applied sciences

MDPI

Article

GLFFNet: A Global and Local Features Fusion Network with Biencoder for Remote Sensing Image Segmentation

Qing Tian, Fuhui Zhao, Zheng Zhang * and Hongquan Qu

School of Information Science and Technology, North China University of Technology, Beijing 100144, China; tianqing@ncut.edu.cn (Q.T.); 2021312100102@mail.ncut.edu.cn (F.Z.); qhqphd@ncut.edu.cn (H.Q.)
* Correspondence: zhangzheng@ncut.edu.cn

Abstract: In recent years, semantic segmentation of high-resolution remote sensing images has been gradually applied to many important scenes. However, with the rapid development of remote sensing data acquisition technology, the existing image data processing methods are facing major challenges. Especially in the accuracy of extraction and the integrity of the edges of objects, there are often problems such as small objects being assimilated by large objects. In order to solve the above problems, based on the excellent performance of Transformer, convolution and its variants, and feature pyramids in the field of deep learning image segmentation, we designed two encoders with excellent performance to extract global high-order interactive features and low-order local feature information. These encoders are then used as the backbone to construct a global and local feature fusion network with a dual encoder (GLFFNet) to effectively complete the segmentation of remote sensing images. Furthermore, a new auxiliary training module is proposed that uses the semantic attention layer to process the extracted feature maps separately, adjust the losses, and more specifically optimize each encoder of the backbone, thus optimizing the training process of the entire network. A large number of experiments show that our model achieves 87.96% mIoU on the Potsdam dataset and 80.42% mIoU on the GID dataset, and it has superior performance compared with some state-of-the-art methods on semantic segmentation tasks in the field of remote sensing.

Keywords: remote sensing image; gated convolution; transformer; atrous convolution

Citation: Tian, Q.; Zhao, F.; Zhang, Z.; Qu, H. GLFFNet: A Global and Local Features Fusion Network with Biencoder for Remote Sensing Image Segmentation. *Appl. Sci.* **2023**, *13*, 8725. https://doi.org/10.3390/app13158725

Academic Editor: Francesco Zirilli

Received: 10 May 2023
Revised: 25 July 2023
Accepted: 26 July 2023
Published: 28 July 2023

1. Introduction

In recent years, with the rapid development of remote sensing technology, the amount of available remote sensing data has increased significantly. Semantic segmentation technology in remote sensing images has made many contributions to environmental monitoring, crop cover and type analysis, tree species in forests, building classification in urban space, land use analysis, and other important scenes that promote sustained economic growth and continuous improvement of life quality. As a result, it has significant implications for further study to design an excellent semantic segmentation network in the field of high-resolution remote sensing [1].

With the development of sensor technology, the spatial resolution of remote sensing images is constantly improving, and a lot of rich and important information can be obtained from remote sensing images. But the technology for extracting useful feature information from remote sensing data still has a lot of room for improvement. The semantic information in the actual scene is very complex, and the objects of the same semantic class may use different materials or different construction methods, which lead to the diversity of color, size, shape, and texture; in addition, objects of different semantic classes can present similar features, such as cement roofs, cement sidewalks, and cement roads. This high within-class variability and low between-class variability make segmentation scenes very complicated [2]. At the same time, objects occupying fewer pixels are often covered by others occupying more pixels, resulting in small objects being ignored and leading to a large number of false segments.

Most of the traditional semantic segmentation techniques first process the image to extract the semantic features of the image and then segment it. Therefore, the accuracy is limited, and the scope of application is very limited. The deep learning method, which adjusts the feature extraction process based on segmentation results, can extract more abstract, high-level feature information and is more suitable for dealing with complex scenes. As shown in Figure 1, the U-Net series [3] and the DeepLab series [4–8] have shown strong performance in the semantic segmentation field of computer vision. In remote sensing image processing, researchers have considered the specific characteristics of remote sensing data and improved the classical semantic segmentation network to further improve its segmentation performance [9–15]. Ziaee et al. [16] proposed a new G2G network based on the Pix2Pix network to segment the edges of special-class objects more precisely. This network upgrades the segmentation accuracy of objects of interest, but it is not suitable for complex scenes in remote sensing. Chen et al. [17] proposed a dense residual neural network, DR-Net, which can obtain more high-level semantic feature information for subsequent segmentation by combining the advantages of dense convolutional neural networks and residual networks. However, it suffers from overfitting problems, and its generalization ability has a lot of room for improvement. Yang et al. [18] proposed AFNet, which includes a multipath encoder structure for extracting multipath input features, a multipath attention fusion module for fusing multipath features, and a fine attention fusion module for fusing high-level abstract features and low-level spatial features. And it has achieved excellent performance on the Potsdam dataset, but its performance for small object segmentation needs to be further improved. Foivos I. et al. [2] proposed ResUNet-a, which combines atrous convolution, residual connection, pyramid principle, and multi-task inference theory to mainly alleviate the problem of height imbalance in remote sensing datasets. But its performance still cannot meet the demand for accuracy in practical application scenarios. Wang et al. [19] proposed a Transformer-based decoder and a novel attention mechanism to improve segmentation performance for real-time remote scenes. The experimental results show that it outperforms other state-of-the-art models in terms of segmentation accuracy. But their work on encoders is still slightly inadequate and needs further research. In general, semantic segmentation methods based on deep learning effectively extract construction features and improve segmentation accuracy. But in the complex scene, with the characteristics of high within-class variability and low between-class variability, the edge accuracy of small object segmentation is low, the segmentation detection error is high, the segmentation efficiency is insufficient, and so on, and the further application of the technology is still hindered.

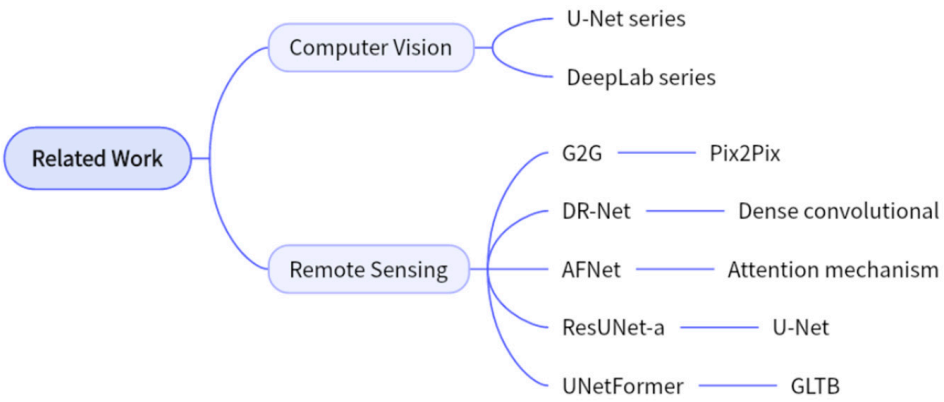

Figure 1. This figure shows some classical semantic segmentation models in the fields of computer vision and remote sensing, which are referred to in this paper.

