard negatives.

The forest remote sensing images are characterized by typical class imbalance and sample difficulty disparity. To address this issue while ensuring training stability, this study

proposes two different loss function strategies for model training, aiming to better explore the model's advantages. See Section 3.2.2 for the specific settings.

3. Results

3.1. Materials

3.1.1. Forest Dataset

The forest dataset is based on the existing publicly available Sentinel-2 remote sensing image data source. Sentinel-2A/B satellites feature a high-resolution multispectral imaging system, utilizing a Multi Spectral Imager (MSI) to support land environmental monitoring. These satellites deliver remote sensing data for diverse land cover types, including vegetation, water bodies, soil, coastal regions, and inland waterways. This is of great significance for improving agricultural and forestry planting, biomass inversion, and monitoring of geological disasters. This study utilized 17 cloud-free or nearly cloud-free remote sensing images from national forest parks, forest lands, and surrounding areas in central and western China (including Hunan, Hubei, Guangxi, Sichuan, and other regions) from January 2024 to February 2025 as the original data source. From these, 15 images were selected to form the forest land classification dataset. For each original image with a size of 10,980 × 10,980 pixels, a representative region of 7840 × 7840 pixels was cropped. Additionally, spectral information was extracted from the 10 m resolution bands (B2, B3, B4, and B8) using different band combinations, generating two types of images: 15 IR-R-G (near-infrared–red–green) images for the IRRG forest RS dataset and 15 R-G-B (red–green–blue) images for the RGB remote sensing dataset. Both datasets were annotated with six semantic segmentation categories: forest, agricultural land, buildings, water bodies, bare land, and roads. Labeling was performed by three trained annotators using GeoLabel. Ambiguous cases (e.g., mixed forest–cropland pixels) were resolved through majority voting with expert consultation. The method of 10% random sample re-annotation ($\kappa > 0.85$) was adopted for quality control.

In the experiments, a scene-level segmentation strategy was adopted: 11 images were selected for model training, and 4 were reserved for testing. All training and testing images were further cropped to a size of 224 × 224 non-overlapping pixels to meet the input requirements of deep learning models. A total of 11,200 images were obtained for training (divided into training and validation sets at a 4:1 ratio), and 4100 images were reserved for testing (1500 images were randomly selected during testing).

3.1.2. GID Dataset

To explore the adaptability of the improved network model, experiments were performed using the publicly available GID (Gaofen Image Dataset). This dataset, specifically designed for land use and land cover classification, comprises 150 Gaofen-2 satellite images covering 60 Chinese cities. Each image has a size of 6908 × 7300 pixels with a ground sampling distance (GSD) of approximately 0.8 m. This collection encompasses a total geographic coverage exceeding 50,000 square kilometers, and the images in the GID dataset have low inter-class separability and high intra-class diversity. GID annotated six types of land use: farmland, forest land, buildings, water bodies, grasslands, and others. A total of 30 images with high forest coverage were randomly selected, with 22 for model training and 8 for validation. These original imagery was cropped to 224 × 224 non-overlapping dimensions. A total of 15,000 images were obtained for training (divided into training and validation sets at a 4:1 ratio), and 4500 images were reserved for testing (2000 images were randomly selected during testing).

3.2. Implementation Details

3.2.1. Training Setting

Our implementation is based on Python 3.8, PyTorch 2.2, and CUDA 11.8. The model was trained using the SGD optimizer with a weight decay of 1×10^{-4} and a momentum item of 0.9. In addition, we adopted the "Poly" decay strategy with an initial learning rate of 0.01. All experiments were conducted on an NVIDIA GeForce RTX 3080 16 GB GPU with 16 GB memory, using a batch size of 8 and a maximum of 110 training epochs. Due to computational constraints, all quantitative results are reported from single training runs. The reported performance gains may exhibit sensitivity to random initialization and data split configurations, potentially leading to variability in replication outcomes.

3.2.2. Loss Function Setting

The Forest and GID datasets demonstrate significant class imbalance, where certain categories are underrepresented, leading the model to disproportionately focus on the majority classes during training. To mitigate this issue, we propose a dual-loss supervision strategy: the first loss function (L1) (following [56]) jointly combines Dice loss and cross-entropy (Equation (25)), while the second loss function (L2) integrates Dice loss with Focal loss (Equation (26)). This approach aims to improve the model's performance on minority classes by addressing the imbalance in data distribution.

To prevent the learning difficulty caused by the unstable gradient of Dice loss during the early stage of training, and to more effectively improve the model's performance in segmentation tasks, especially in cases where the target region is small or unevenly distributed, the composite function L1 is set as shown in Equation (25).

$$\mathrm{L1} = 0.4L_{CE} + 0.6L_{Dice} \tag{25}$$

To reduce the weights of easy-to-classify samples in the case of sample imbalance, focus on hard samples and improve the detection and segmentation capabilities for small or rare objects, while also paying attention to the overlap degree between predictions and real objects, a combined function L2 is set as shown in Equation (26).

$$\mathrm{L2} = 0.6L_{Dice} + 0.4L_{Focal} \tag{26}$$

3.2.3. Evaluation Index

To comprehensively assess the model's performance, we employed average F1 score (Ave.F1) and the mean Intersection over Union (MIoU). Both metrics are derived from the confusion matrix and contain four items: true negative (TN), false negative (FN), true positive (TP), and false positive (FP). The F1 score of each category was calculated as follows:

$$\mathrm{F1} = 2 \times \frac{precision \times recall}{precision + recall} \tag{27}$$

where precision is TP/(TP + FP) and recall is TP/(TP + FN).

For each category, IoU metric is formally defined as the ratio between the intersection area and union area of the predicted segmentation and ground truth, as mathematically expressed in Equation (28).

$$\mathrm{IoU} = \frac{TP}{TP + FP + FN} \tag{28}$$

Furthermore, Ave.F1 score and the MIoU are the average of F1 and IoU from all categories, respectively.

The greater the IoU and F1 indices, the greater the overlap between the predicted area and the true annotated area, and the better the positioning or segmentation accuracy of

the model. The greater the F1 score, the stronger the overall performance of the model in segmentation tasks, especially in scenarios where data is imbalanced and there is a need to balance the risks of false negatives and false positives.

3.2.4. Network Structure Parameter Settings

In the domain of computer vision, commonly used Transformers typically have isotropic structures (such as ViT and ResMLP), where the size of the model block's features remains unchanged from start to finish. The feature map of the pyramid structure model gradually becomes smaller as the network deepens, and has been proven to be an effective performance improvement method by many visual Transformers such as Swin [20], PVT v2 [28], etc. In the experiments, both the main and auxiliary encoders in the GSwin–Unet model adopted a pyramid structure, and the model's network architecture parameters are summarized in Table 1. The hidden layer dimension C was configured to 96, and the window size of the Swin Transformer blocks was configured to 7×7.