In order to solve the problems above, we conducted a comprehensive analysis of advanced technologies in related fields. We find that Transformer contributes greatly to the improvement of accuracy in the field of image segmentation, but its high computational

cost also brings great difficulties to the practical application of relevant models. In this regard, convolution has many advantages. Later, we discovered that some research on implementing Transformer's strong spatial interaction capability with convolution has achieved good results. However, we found that it is difficult to completely replace Transformer with convolution, so we combined the two to complement each other. Specifically, based on the Transformer [20], atrous convolution [5], and gated convolution [21] models, this paper proposes a global and local feature fusion network with a biencoder (GLFFNet) for high-resolution remote sensing image segmentation.

Most importantly, the main contributions of this work are as follows:

1. In order to solve the problem of high within-class variability of objects of interest in complex remote sensing images, an efficient spatial feature aggregation encoder (SFA encoder) is proposed for feature extraction. The attention module of the Transformer is incorporated into the recursive gated convolution [22] to perform self-attention operations on feature maps, overcoming the limitations of convolutional networks, such as poor global modeling capability and insufficient exploitation of spatial location information. It is used to extract more abundant multi-scale feature information and improve the segmentation accuracy of the whole model.
2. Aiming at the low between-class variability of objects of interest in complex scenes of high-resolution remote sensing and the related problems they raise. In this paper, a spatial pyramid atrous convolutional encoder (SPAC encoder) is proposed to extract shallow feature information from remote sensing images using spatial pyramid structure and atrous convolutional to maximize the retention of local semantic information for each pixel. This is used to improve the model's recognition of small objects.
3. In order to better transfer the weight of the backbone pre-trained on classification datasets to segmentation tasks, this paper proposes an auxiliary training module called Multi-head Loss Block. This module employs multiple semantic attention layers and four lightweight decoders of the segmentation task to fine-tune the weights of each encoder in the backbone. Through this auxiliary model, encoders can extract more abundant multi-scale feature maps for objects of different sizes, which makes the whole training process convergence faster and more stable.

2. Method

2.1. Overview Structure

The overall structure of GLFFNet proposed in this paper is shown in Figure 2, which contains four key modules: (1) SFA Encoder, which is mainly composed of recursive gated convolution with attention block (RG Convolution with Attention), (2) SPAC Encoder, which is mainly composed of atrous convolution and based on spatial pyramid architecture, (3) Decoder, which is mainly composed of the global-local Transformer block (GLTP) and Feature refinement head (FRH) [19], and (4) Multi-head Loss Block, which is an auxiliary training module. The backbone of GLFFNet consists of the SFA Encoder and the SPAC Encoder.

GLFFNet mainly uses atrous convolution, gated convolution, and Transformer to extract multi-layer semantic feature information from images. Then the feature information is fused and decoded to generate the final segmentation results by decoder, so as to achieve accurate description of objects of different sizes. In addition, GLFFNet uses Multi-head Loss Block-assisted training to better transfer the weight of backbone pre-trained on classification datasets to segmentation tasks.

In this paper, GLFFNet is applied to semantic segmentation task of remote sensing images. The backbone of GLFFNet consists of four SFA encoders and a SPAC Encoder. Four SFA Encoder blocks sequentially processed the images after data enhancement and extracted the feature maps of 4, 8, 16, and 32 times downsampling, respectively. A SPAC Encoder block extracts a 4-times downsampled feature map directly from the image after data enhancement. Then, the 4-times downsampled feature maps extracted by SPAC Encoder and SFA Encoder are weighted and fused. Finally, feature maps of these four

scales are used as the output of backbone and input decoder for decoding. The decoder of GLFFNet is the decoder of UNetFormer, a lightweight decoder based on Transformer, which will fuse the input feature maps of four scales to generate the final segmentation result. Especially, Multi-head Loss Block is used in the training process, and the auxiliary training module consists of semantic attention layer and lightweight decoder of segmentation tasks and is used to calculate a weighted loss for optimization of model.

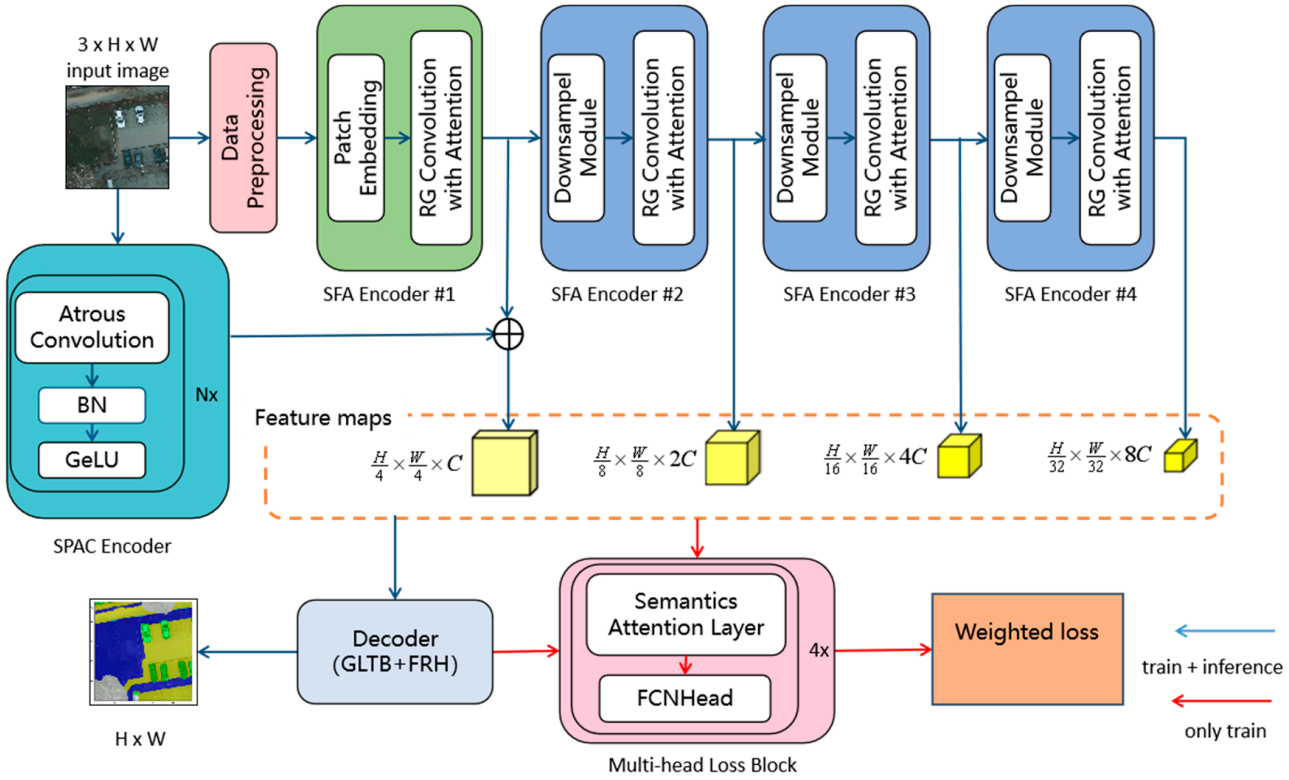

Figure 2. The overall structure of GLFFNet.

2.2. SFA Encoder

In order to obtain rich and important feature information from remote sensing images, this paper combines the attention mechanism and the recursive gated convolution with efficient high-order interaction ability, and proposes the recursive gated convolution with attention block, which is the core module of SFA Encoder.

The structure of the recursive gated convolution with attention block is shown in Figure 3, which consists of a window attention block and a recursive gated convolution block. In the window attention block, we use LN to normalize the feature map before and after attention operations and use the dropout layer and the global average pooling [23] in the MLP block to prevent overfitting. In the recursive gated convolution with attention block, the feature maps first go through the window-attention block for global modeling, extracting features, and synthesizing the global information. Then they go through a recursive gated convolution block to extract spatial and local bias information through efficient higher-order spatial interactions. Finally, some feature maps with more comprehensive and detailed feature information are output.

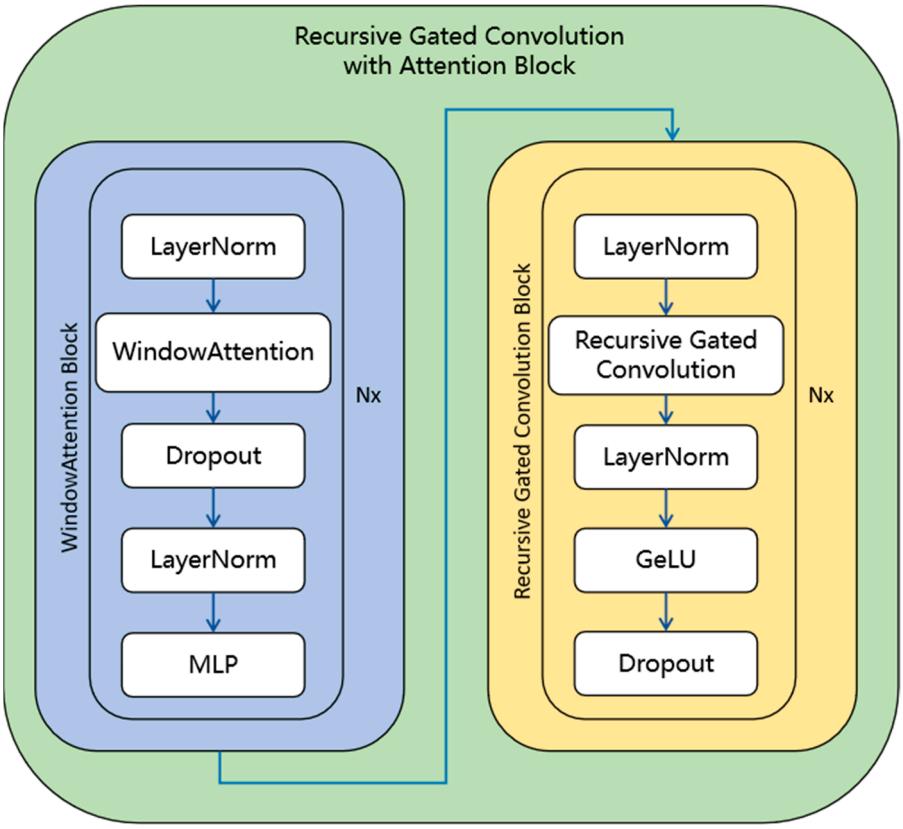

Figure 3. The structure of the recursive gated convolution with attention block.

SFA Encoder uses attention mechanism [24] of Transformer to achieve globe modeling. The expression of attention mechanism is as follows:

$$Attention(Q, K, V) = Softmax(\frac{Q \times K^T}{\sqrt{d_k}}) \times V \tag{1}$$

where Q, K, and V are, respectively, from the mapping of the input $X \in R^{N \times C}$ to different feature spaces. The expression of a linear transformation is as follows:

$$Q = X \times W_Q,\ K = X \times W_K,\ V = X \times W_V \tag{2}$$

SFA Encoder uses recursive gated convolution to achieve high-order spatial interaction. Firstly, we use linear layer φ_{in} to obtain a set of projected feature maps p_0 and $\{q_k\}_{k=0}^{n-1}$:

$$\left[p_0^{C_0 \times H \times W}, q_0^{C_0 \times H \times W}, \cdots, q_{n-1}^{C_{n-1} \times H \times W}\right] = \varphi_{in}(x) \in R^{(C_0 + \sum_{0 \le k \le n-1} C_k) \times H \times W} \tag{3}$$

then we process the set of projected feature maps by the recursive gated convolution,

$$p_{k+1} = f_k(q_k) \otimes g_k(p_k)/\alpha = 0, 1, 2, \cdots, n-1 \tag{4}$$

where f_k is the gated convolution and g_k is a function to match the dimension in different orders. Finally, we use linear layer φ_{out} to obtain the feature maps:

$$y = \varphi_{out}(p_{n-1}) \in R^{C \times H \times W} \tag{5}$$

The first SFA Encoder is composed of patch embedding layer and recursive gated convolution with attention block, which generates a 4-times downsampled feature map. The other three SFA Encoders are composed of downsample module and recursive gated

convolution with attention block, which generate an 8, 16, and 32-times downsampled feature map, respectively. The downsampling operation of SFA Encoder is completed by patch embedding layer and downsample module. And the recursive gated convolution with attention block is used to extract features from the input feature map and does not change the size or channel number of the feature map.

2.3. SPAC Encoder

A spatial pyramid atrous convolution encoder (SPAC Encoder) is proposed to obtain the spatial features of objects of various sizes in remote sensing images and avoid the disappearance of small objects due to the loss of information in the convolution process. Based on the spatial pyramid principle, SPAC Encoder extracts and fuses feature information from feature maps by using atrous convolution with different dilation rates. This module is simple and has fewer convolutional layers. It aims to extract relatively low-order pixel-level features from original images and retain the semantic information as much as possible.

The structure of SPAC Encoder proposed in this paper is shown in Figure 4. The processing of input image by SPAC Encoder mainly includes three parts. In the first part, the original image is processed with the atrous convolution of $d = 2$ (dilation rate). The expression is as follows:

$$y_1 = f_{d=2}(x), x \in R^{C \times H_0 \times W_0} \tag{6}$$

where $f_{d=2}$ is atrous convolution of $d = 2$. In the second part, the feature maps in the first part of the output are respectively processed with three different expansion rates of the atrous convolution, and then using the global average pooling operation, a unified output size is output, resulting in three of the same size feature maps. The expression is as follows:

$$y_{2k} = f_{d=2\times k}(y_1), k = 1, 2, 3 \tag{7}$$

where $y_{2k} \in R^{C \times H_{2k} \times W_{2k}}$ is output of atrous convolution with $d = 2 \times k$ in the part. In the third part, the cat function is first used to join the three feature maps in the second part of the output, and then 1×1 convolution is used to adjust the channel number of feature maps. The expression is as follows:

$$y_3 = g(y_{21} \oplus y_{22} \oplus y_{23}) \tag{8}$$

where $\oplus$ is concatenation with channel, g is 1×1 convolution to adjust the number of channels, and $y_3 \in R^{C \times \frac{H_0}{4} \times \frac{W_0}{4}}$ is final output of SPAC Encoder. To sum up, a feature map with the same output size as the 4-times downsampled feature map extracted by SFA Encoder is finally obtained by SPAC Encoder, which is involved in the subsequent fusion. In addition, in the SPAC Encoder, each convolutional layer is normalized by BN [25] and activated by GeLU [26].