Table 1. Network structure parameter settings.

Stage	Pyramid ViG-Tiny	Output Size	Swin Transformer	Output Size
			Patch Partition	$\frac{H}{4} \times \frac{W}{4} \times 48$
Stem	Conv × C	$\frac{H}{4} \times \frac{W}{4} \times C$	Linear Embedding	$\frac{H}{4} \times \frac{W}{4} \times 48$
Stage 1	$\begin{bmatrix} D=48 \\ E=4 \\ K=9 \end{bmatrix} \times 2$ *	$\frac{H}{4} \times \frac{W}{4} \times C$	Swin Transformer Block × 2	$\frac{H}{4} \times \frac{W}{4} \times C$
Downsample	Conv × 2C	$\frac{H}{8} \times \frac{W}{8} \times 2C$	Patch Merging	$\frac{H}{8} \times \frac{W}{8} \times 2C$
Stage 2	$\begin{bmatrix} D=96 \\ E=4 \\ K=9 \end{bmatrix} \times 2$	$\frac{H}{8} \times \frac{W}{8} \times 2C$	Swin Transformer Block × 2	$\frac{H}{8} \times \frac{W}{8} \times 2C$
Downsample	Conv × 4C	$\frac{H}{16} \times \frac{W}{16} \times 4C$	Patch Merging	$\frac{H}{16} \times \frac{W}{16} \times 4C$
Stage 3	$\begin{bmatrix} D=240 \\ E=4 \\ K=9 \end{bmatrix} \times 6$	$\frac{H}{16} \times \frac{W}{16} \times 4C$	Swin Transformer Block × 2	$\frac{H}{16} \times \frac{W}{16} \times 4C$

* D: Feature dimension (size); E: Ratio of hidden dimensions in FFN; K: Neighbors in GCN (receptive field); $H \times W$: Input image size.

3.3. Ablation Study

To assess the proposed network architecture's performance, FAMs, Zero-DCE modules, and loss function, we applied Swin–Unet as the baseline network to perform ablation experiments on the RGB forest dataset. In GSwin–Unet, the main encoder employed Swin, complemented by a Pyramid VIG-Tiny auxiliary encoder with a hidden size of 48, 9×9 receptive fields, and a {2, 2, 6} layer configuration.

3.4. Implementation Results

3.4.1. Effect of Dual-Encoder Structure

The results of the dual-encoder structure are presented in the second row of Table 2. The results demonstrate that the integration of GCN enhances the segmentation performance of Swin–Unet, with an increase of 5.48% in MIoU. By performing element-wise summation of the main- and auxiliary-encoder features at each encoding stage (following [57]), the semantic segmentation model achieves significant improvements in the segmentation accuracy of agricultural land, buildings, and roads (5.4%, 7.47%, and 7.5%, respectively), as well as a moderate improvement in bare land (2.82%). However, the model shows limited

effectiveness for forest and water bodies (0% and −0.31%, respectively). The study demonstrates that the dual-encoder structure based on pixel-wise summation can aggregate more information through hierarchical cascading, which is beneficial for semantic prediction, but it still has limitations for specific land cover types.

Table 2. Ablation experiment of the proposed modules using the RGB forest dataset.

Model Name	IoU (%)						Evaluation Index	
	Forest	Agricultural Land	Water	Building	Bare Land	Road	MIoU (%)	Ave.F1 (%)
Swin–Unet	67.23	70.63	67.26	66.67	62.83	62.50	66.19	79.62
Swin–Unet + GCN + L1	67.23	76.03	66.95	74.14	75.65	70.00	71.67	80.43
Swin–Unet + GCN + FAM + L1	70.09	76.27	72.07	75.63	74.79	74.53	73.90	84.80
Swin–Unet + GCN + FAM + ZD + L1	70.34	77.59	72.07	75.63	76.07	74.77	74.41	85.30
Swin–Unet + GCN + FAM + ZD + L2	73.04	78.63	68.64	76.47	78.26	75.70	75.13	85.96

3.4.2. Effect of Feature Aggregation Module

Table 2 also illustrates that the segmentation results increase by 2.23% for MIoU and 4.37% for Ave.F1 when the FAM was incorporated into the GSwin–Unet framework. In particular, the "Water" category exhibits the most significant segmentation accuracy improvement (+5.12% IoU), followed by "Road" (+4.53%) and "Forest" (+2.56%). Visual segmentation comparisons in Figure 8 further demonstrate these enhancements. In row 1, segmentation errors induced by skylight interference and illumination variations in forested areas are effectively mitigated through RAM implementation. In rows 2–3, the road adjacent to the buildings, which is very similar to the "Building" category, and the water surrounding agricultural land, which is very similar to the "Agricultural land", are still distinguished after using FAM. This demonstrates that FAM-enhanced local-feature embedding improves segmentation accuracy for objects with high similarity.