2.4. Multi-Head Loss Block

With the rapid development of deep learning, model training requires more and more data. However, so far, the labeling cost of semantic segmentation datasets is still very large, and the semantic segmentation datasets that can be used for training obviously cannot meet the needs of the training of semantic segmentation models based on deep learning. Therefore, semantic segmentation networks based on deep learning almost always replace their own decoder with the decoder of the classification network first and do pre-training on the classification dataset to obtain the weights of that network's backbone. Then it is swapped back to the decoder of semantic segmentation network, fine-tuned on the semantic segmentation dataset. And finally, a complete semantic segmentation model is obtained. However, there are obvious differences between semantic segmentation and classification tasks. This training method can lead to some designs in semantic segmentation network models failing to fully realize their original design intention.

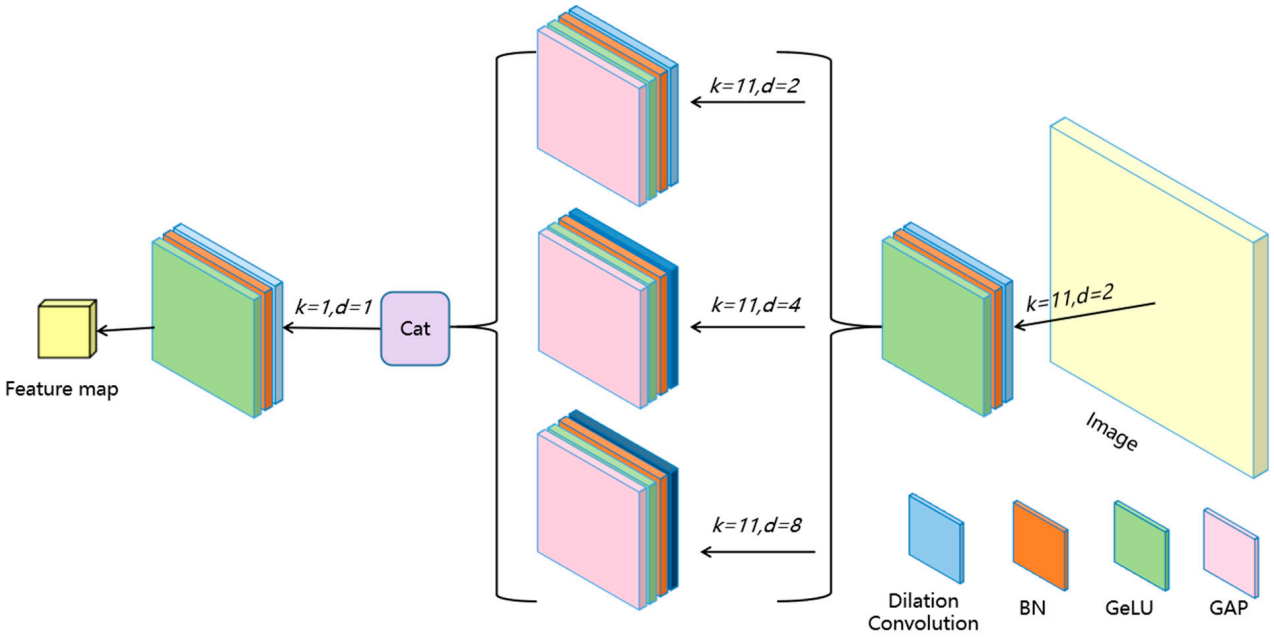

Figure 4. The overall structure of the spatial pyramid atrous convolution encoder.

Specifically, in the semantic segmentation task, an image contains objects of various sizes. However, most models now try to obtain as much global information as possible by constantly reducing the size of the feature map. This makes the models pay too much attention to global information, which in turn makes the strong object with too many pixels in the image completely cover the weak object with too few pixels. This eventually leads to, in the semantic segmentation task, the problem of small objects being ignored and unevenly segmented edges being assimilated by large objects. In order to solve this problem, most common semantic segmentation networks use a multi-scale structure to retain the local information of pixels as much as possible by adding large-size feature maps so as to detect small objects [27]. However, due to the limitations of the dataset, the network using the multi-scale structure still needs to conduct pre-training on the classification dataset. As a result of this training method, encoders for extracting large-size feature maps still pay more attention to global information, and the fine-tuning of semantic segmentation datasets has little effect on alleviating this problem. These ultimately led to the failure of this multi-scale design to achieve the desired effect.

In order to alleviate the above-mentioned problem of poor performance of multi-scale structures caused by pre-training on a classification dataset, this paper proposes a Multi-head Loss Block as an auxiliary training module. It can calculate the auxiliary losses of the feature maps of each output scale, and then these auxiliary losses are weighted and fused with the loss calculated by the main decoder for the back propagation during network training. The auxiliary training module proposed in this paper, Multi-head Loss Block, specifically optimizes the encoder for extracting feature maps of each scale so that the feature maps of each scale extracted by the network can save more effective information.

In GLFFNet, Multi-head Loss Block makes GLFFNet output feature maps of different scales as close as possible to the desired semantic segmentation results, thus improving the overall performance of GLFFNet. The feature maps of different scales have different values for improving the final performance of semantic segmentation networks, which are expressed by different weights. In order to avoid the influence caused by subjective weight setting, the weight is automatically obtained through network training. In this method, the feature maps of different scales of the network are obtained during training, the feature maps of each scale are processed, and loss is calculated. The loss calculated by the feature maps of various scales is weighted and fused to obtain the final weighted loss. Then, the weighted loss is used in the gradient descent process to update the model parameters until

the model converges and the final model is obtained. Since this method only changes the calculation method of loss in the model training but does not change the structure of the model or its inference process, there is no extra cost in the model inference process. This method is easy to implement and can effectively improve the performance of the model without increasing its complexity or overhead.

The Multi-head Loss Block structure used in GLFFNet is shown in Figure 5, which mainly includes semantic attention layer and lightweight decoder of the segmentation task, FCNHead. For the feature map of each input scale, Multi-head Loss Block first extracts the semantic information of the feature map with semantic attention layer and then inputs it into FCNHead for decoding and output segmentation results. After calculating the CE loss of five decoders, the weight of loss calculated by main decoder is artificially set to 1.0, a full connection layer is used to learn the weights of the other four auxiliary losses in the weighted fusion, and the weight range is specified as 0.1 to 0.4. Finally, the weighted loss is obtained by weight fusion.