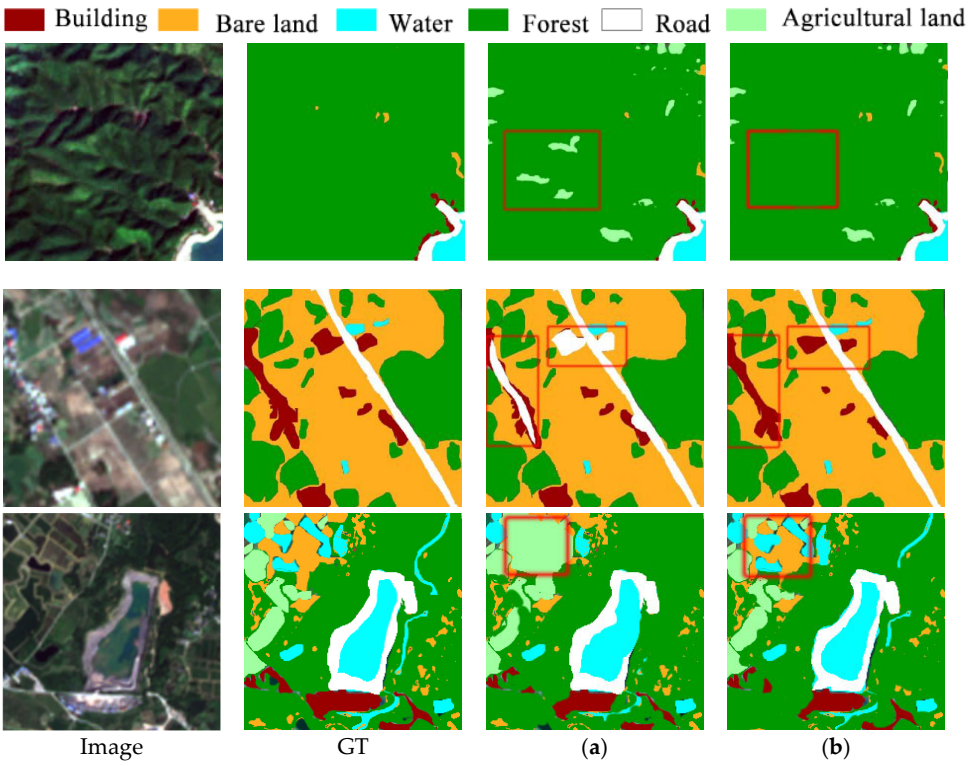

Figure 8. Segmentation performance comparison: (**a**) Swin–Unet + GCN + L1. (**b**) Swin–Unet + GCN + FAM + L1.

3.4.3. Effect of Zero-DCE (ZD) Module

As shown in the fourth row of Table 2, after using ZD, the model achieves improvements of 0.51% in MIoU and 0.5% in Ave.F1. The model shows relatively large improvements in the "Bare land" and "Agricultural land" categories (1.28% and 1.32%, respectively), while the improvements are smaller for "Forest" and "Road" (0.25% and 0.24%, respectively). As illustrated in the first row of Figure 9, the darkness of the image blurs the texture features of the forest, causing the model to fail in judging the boundary of bare land within the forest and mistakenly recognize it as forest. After incorporating the ZD module, the model can accurately identify the bare land boundary. In the second row, the visual similarity between the "Forest" and "Water" categories—where water bodies are embedded within forested regions—poses inter-class discrimination challenges for the model. Before using ZD, the model misclassifies a large area of water bodies as forest. However, after incorporating ZD, the model almost completely extracts the water bodies from the forest. Figure 9 demonstrates that the integration of ZD effectively enhances the discriminability between objects with blurred boundaries by improving contrast. Although the accuracy improvement in this test set is not very significant, the performance of the Zero-DCE module in low-light images is evident.

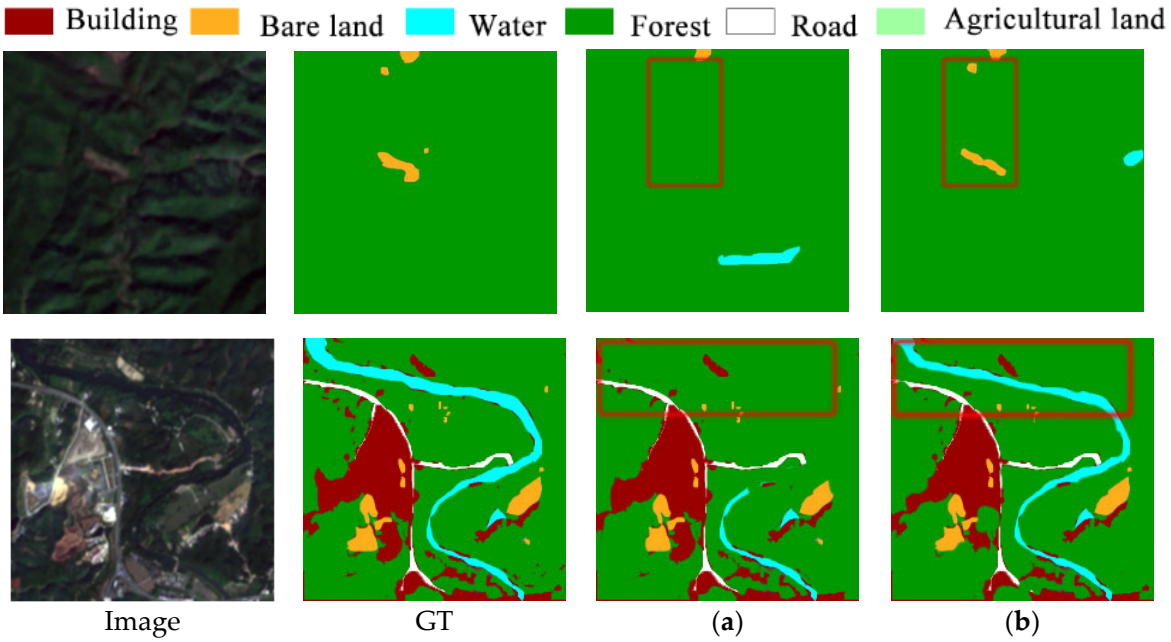

Figure 9. Zero-DCE (ZD) module efficacy: (**a**) Swin–Unet + GCN + FAM + L1. (**b**) Unet + GCN + FAM + ZD + L1.

3.4.4. Effect of Loss Functions

Ablation studies on loss function efficacy were conducted using the RGB forest dataset, with Figure 10 quantitatively presenting the results. Compared to L1 loss (CE + Dice) (see the fourth row in Table 2), L2 loss (Dice + Focal) (see the fifth row in Table 2) improves the segmentation performance of "Forest" and "Bare land", but slightly decreases that of "Water", leading to an overall MIoU improvement of 0.72%. The results indicate that L1 and L2 losses perform comparably in complex environments with class imbalance, with L2 being slightly superior. This advantage can be further investigated in subsequent studies.

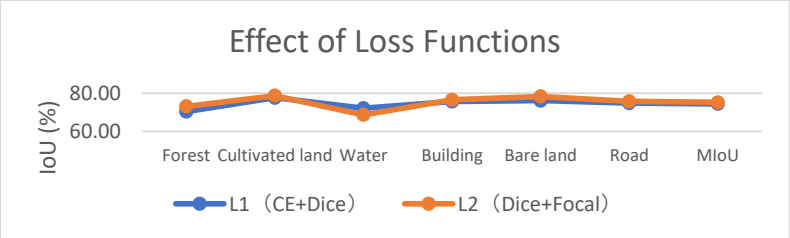

Figure 10. Segmentation performance comparison across Loss-Functions.

3.4.5. Comparison with Other Methods

We compared the proposed GSwin–Unet against several established segmentation methods: TransUnet [10], Swin–Unet [4], Unet [1], and DeepLab V3+ [58]. The first two methods leverage Transformer architecture, while the latter two employ traditional CNN. Specifically, Swin–Unet implements a Unet framework constructed with Swin Transformer blocks, and TransUnet employs a serial hybrid architecture integrating Transformer and CNN modules, whereas our GSwin–Unet adopts a parallel mode of GCN and Swin Transformer.