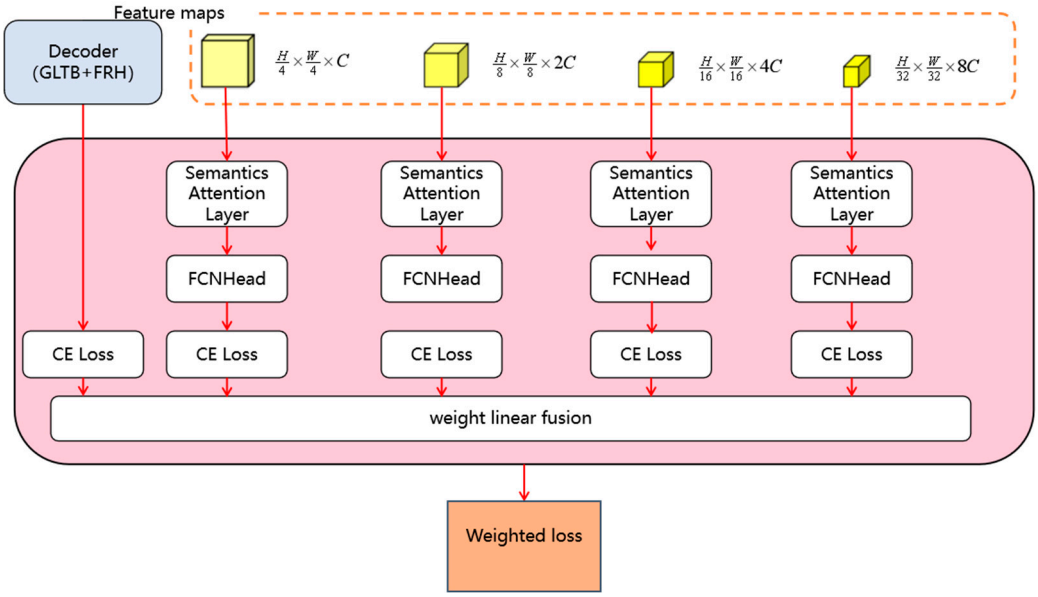

Figure 5. The overall structure of the Multi-head Loss Block.

The formula for calculating weighted loss in Multi-head Loss Block is as follows:

$$L_{Multi-head} = \alpha_0 L_{CE\_0} + \sum_{s=1}^{S} \alpha_s L_{CE\_s} \quad (9)$$

where $L_{Multi-head}$ is the weighted loss ultimately used for gradient descent in the whole model training process, α_0 is loss weight of the output result of main decoder and is set as 1.0, L_{CE_0} is loss calculated by the output of main decoder, α_s is the loss weight of the output result of the s-th decoder and obtained by learning, and its range is from 0.1 to 0.4, and L_{CE_s} is loss calculated by the output of the s-th decoder.

3. Experiment

3.1. Evaluation Criteria

In order to provide a comprehensive assessment of the performance of different structural models in the field of semantic segmentation for high-resolution remote sensing, we use four metrics, including mean intersection over union (MIoU) [28], mean accuracy (MAcc) [29], pixel accuracy (PAcc) [30], and inference time (ms). Where MIoU provides a more accurate measure of the segmentation accuracy of the model at the pixel level and is robust to problems such as category imbalance that may occur in the dataset. Therefore, our

experiments use MIoU as the main evaluation criteria. The formulas for the three accuracy evaluation metrics are as follows:

$$MIoU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{TP}{FN+FP+TP} \tag{10}$$

$$MAcc = \frac{1}{k+1}\sum_{i=0}^{k}\frac{TP}{FP+TP} \tag{11}$$

$$PAcc = \frac{TP+TN}{TP+TN+FP+TN} \tag{12}$$

3.2. Datasets

In our experiments, the performance of each model on the ISPRS Potsdam dataset is analyzed in detail. The Potsdam dataset contains 38 images taken from the Potsdam region, all of which are 6000 × 6000 in size. Each image was semantically labeled into six categories of objects. Due to the limitations of input image size and GPU memory for model processing, we cropped the original image and obtained a 512 × 512 image for the experiment. Then we randomly divide the resulting dataset, with 80% as the training set and 20% as the test set. In addition to this, in order to validate the usefulness of the model for the problem of object diversity in the image, the GID high-resolution remote sensing dataset was chosen as an auxiliary dataset for further evaluation of the model. The processing of the GID dataset is the same as that of the Potsdam dataset and will not be repeated. Finally, the training set of the Potsdam dataset consists of 4263 images, and the test set consists of 1065 images. The training set of the Potsdam dataset consists of 10,379 images, and the test set consists of 2587 images.

3.3. Compare Models and Experimental Design Details

In this section, the performance of GLFFNet on the Potsdam and GID datasets is compared with that of six representative semantic segmentation models in recent years to prove the superiority of GLFFNet proposed in this paper. The networks selected are as follows: DeepLabv3+ [8] that is a classical semantic segmentation network using deeply separable convolution, CAE(ViT-L, UperNet) [31] that uses a new mask image modeling and context auto encoder, ABCNet [32] that uses Bezier curves to make a new concise parametric representation of the curved scene text to accurately locate the oriented and curved scene text in the image, BANet [33] that proposed a new bilateral structure consisting of a convolutional layer and Transformer block with a view to capturing both local texture information and full-length dependency information in the feature network, SegFormer [34] that used a novel positional-encoding-free and hierarchical Transformer encoder, and UNetFormer [19] that uses a Transformer-based decoder for real-time urban scene segmentation. In order to ensure fairness, this paper chooses the configuration of each network as closely as possible.

In our experiments, we fully train all models on the same dataset as described above and ensure the fairness of the experiments by setting up the same training environment. The three most important parameters in the training process are batch size, learning rate, and number of training iterations. Since some models are so large that the training process requires a lot of GPU memory, we uniformly set the batch size to 8. Then we use 1×10^{-5} as the initial learning rate to start the training, and the learning rate decays with the training process. Finally, after completing the training of 100 K iterations, the model with the best results is selected for testing, and the model's performance is evaluated using accuracy and inference speed as criteria, respectively. Beyond the initial, we use Adam [35] as the optimization strategy during the training process. All code is based on the MMSegmentation implementation, and all training is done on NVIDIA RTX2080TI GPUs.

3.4. Ablation Experiment

In order to verify the performance of GLFFNet, we selected these corresponding modules with excellent performance, replaced the key modules of GLFFNet, and designed an ablation experiment, including: the SFA Encoder of GLFFNet was replaced with the encoder module of HorNet [22] and Mask2Former [36], the SPAC Encoder module of GLFFNet was removed, and the auxiliary training method, Multi-head Loss Block, was replaced with the common single loss training method.

3.4.1. Effect of SFA Encoder

In order to verify the effectiveness of the SFA Encoder, the paper used the SFA Encoder, the encoder (based on recursive gated convolution) of HorNet, and the Swin Transformer [37] of Mask2Former, respectively, as the backbone of GLFFNet, and other modules of experiments were kept consistent (specifically, to simply compare the advantages of the SFA Encoder, the SPAC Encoder proposed in this paper is not added to any of the three backbones). The experimental results are shown in Table 1. Compared with the other two models, the SFA Encoder shows a significant improvement in terms of these metrics.

Table 1. SFA Encoder ablation experiment results.

Backbone	MIoU (%)	MAcc (%)	PAcc (%)
Backbone (HorNet)	86.15	92.57	93.64
Swin Transformer (Mask2Former)	86.09	92.62	93.81
SFA Encoder	**87.37**	**93.36**	**94.24**

3.4.2. Effect of SPAC Encoder

In order to verify the effectiveness of the SPAC Encoder module designed in this paper, SFA Encoder and SFA Encoder + SPAC Encoder are respectively acting as the backbone of GLFFNet, and other modules are consistent for ablation experiments. The experimental results are shown in Table 2. From these three metrics, the SPAC encoder has made a certain contribution to the performance improvement of the model. Specific to the five classes of foreground objects in the Potsdam dataset, the experimental results are shown in Table 3 (MIoU (%) as the metrics). The segmentation effect of GLFFNet with the SPAC Encoder module is slightly reduced for building, but improved for other classes. In particular, the recognition ability of low-vegetation objects has been greatly improved. To sum up, the comprehensive performance of GLFFNet with the SPAC Encoder module has been improved.

Table 2. SPAC Encoder ablation overall experiment results.

Backbone	MIoU (%)	MAcc (%)	PAcc (%)
SFA Encoder	87.37	93.36	94.24
SFA Encoder + SPAC Encoder	**87.96**	**93.84**	**94.75**

Table 3. SPAC Encoder ablation experiment detailed results.

Backbone	Impervious-Surface	Building	Low-Vegetation	Tree	Car
SFA Encoder	85.45	**87.64**	74.18	93.34	96.25
SFA Encoder + SPAC Encoder	**85.84**	87.34	**76.36**	**93.68**	**96.58**

3.4.3. Effect of Multi-Head Loss Block

In order to verify the contribution of the Multi-head Loss Block proposed in this paper to the model accuracy and stability of the training process, we designed controlled experiments. We use two different methods for training. One is to train using the complete

model structure as described in Figure 2. The other is to remove the red line portion as described in Figure 2, without using the Multi-head Loss Block to normalize the feature map of each scale separately and using only the results of the main decoder to compute the loss used for backpropagation. After completing the experiment, we analyzed the results statistically. Figure 6 shows the changing trend of MIoU in two training ways. Table 4 records the highest evaluation criteria of the two training methods, and it is clear that Multi-head Loss Block contributes to both training efficiency and final model accuracy.

Figure 6. The training result of multi-head loss.

Table 4. Comparison of results of Multi-head Loss Block ablation experiment.

Method	MIoU (%)	MAcc (%)	PAcc (%)
single-head loss	86.74	92.57	93.64
multi-head loss	**87.96**	**93.84**	**94.75**

3.5. Comparison with State-of-the-Art Methods

In this module, we compare the GLFFNet proposed in this paper with some excellent similar models. The experiment results of all models on the Potsdam dataset are shown in Table 5. The experiment results show that the GLFFNet proposed achieves an excellent result of 87.96% on MIoU, surpassing other models. It should be additionally noted that our research is mainly aimed at improving the accuracy of the model and therefore has not yet reached a very satisfactory state in terms of inference time. However, compared with most of the other Transformer-like large models, our inference time still has some advantages since our auxiliary training module is not involved in inference. Of course, we will continue to improve the model to further enhance its overall performance.

The experiment results of all models on the GID dataset are shown in Table 6. The experiment results show that the GLFFNet proposed achieves an excellent result of 80.42% on MIoU. This result further demonstrates the usefulness of GLFFNet for the problem of object diversity in the image and how GLFFNet can be applied to a general environment.

Table 5. Comparative experimental results on the Potsdam dataset with other methods.

Method	MIoU (%)	MAcc (%)	PAcc (%)	Inference Time (ms)
DeepLabv3+ [8]	80.64	88.38	89.26	25.87
CAE(ViT-L, UperNet) [31]	79.67	87.51	88.34	85.14
ABCNet [32]	84.02	89.64	91.15	78.34
BANet [33]	85.67	89.91	91.34	73.51
SegFormer [34]	85.43	91.95	92.33	82.65
UNetFormer [19]	86.80	93.23	93.89	53.41
GLFFNet	**87.96**	**93.84**	**94.75**	**56.37**

Table 6. Quantitative comparison results on the GID dataset with some state-of-the-art methods.

Method	MIoU (%)	MAcc (%)	PAcc (%)
DeepLabv3+ [8]	72.25	79.22	81.07
CAE(ViT-L, UperNet) [31]	71.90	78.27	80.99
ABCNet [32]	76.57	82.39	82.82
BANet [33]	77.38	82.59	82.43
SegFormer [34]	78.43	83.24	83.37
UNetFormer [19]	79.35	84.57	85.97
GLFFNet	**80.42**	**85.08**	**86.23**

Figure 7 shows the partial segmentation result on the Potsdam dataset. Due to limited space, only the four segmentation results of GLFFNet (ours) and UNetFormer that are the best effects in the selected comparison model are shown. Figure 7 consists of four columns. The first column is the original image, the second column is the labeled image, the third column is the segmentation result of GLFFNet, and the fourth column is the segmentation result of UNetFormer. To make a better comparison, we have circled the key positions with red ovals.

In the first set of images, the two cars on the road in the upper right are very closely spaced, which makes accurate recognition by the model very difficult. The presence of the interval between the two cars is clearly visible in the recognition results of our GLFFNet (third column), while in the recognition results of UNetFormer (fourth column), the road between these two cars is mis-segmented. In the second set of images, the object in the top right corner is too small to be recognized with any difficulty by the human eye in time. But the features of the pixels occupied by these small objects, which are different from their surroundings, are still captured by our GLFFNet. In the third set of images, a white car at the bottom edge is slightly obscured by these tree branches next to it, and even I am confused as to whether to mark it all as a car here or show the presence of the branch. Of course, according to the officially labeled data, there are tree branches here. So in our recognition results, there are these branches present, but of course, this is too small, and the outline is still not perfect. From the recognition results of UNetFormer, the car is accurately identified, completely ignoring the tree branches on it. In the actual scenario, there should be different specific uses for ignoring the tree branches or not. In the next two sets of images, the UNetFormer still has the problem of ignoring small objects, but, in this case, our model still does some justice to these small objects, although the segmented edges are not complete enough.

In conclusion, it can be seen from the red marks in the image that GLFFNet has better segmentation ability for small and weak objects in complex scenes in remote sensing images.