Unet, DeepLab V3+, and TransUnet employ ResNet50 (non-pretrained) as their backbone, Swin–Unet and GSwin–Unet utilize Swin-Tiny (non-pretrained) for their main encoders, while GSwin–Unet use Pyramid ViG-Tiny as the auxiliary encoder.

Results on RGB Forest Dataset

We validated the effectiveness of the proposed GSwin–Unet on two forest datasets with different bands combinations. The numerical results of each semantic segmentation method on RGB forest dataset are listed in Table 3. Our GSwin–Unet achieves 75.13% in MIoU and 83.92% in Ave.F1, outperforming other methods. DeepLab V3+ with an atrous spatial pyramid pooling module captures global context through expanding receptive field, and experimental data demonstrate that DeepLab V3+ with dilated convolution is inferior to GSwin–Unet in global context modeling. Unet achieves slightly better segmentation results than DeepLab V3+ by continuously integrating low-level spatial features via skip connections. GSwin–Unet surpasses Unet with 5.12% MIoU and 1.72% Ave.F1 improvements. Compared to DeepLab V3+ and Unet, Swin–Unet delivers unsatisfactory performance despite utilizing a pure Swin Transformer backbone. Stacking Swin Transformer blocks proves insufficient for forest RS images, even with excellent global modeling capability. TransUnet sequentially combines Transformer and convolutional layers, achieving 6.25% MIoU gain over Swin–Unet. GSwin–Unet outperforms TransUnet with 2.69% MIoU and 3.69% Ave.F1 improvements, validating the efficacy of integrating GCN with Swin Transformer.

The segmentation results of the methods in Table 3 are shown in Figure 11. Swin–Unet has limited ability to detect small targets, leading to a disregard for the categories with a small proportion, such as bare land and buildings. Comparatively, GSwin–Unet reduces segmentation errors for geometrically complex and high-similarity features—including roads and water in rivers, forests, and agricultural land. In row 2, due to similar colors, other methods mistakenly identified “Agricultural land” as “Forest” and “Water” as “Forest”, while our GSwin–Unet made comparatively precise judgments. Furthermore, the example in row 4 shows that GSwin–Unet demonstrates superior efficacy in complex textured scenes and specific shaped ground objects, such as interwoven and nested forests, agricultural land, and roads.

Table 3. Segmentation results on the RGB forest dataset.

Model Name	IoU (%)						Evaluation Index	
	Forest	Agricultural Land	Water	Building	Bare Land	Road	MIoU (%)	Ave.F1 (%)
DeepLab V3+	57.60	68.66	61.67	78.38	78.07	63.96	68.06	80.73
Unet	61.67	73.17	70.54	79.46	73.17	62.07	70.01	82.20
Swin–Unet	67.23	70.63	67.26	66.67	62.83	62.50	66.19	79.62
TransUnet	66.39	75.21	68.10	76.52	79.65	68.75	72.44	80.23
GSwin–Unet	73.04	78.63	68.64	76.47	78.26	75.70	75.13	83.92
MIoU (%)	65.19	73.26	67.24	75.5	74.40	66.60		

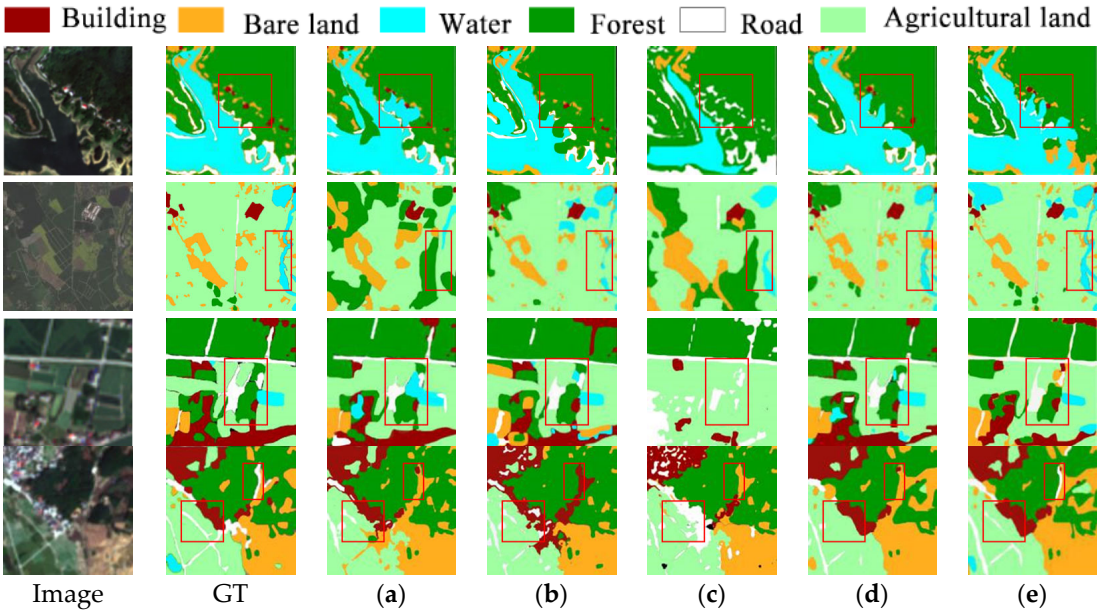

Figure 11. Semantic segmentation instances on the RGB forest dataset. (**a**) DeepLab V3+. (**b**) Unet. (**c**) Swin–Unet. (**d**) TransUnet. (**e**) GSwin–Unet.

Results on NIRRG Forest Dataset

The segmentation performance of each method on the NIRRG forest dataset is reported in Table 4. GSwin–Unet achieves 72.92% MIoU and 84.3% Ave.F1, surpassing all other methods. Owing to distinct spectral compositions, segmentation accuracy on the NIRRG forest dataset is slightly lower than that on the RGB forest dataset. However, the mean IoUs of forest, water, and road in the methods involved in Table 4 are 65.41%, 73.62% and 68.45%, respectively, which are higher than those on the RGB dataset (65.19%, 67.24%, and 66.60, respectively). This demonstrates the advantages of near-infrared in vegetation and water recognition. It should be pointed out that Swin–Unet is nevertheless inferior to CNN-based models, while TransUnet with a hybrid structure has a higher segmentation accuracy than the CNN-based models in Table 4. This confirms that the ability to model both local details and global context is equally important for large-scale forest RS images. Compared with Swin–Unet, GSwin–Unet increases MIoU by 6.79% and Ave.F1 by 4.72%. Compared with other methods, our GSwin–Unet demonstrates superior capability in capturing narrow linear features (e.g., roads and river) and delineating boundaries between forest and agricultural land, as shown in Figure 12 (Row 4). However, its segmentation performance for agricultural land is slightly lower than Unet and DeepLab V3+, and for water, slightly lower than TransUnet. This may be attributed to the model's limitations in distinguishing local fine-scale features.

Table 4. Segmentation results on the NIRRG forest dataset.

Model Name	IoU (%)						Evaluation Index	
	Forest	Agricultural Land	Water	Building	Bare Land	Road	MIoU (%)	Ave.F1 (%)
DeepLab V3+	63.39	73.21	70.23	61.90	63.33	70.94	67.17	80.28
Unet	60.87	76.58	72.22	65.04	71.68	70.83	69.54	81.92
Swin–Unet	66.12	67.21	71.20	66.12	64.96	61.21	66.13	79.58
TransUnet	66.96	67.20	77.78	71.93	70.69	68.33	70.48	82.63
GSwin–Unet	69.72	69.67	76.67	74.14	76.36	70.94	72.92	84.30
MIoU (%)	65.41	70.77	73.62	67.83	69.40	68.45		

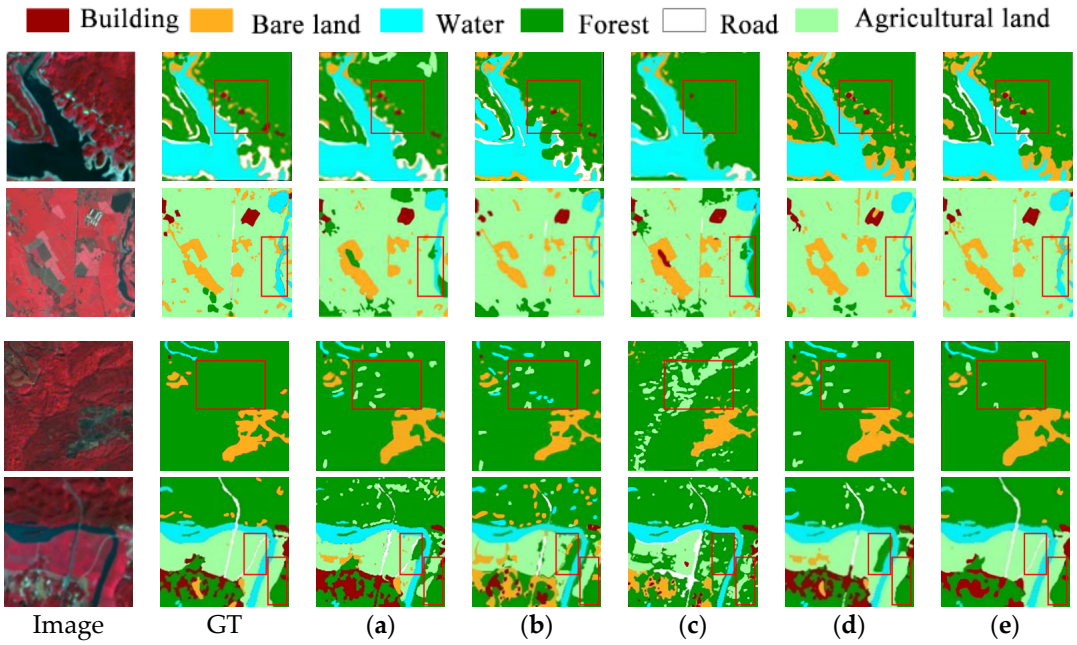

Figure 12. Examples of semantic segmentation results on the NIRRG forest dataset. (**a**) DeepLab V3+. (**b**) Unet. (**c**) Swin–Unet. (**d**) TransUnet. (**e**) GSwin–Unet.

Figure 12 visually presents the segmentation results. In rows 1–2, compared with corresponding rows in Figure 11, the methods mentioned in Table 4 can better distinguish "Agricultural land" and "Water" in the "Forest" due to the marked spectral discriminability between "Agricultural land" and "Water" in NIRRG imagery, and the contours recognized by the GSwin–Unet method are clearer. In row 3, "Forest" and "Agricultural land" have alike colors, and because of the specific angle of sunlight exposure, the "Forest" with high brightness is prone to misclassification as "Agricultural land". By aggregating discriminative features from both local neighbor nodes and global contextual texture features, GSwin–Unet maintains relatively accurate inference performance in this scenario. In row 4, our model better distinguishes between "Forest" and "Agricultural land".

Efficiency Analysis

The parameters and inference speed of the methods for comprehensive comparisons are listed in Table 5. In the table, "Speed" refers to the model's image processing rate per second, measured in FPS (frames per second). Regarding speed (computational efficiency), models with pure CNN structures are generally faster than models with a Transformer or Swin Transformer block. Specifically, our proposed GSwin–Unet achieves only 8 FPS on the RGB forest dataset and 9 FPS on the NIRRG dataset. Furthermore, as GSwin–Unet

adopts a parallel hybrid architecture combining Unet and Swin Transformer, its parameter count exceeds that of other methods.

Table 5. Model speed, accuracy, and parameters.

Method	Parameters	RGB Forest		NIRRG Forest	
		Speed (FPS)	MIoU (%)	Speed (FPS)	MIoU (%)
DeepLab V3+	38.48 MB	76	68.06	74	67.17
Unet	25.13 MB	239	70.01	232	69.54
Swin–Unet	25.89 MB	63	66.19	60	66.13
TransUnet	100.44	39	72.44	37	70.48
GSwin–Unet	160.97	9	75.13	8	72.92

Exploration of the Adaptability of GSwin–Unet to Fresh Datasets

To explore the adaptability of the proposed enhanced model to fresh datasets, we performed experiments on the publicly available RS GID-5. The segmentation results are compared in Table 6.

Table 6. Comparison of segmentation results on the GID dataset.

Model Name	IoU (%)						Evaluation Index	
	Forest	Farmland	Water	Building	Grass	Others	MIoU (%)	Ave.F1 (%)
DeepLab V3+	57.48	70.00	64.71	78.57	76.32	63.96	67.17	80.28
Unet	61.67	72.13	67.83	79.46	73.17	62.93	69.54	81.92
Swin–Unet	65.57	69.60	66.09	65.32	60.68	59.06	66.13	79.58
TransUnet	65.83	69.92	66.95	75.86	75.65	68.75	70.48	82.63
GSwin–Unet	71.30	74.79	71.43	73.33	73.55	72.90	72.92	84.30

It can be observed that GSwin–Unet achieves a relatively ideal accuracy in identifying various categories in the GID dataset. In comparison to TransUnet, the overall accuracy of GSwin–Unet is 2.44% higher, with significant improvements in the categories of Forest, Farmland, and Water, all of which have increased by 4%.