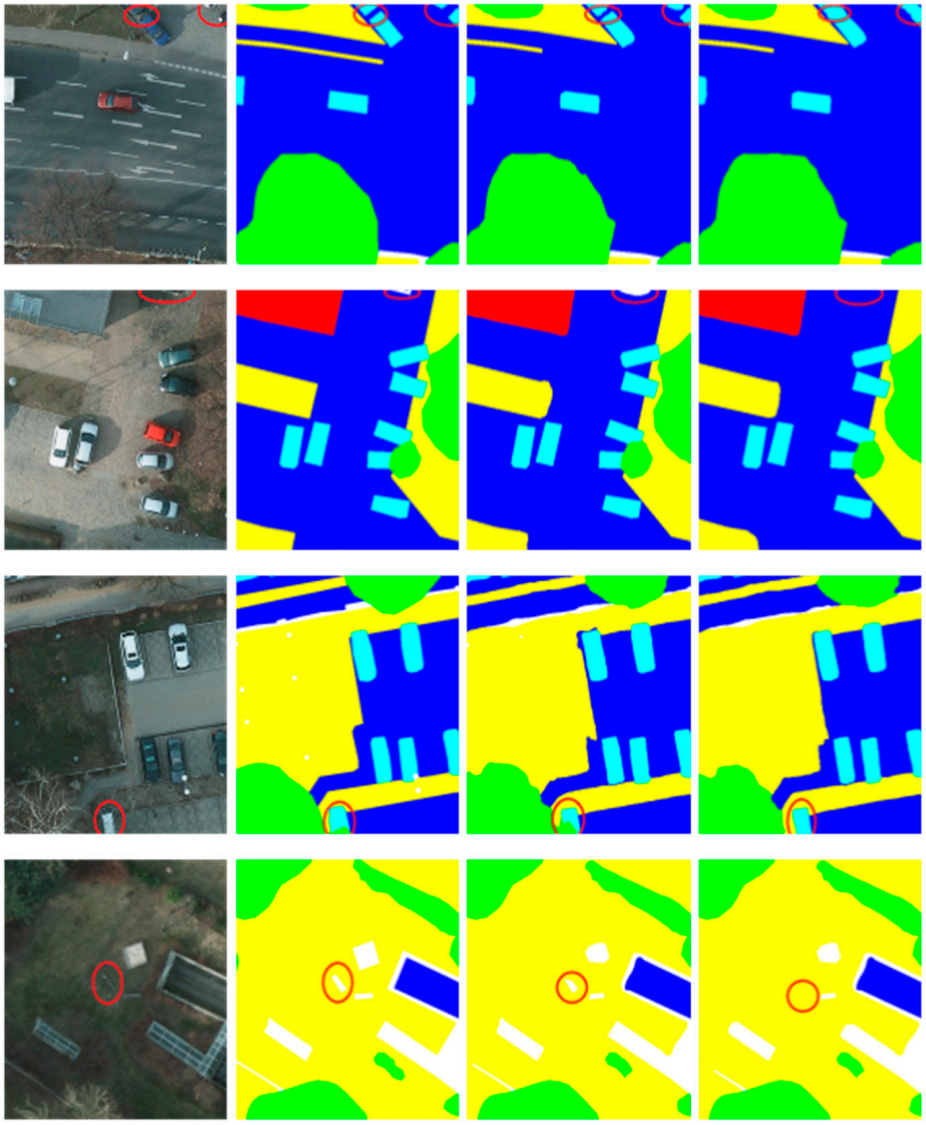

Figure 7. Partial segmentation result. The first column is the original image, the second column is labeled image, the third column is the segmentation result of GLFFNet, and the fourth column is the segmentation result of UNetFormer.

4. Discussion

4.1. The Design and Analysis of the SFA Encoder

Since AlexNet [38] came into the world in 2012, convolutional networks have dominated computer vision and related fields. A large number of models to obtain SOTA performance are designed based on convolutional networks and their variants [39–41]. However, convolution has a limited receptive field and insufficient global modeling ability, which hinders its further development in vision and related fields. Later, ViT [42] introduced Transformer, a dark horse in the NLP field in 2017, into the computer vision field. The powerful global modeling capabilities of Transformer take computer vision in a new direction, but the computing costs of the attention mechanism are a headache for scholars.

The SFA Encoder used in this paper uses recursive gated convolution and Transformer for high-order interaction to obtain global information about images, so as to deal with the problem of high within-class variability in complex scenes of remote sensing images. Compared with Transformer, the advantages of convolution are computational simplicity and the need for less data to train, while the disadvantage is that the receptive field is not large enough. The recursive gated convolution enables the global information to

interact, simulates the self-attention operation of Transformer, expands the receptive field, and obtains more global feature information. It has been experimentally proven that it is feasible to use recursive gated convolution instead of the self-attentive operation of Transformer, but the accuracy is not enough. Therefore, this paper introduces the attention mechanism of Transformer to recursive gated convolution, integrates the advantages of recursive gated convolution and Transformer, and designs the SFA encoder. As shown in Table 1, the results demonstrate that the SFA encoder outperforms encoders based mainly on recursive gated convolution, or Transformer.

4.2. The Design and Analysis of the SPAC Encoder

In addition to the above high within-class variability, another important problem in complex scenes of remote sensing images is that low between-class variability and small objects are often ignored or mis-segmented. To solve this problem, this paper designs the SPAC Encoder based on spatial pyramid architecture and atrous convolution, which is a convolution variant used by most classical semantic segmentation networks. It directly extracts the shallow feature information of a remote sensing image and preserves the local information of each pixel in the image as much as possible. In this way, the proportion of small object feature information in the final segmentation result is increased.

Compared with DeepLabV3+ and ResUNet-a, these networks all use atrous convolution as the main method to extract features. With the vigorous development of computer vision and related fields, models based on other convolutional variants or Transformer have shown better performance. In the face of advanced features in complex scenes, it seems that atrous convolution can no longer break through. However, the advantage that atrous convolution actively improves the convolutional receptive field too small should still not be ignored, and it still has an advantage in the extraction of low-order global information. Therefore, we use the spatial pyramid principle and atrous convolution to design a simple SPAC encoder to extract low-order feature information and fuse it with the high-order feature information extracted by the SFA encoder to obtain better performance. As shown in Tables 2 and 3, these ablation experiments prove that the accuracy of the segmentation of specific classes of objects is significantly improved by using the SPAC Encoder.

4.3. The Effectiveness of the Multi-Head Loss Block

In all kinds of segmentation tasks, the problem that small objects are easy to ignore has puzzled scholars. To solve this problem, multi-scale feature maps should be an effective and classical method. However, only loss is calculated for the final segmentation results during back propagation. Each encoder that extracts multi-scale feature maps cannot be accurately optimized separately in the training process, and the pre-training on the classification dataset aggravates this problem. Aiming at the above problems, this paper designs a Multi-head Loss Block to assist training. It involves the feature maps of each scale directly in the calculation of the loss, which affects the subsequent back-propagation process, so as to accurately optimize each encoder so that multi-scale feature maps can be extracted. Multi-head Loss Block enables the encoder weight of the backbone, pre-trained on classification datasets, to migrate to segmentation tasks better and faster. As shown in Table 4, experiments show that GLFFNet with a Multi-head Loss Block has a faster and better convergence process.