4. Discussion

The Zero-DCE module introduces an additional 0.303 MB of parameters and demonstrates exceptionally high inference speed for 224 × 224 images, while exerting negligible impact on overall model efficiency. Experimental results indicate that Zero-DCE effectively enhances image brightness, contrast, and detail in low-light environments, particularly excelling in scenarios with uneven or severely insufficient illumination. However, under well-lit conditions with high image quality, its enhancement effects are limited, and over-enhancement may even introduce noise or color distortion, leading to insignificant or slightly decreased quantitative metrics. Subsequent research will categorize forest RS images based on illumination intensity (e.g., low-light and normal illumination) to compute performance indicators for each subset, thereby clarifying the applicability of Zero-DCE under varying forest lighting conditions.

The combined function of Dice and Focal achieves a 0.72% higher segmentation accuracy than the combined function of CE and Dice. This indicates that in complex forest-like scenarios, more attention should be paid to hard samples. In the future, more combinations of loss functions and different weight settings will be explored to find more suitable loss functions for forest scenarios.

The conclusion of this ablation study is based on results from a single random seed (seed 42) on the RGB forest dataset. Although the experiments observed improvements in MIoU (approximately 0.5–2%) with the FAM, Zero-DCE, and the alternative loss (L2), this improvement may be influenced by random factors such as model initialization and data partitioning. Therefore, the small performance gains in the current results should be interpreted with caution, as their robustness may be limited. To strengthen the study, future work will involve multi-seed validation, cross-dataset generalization analysis, and statistical testing to ensure the reliability of the conclusions.

The GSwin–Unet model demonstrates superior semantic segmentation accuracy compared to baseline models such as DeepLab V3+, Unet, and Swin–Unet on both the self-developed forest dataset and selected GID data subsets. TransUnet performs exceptionally on medical images but falls short of GSwin–Unet on remote sensing images. This may be due to the large-scale variations in remote sensing imagery, where the task focuses more on macro-level object classification and global modeling capabilities, whereas medical tasks require precise segmentation of lesion boundaries, demanding higher local sensitivity. The superior global modeling ability, multi-scale fusion, and loss function of GSwin–Unet make it better suited for remote sensing data characteristics. This performance improvement holds substantial significance for forest resource management applications. In forest fire scenarios, the high-precision segmentation outputs provide critical decision support for firefighting operations, particularly in evacuation route optimization and firebreak deployment. Furthermore, when integrated with LiDAR/SAR data, the model may enhance the accuracy of parameter inversion for tree height and canopy density measurements, thereby improving the reliability of carbon stock estimation. In future research, we will systematically explore the model's adaptability across diverse forest types (e.g., dense evergreen forests, plantations, and mountainous forests) through targeted data subsetting, and evaluate its robustness under challenging conditions (e.g., partial occlusion from atmospheric effects or cloud coverage), with the aim of fully unlocking its application potential in forest management.

The current inference speed of the model (see Table 5) is 9 FPS, rendering it suitable for offline mapping workflows, such as high-precision map generation, with a processing time of up to 10 min. For real-time applications requiring sub-second response (e.g., forest fire monitoring), we suggest further optimizing the model or employing hardware acceleration to reduce latency. Existing findings demonstrate that pruning 20% of the channels reduces the model size by 35% with negligible MIoU loss, offering a viable trade-off for resource-constrained environments [59].

Overall, the prediction results of the exploratory experiments demonstrate that the proposed improved model exhibits good adaptability to fresh datasets and is well-suited for sematic segmentation tasks in forest RS images. The segmentation performance of the GID-5 dataset is slightly lower than that of Sentinel-2 forest dataset. This is most likely because Sentinel-2 images have a larger field of view, can present a wider range of scenes or objects, and have richer texture and shape features. Additionally, the GID-5 dataset used in the experiment is a large-scale classification set, which is not precise enough for classifying land features. Although the two limitations—efficiency and parameter size—may restrict GSwin–Unet's applicability in resource-constrained environments like compact mobile devices, it remains valuable for investigating GCN and Swin Transformer's potential in forest RS semantic segmentation tasks.

There are many possibilities for improvement through various auxiliary functions: (1) optimize the deep learning model through further exploring encoding methods for the small ground features embedded within predominant land cover types; (2) adopt a progressive segmentation method layer by layer by extracting different ground features

step by step from different band combination images of the same area; (3) train specialized sub-models for different regions and seasons to avoid performance degradation caused by a one-size-fits-all approach; (4) implement more precise pixel-level annotation and create refined forest land cover datasets; and (5) perform model compression to improve segmentation efficiency.

5. Conclusions

To obtain the global texture features and understand the complex relationships between different ground objects in forest RS imagery, and to enhance feature discriminability, we proposed and evaluated a semantic segmentation framework called GSwin–Unet. This framework features a dual-encoder structure by combining GCN and Swin Transformer. It represents a new attempt at combining model frameworks. Specifically, the Zero-DCE module improves the brightness and contrast of forest RS images, especially the details of texture feature of forest. The proposed feature aggregation module (FAM) leverages local aggregation features to guide the main encoder in extracting more discriminative representations; a joint loss function combining Dice loss (L_{Dice}) and Focal loss (L_{Focal}) is constructed to supervise the model to alleviate the problem caused by the imbalanced category proportions. Moreover, we constructed two forest datasets using different bands combinations and verified through comparative experiments that the NIRRG dataset has higher recognition accuracy in forests, water, and road. In contrast, the RGB dataset is superior in distinguishing agricultural land, buildings, and bare land. By integrating the discriminative advantages of the two different band combination images, we can achieve higher overall accuracy. Although the proposed framework has certain limitations in performance, our research findings have significant implications in the domain of complex RS image segmentation.

Author Contributions: Conceptualization: G.Z. and P.L.; Methodology: P.L. and J.L.; Data Collection: P.L.; Formal Analysis: P.L. and J.L.; Writing—Original Draft: P.L.; Writing—Review and Editing: G.Z.; Funding Acquisition: G.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the following grants: (1) Scientific Research Fund of Hunan Provincial Education Department (Grants 23B0244); (2) Science and Technology Innovation Platform and Talent Plan Project of Hunan Province (Grant 2017TP1022); (3) Natural Science Foundation of Hunan Province (Grant 2024JJ7645); and (4) Field Observation and Research Station of Dongting Lake Natural Resource Ecosystem, Ministry of Natural Resources.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: We thank the Mathematical Experiment Center of the School of Mathematics and Statistics at Central South University for their provision of support in data processing.

Conflicts of Interest: Author Jianzhong Li was employed by the company Changsha Shuobo Electronic Technology Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; Proceedings, Part III 18; Springer International Publishing: Berlin/Heidelberg, Germany, 2015; pp. 234–241. [CrossRef]
2. Hatamizadeh, A.; Tang, Y.; Nath, V.; Nath, V.; Yang, D.; Myronenko, A.; Landman, B.; Roth, H.; Xu, D. Unetr: Transformers for 3d medical image segmentation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2022; pp. 574–584. [CrossRef]
3. Wang, W.; Chen, C.; Ding, M.; Yu, H.; Zha, S.; Li, J. TransBTS: Multimodal Brain Tumor Segmentation Using Transformer. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021, Proceedings of the 24th International Conference, Strasbourg, France, 27 September–1 October 2021*; Lecture Notes in Computer Science; De Bruijne, M., Cattin, P.C., Cotin, S., Eds.; Springer: Cham, Switzerland, 2021; Volume 12901. [CrossRef]
4. 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, Tel Aviv, Israel, 23–27 October 2022; Springer Nature: Cham, Switzerland, 2022; pp. 205–218. [CrossRef]
5. Gong, Z.; French, A.P.; Qiu, G.; Xin, C. CTranS: A Multi-Resolution Convolution-Transformer Network for Medical Image Segmentation. In Proceedings of the 2024 IEEE International Symposium on Biomedical Imaging (ISBI), Athens, Greece, 27–30 May 2024; pp. 1–5. [CrossRef]
6. Huang, X.; Deng, Z.; Li, D.; Yuan, X. Missformer: An effective medical image segmentation transformer. *arXiv* **2021**, arXiv:2109.07162.
7. Hatamizadeh, A.; Nath, V.; Tang, Y.; Yang, D.; Roth, H. Swin unetr: Swin transformers for semantic segmentation of brain tumors in MRI images. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, Proceedings of the 7th International Workshop, BrainLes 2021, Virtual, 27 September 2021*; Springer International Publishing: Cham, Switzerland, 2021; pp. 272–284. [CrossRef]
8. Azad, R.; Jia, Y.; Aghdam, E.K.; Cohen-Adad, J.; Merhof, D. Enhancing medical image segmentation with TransCeption: A multi-scale feature fusion approach. *arXiv* **2023**, arXiv:2301.10847.
9. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.; Luo, P. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12077–12090.
10. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.; Zhuo, Y. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv* **2021**, arXiv:2102.04306.
11. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3146–3154. [CrossRef]
12. Huang, Z.; Wang, X.; Huang, L.; Huang, C.; Wei, Y.; Liu, W. CCNet: Criss-cross attention for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October 2019; pp. 603–612. [CrossRef]
13. Zhao, B.; Hua, L.; Li, X.; Lu, X.; Wang, Z. Weather recognition via classification labels and weather-cue maps. *Pattern Recognit.* **2019**, *95*, 272–284. [CrossRef]
14. 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. [CrossRef]
15. Xiao, T.; Liu, Y.; Zhou, B.; Jiang, Y.; Sun, J. Unified perceptual parsing for scene understanding. In *Computer Vision–ECCV 2018, Proceedings of the 15th European Conference, Munich, Germany, 8–14 September 2018*; Springer: Cham, Switzerland, 2018; pp. 418–434. [CrossRef]
16. Mou, L.; Hua, Y.; Zhu, X.X. Relation matters: Relational context-aware fully convolutional network for semantic segmentation of high-resolution aerial images. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 7557–7569. [CrossRef]
17. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Matthias, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
18. Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 6881–6890. [CrossRef]
19. Cheng, B.; Misra, I.; Schwing, A.G.; Alexander, K.; Rohit, G. Masked-attention mask transformer for universal image segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 1290–1299. [CrossRef]

20. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022. [CrossRef]
21. Chen, Z.M.; Wei, X.S.; Wang, P.; Guo, Y. Multi-label image recognition with graph convolutional networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5177–5186. [CrossRef]
22. Jung, H.; Park, S.Y.; Yang, S.; Kim, J. Superpixel-Based Graph Convolutional Network for Semantic Segmentation. 2021. Available online: http://vip.joonseok.net/courses/mlvu_2021_2/projects/final_papers/13.pdf (accessed on 2 June 2024).
23. Han, K.; Wang, Y.; Guo, J.; Tang, Y.; Wu, E. Vision GNN: An image is worth graph of nodes. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 8291–8303. Available online: https://proceedings.neurips.cc/paper_files/paper/2022/file/3743e69c8e47eb2e6d3afaea80e439fb-Paper-Conference.pdf (accessed on 12 June 2024).
24. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. Advances in neural information processing systems. *arXiv* **2017**, arXiv:1706.03762.
25. Wang, W.; Xie, E.; Li, X.; Fan, D.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 568–578. [CrossRef]
26. Han, K.; Xiao, A.; Wu, E.; Guo, J.; Xu, C.; Wang, Y. Transformer in transformer. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 15908–15919.
27. Yuan, L.; Chen, Y.; Wang, T.; Yu, W.; Shi, Y.; Jiang, Z.; Tay, F.E.H.; Feng, J.; Yan, S. Tokens-to-token vit: Training vision transformers from scratch on ImageNet. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 558–567. [CrossRef]
28. Wang, W.; Xie, E.; Li, X.; Fan, D.; Song, K.; Liang, D. Pvt v2: Improved baselines with pyramid vision transformer. *Comput. Vis. Media* **2022**, *8*, 415–424. [CrossRef]
29. Lin, A.; Chen, B.; Xu, J.; Zhang, Z.; Lu, G.; Zhang, D. DS-TransUnet: Dual swin transformer u-net for medical image segmentation. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 4005615. [CrossRef]
30. Zhang, Y.; Liu, H.; Hu, Q. Transfuse: Fusing transformers and CNNS for medical image segmentation. In Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, 27 September–1 October 2021; Proceedings, Part I 24; Springer International Publishing: Cham, Switzerland, 2021; pp. 14–24. [CrossRef]
31. Khan, A.R.; Khan, A. MaxViT-UNet: Multi-axis attention for medical image segmentation. *arXiv* **2023**, arXiv:2305.08396.
32. Kipf, T.N.; Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv* **2016**, arXiv:1609.02907.
33. Wan, S.; Gong, C.; Zhong, P.; Pan, S.; Li, G.; Yang, J. Hyperspectral image classification with context-aware dynamic graph convolutional network. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 597–612. Available online: https://arxiv.org/pdf/1909.11953 (accessed on 5 July 2024). [CrossRef]
34. Cai, Y.; Zhang, Z.; Cai, Z.; Liu, X.; Jiang, X.; Yan, Q. Graph convolutional subspace clustering: A robust subspace clustering framework for hyperspectral image. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 4191–4202. [CrossRef]
35. Liu, J.; Li, T.; Zhao, F.; Liu, Y. Dual Graph Convolutional Network for Hyperspectral Images with Spatial Graph and Spectral Multi-graph. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 5506105. [CrossRef]
36. Hong, D.; Gao, L.; Yao, J.; Zhang, B.; Plaza, A.; Chanussot, J. Graph convolutional networks for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 5966–5978. [CrossRef]
37. Ding, Y.; Guo, Y.; Chong, Y.; Pan, S. Global consistent graph convolutional network for hyperspectral image classification. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 5501516. [CrossRef]
38. Cheng, J.; Zhang, F.; Xiang, D.; Yin, Q.; Zhou, Y. PolSAR image classification with multiscale superpixel-based graph convolutional network. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5501516. [CrossRef]
39. Liu, H.; Zhu, T.; Shang, F.; Liu, Y.; Lv, D.; Yang, S. Deep fuzzy graph convolutional networks for PolSAR imagery pixelwise classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *14*, 504–514. [CrossRef]
40. Ren, S.; Zhou, F. Semi-supervised classification for PolSAR data with multi-scale evolving weighted graph convolutional network. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 2911–2927. [CrossRef]
41. Wu, D. Research on Remote Sensing Image Feature Classification Method Based on Graph Neural. Master's Thesis, Shandong Jianzhu University, Jinan, China, 2022. [CrossRef]
42. Chen, X. Vehicle Video Object Semantic Segmentation Method Based on Graph Convolution. Master's Thesis, Beijing Jianzhu University, Beijing, China, 2022. [CrossRef]
43. Ma, F.; Gao, F.; Sun, J.; Zhou, H.; Hussain, A. Attention graph convolution network for image segmentation in big SAR imagery data. *Remote Sens.* **2019**, *11*, 2586. [CrossRef]
44. Liu, Q.; Xiao, L.; Yang, J.; Wei, Z. CNN-enhanced graph convolutional network with pixel-and superpixel-level feature fusion for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 8657–8671. [CrossRef]

45. Liu, Q.; Kampffmeyer, M.; Jenssen, R. Self-constructing graph convolutional networks for semantic labeling. In Proceedings of the IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium, Waikoloa, HI, USA, 26 September–2 October 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1801–1804. [CrossRef]
46. Zhang, X.; Tan, X.; Chen, G.; Zhu, K.; Liao, P.; Wang, T. Object-based classification framework of remote sensing images with graph convolutional networks. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 8010905. [CrossRef]
47. Liu, Q.; Xiao, L.; Yang, J.; Wei, Z. Multilevel superpixel structured graph U-Nets for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5516115. [CrossRef]
48. Bandara, W.G.C.; Valanarasu, J.M.J.; Patel, V.M. Spin road mapper: Extracting roads from aerial images via spatial and interaction space graph reasoning for autonomous driving. In Proceedings of the 2022 International Conference on Robotics and Automation (ICRA), Philadelphia, PA, USA, 23–27 May 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 343–350. [CrossRef]
49. Liu, Q.; Kampffmeyer, M.C.; Jenssen, R.; Salberg, A. Multi-view self-constructing graph convolutional networks with adaptive class weighting loss for semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 16–18 June 2020; pp. 44–45. Available online: https://openaccess.thecvf.com/content_CVPRW_2020/papers/w5/Liu_Multi-View_Self-Constructing_ Graph_ Convolutional_Networks_With_Adaptive_Class_Weighting_Loss_CVPRW_2020_paper.pdf (accessed on 15 July 2024).
50. Cui, W.; He, X.; Yao, M.; Wang, Z.; Hao, Y.; Li, J.; Wu, W.; Zhao, H.; Xia, C.; Li, J.; et al. Knowledge and spatial pyramid distance-based gated graph attention network for remote sensing semantic segmentation. *Remote Sens.* **2021**, *13*, 1312. [CrossRef]
51. Zeiler, M.D.; Fergus, R. Visualizing and understanding convolutional networks. In Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Proceedings, Part I 13. Springer International Publishing: Cham, Switzerland, 2014; pp. 818–833. [CrossRef]
52. 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. [CrossRef]
53. Guo, C.; Li, C.; Guo, J.; Loy, C.; Hou, J.; Kwong, S.; Cong, R. Zero-reference deep curve estimation for low-light image enhancement. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 1780–1789. [CrossRef]
54. Fidon, L.; Li, W.; Garcia-Peraza-Herrera, L.C.; Ekanayake, J.; Kitchen, N.; Ourselin, S.; Vercauteren, T. Generalised wasserstein dice score for imbalanced multi-class segmentation using holistic convolutional networks. In *Brain Lesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, Proceedings of the Third International Workshop, BrainLes 2017, Held in Conjunction with MICCAI 2017, Quebec City, QC, Canada, 14 September 2017*; Revised Selected Papers 3; Springer International Publishing: Cham, Switzerland, 2018; pp. 64–76. [CrossRef]
55. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollar, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988. Available online: https://openaccess.thecvf.com/content_ICCV_2017/papers/Lin_Focal_Loss_for_ICCV_2017_paper.pdf (accessed on 5 July 2024).
56. Zhu, Q.; Zheng, Y.; Jiang, Y.; Yang, J. Efficient multi-class semantic segmentation of high resolution aerial imagery with dilated linknet. In Proceedings of the IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; pp. 1065–1068. [CrossRef]
57. He, X.; Zhou, Y.; Zhao, J.; Zhang, D.; Yao, R.; Xue, Y. Swin transformer embedding UNet for remote sensing image semantic segmentation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4408715. [CrossRef]
58. 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. Available online: https://openaccess.thecvf.com/content_ECCV_2018/papers/Liang-Chieh_Chen_Encoder-Decoder_with_Atrous_ECCV_2018_paper.pdf (accessed on 17 October 2024).
59. Liu, Z.; Li, J.; Shen, Z.; Huang, G.; Yan, S.; Zhang, C. Learning efficient convolutional networks through network slimming. In Proceedings of the IEEE International on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2736–2744. [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.