5. Conclusions

In this paper, a global and local feature fusion network with a biencoder (GLFFNet) is proposed in order to solve the problems of low edge accuracy, small segmentation error, and assimilation of small objects by large objects in complex remote sensing scenes. Based on Transformer structure, atrous convolution, gated convolution, and the spatial pyramid architecture, this paper proposes SFA and SPAC encoders for multi-scale feature extraction and then fusion feature information for subsequent decoding to generate segmentation results. At the same time, in order to transfer the pre-training weight of the classification

dataset to the segmentation model better and give full play to the performance of the multi-scale structure of the semantic segmentation model, a multi-head loss block is designed to assist training. Finally, the experimental results of GLFFNet reaching 87.96% mIoU on the Potsdam dataset show that when GLFFNet is applied to the semantic segmentation task of high-resolution remote sensing, GLFFNe achieves higher accuracy than most of the semantic segmentation models released so far. Therefore, this paper believes that GLFFNet has its rationality and necessity in the existing semantic segmentation models in the field of remote sensing. In the future, we will continue to explore new ways to integrate convolution and Transformer and continue to optimize the model to further improve the overall training cost and interface time of the model so that it can be applied in important fields such as environmental monitoring.

Author Contributions: Conceptualization, Q.T., F.Z. and Z.Z.; methodology, Z.Z. and F.Z.; software, F.Z.; validation, Q.T. and F.Z.; formal analysis, F.Z. and H.Q.; investigation, Z.Z.; resources, Z.Z. and F.Z.; data curation, Z.Z. and F.Z.; visualization, F.Z.; writing—original draft preparation, F.Z.; writing—review and editing, Z.Z. and F.Z.; supervision, Z.Z.; project administration, Z.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research is funded by National key research and development program of China (2020YFB1600702).

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Yuan, X.; Shi, J.; Gu, L. A Review of Deep Learning Methods for Semantic Segmentation of Remote Sensing Imagery. *Expert Syst. Appl.* **2021**, *169*, 114417. [CrossRef]
2. Diakogiannis, F.I.; Waldner, F.; Caccetta, P.; Wu, C. ResUNet-a: A Deep Learning Framework for Semantic Segmentation of Remotely Sensed Data. *ISPRS J. Photogramm. Remote Sens.* **2020**, *162*, 94–114. [CrossRef]
3. 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.
4. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. In Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.
5. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef] [PubMed]
6. Chen, L.-C.; Zhu, Y.; Wang, H.; Dabagia, M.; Cheng, B.; Li, Y.; Liu, S.; Adam, H.; Yuille, A.L. DeepLab2: A TensorFlow Library for Deep Labeling. *arXiv* **2021**, arXiv:2106.09748.
7. Chen, L.-C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking Atrous Convolution for Semantic Image Segmentation. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2502–2511.
8. 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. 833–851.
9. Xu, L.; Liu, Y.; Yang, P.; Chen, H.; Zhang, H.; Wang, D.; Zhang, X. HA U-Net: Improved Model for Building Extraction From High Resolution Remote Sensing Imagery. *IEEE Access* **2021**, *9*, 101972–101984. [CrossRef]
10. Liu, R.; Tao, F.; Liu, X.; Na, J.; Leng, H.; Wu, J.; Zhou, T. RAANet: A Residual ASPP with Attention Framework for Semantic Segmentation of High-Resolution Remote Sensing Images. *Remote Sens.* **2022**, *14*, 3109. [CrossRef]
11. Chen, Z.; Li, D.; Fan, W.; Guan, H.; Wang, C.; Li, J. Self-Attention in Reconstruction Bias U-Net for Semantic Segmentation of Building Rooftops in Optical Remote Sensing Images. *Remote Sens.* **2021**, *13*, 2524. [CrossRef]
12. Huang, L.; Zhu, J.; Qiu, M.; Li, X.; Zhu, S. CA-BASNet: A Building Extraction Network in High Spatial Resolution Remote Sensing Images. *Sustainability* **2022**, *14*, 11633. [CrossRef]
13. Zhang, Z.; Xu, Z.; Liu, C.; Tian, Q.; Wang, Y. Cloudformer: Supplementary Aggregation Feature and Mask-Classification Network for Cloud Detection. *Appl. Sci.* **2022**, *12*, 3221. [CrossRef]
14. Zhang, Z.; Xu, Z.; Liu, C.; Tian, Q.; Zhou, Y. Cloudformer V2: Set Prior Prediction and Binary Mask Weighted Network for Cloud Detection. *Mathematics* **2022**, *10*, 2710. [CrossRef]

15. Zhang, Z.; Miao, C.; Liu, C.; Tian, Q.; Zhou, Y. HA-RoadFormer: Hybrid Attention Transformer with Multi-Branch for Large-Scale High-Resolution Dense Road Segmentation. *Mathematics* **2022**, *10*, 1915. [CrossRef]
16. Ziaee, A.; Dehbozorgi, R.; Döller, M. A Novel Adaptive Deep Network for Building Footprint Segmentation. *arXiv* **2021**, arXiv:2103.00286.
17. Chen, M.; Wu, J.; Liu, L.; Zhao, W.; Tian, F.; Shen, Q.; Zhao, B.; Du, R. DR-Net: An Improved Network for Building Extraction from High Resolution Remote Sensing Image. *Remote Sens.* **2021**, *13*, 294. [CrossRef]
18. Yang, X.; Li, S.; Chen, Z.; Chanussot, J.; Jia, X.; Zhang, B.; Li, B.; Chen, P. An Attention-Fused Network for Semantic Segmentation of Very-High-Resolution Remote Sensing Imagery. *ISPRS J. Photogramm. Remote Sens.* **2021**, *177*, 238–262. [CrossRef]
19. Wang, L.; Li, R.; Zhang, C.; Fang, S.; Duan, C.; Meng, X.; Atkinson, P.M. UNetFormer: A UNet-like Transformer for Efficient Semantic Segmentation of Remote Sensing Urban Scene Imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *190*, 196–214. [CrossRef]
20. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In Proceedings of the Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
21. Lu, Y.; Wu, J.; Shen, C.; van den Hengel, A. Gated Convolutional Networks with Hybrid Connectivity for Image Classification. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), New York, NY, USA, 7–12 February 2020; pp. 12241–12248.
22. Rao, Y.; Lu, J.; Zhou, J.; Tian, Q. HorNet: Efficient High-Order Spatial Interactions with Recursive Gated Convolutions. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual Event, 25 April–1 May 2022; pp. 1–16.
23. Lin, M.; Chen, Q.; Yan, S. Network In Network. In Proceedings of the International Conference on Learning Representations (ICLR), Banff, AB, Canada, 14–16 April 2014; pp. 1–10.
24. Fukui, H.; Hirakawa, T.; Yamashita, T.; Fujiyoshi, H. Attention Branch Network: Learning of Attention Mechanism for Visual Explanation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 10705–10714.
25. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In Proceedings of the International Conference on Machine Learning (ICML), Lille, France, 6–11 July 2015; pp. 448–456.
26. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
27. Cheng, B.; Schwing, A.G. Per-Pixel Classification is Not All You Need for Semantic Segmentation. In Proceedings of the Advances in Neural Information Processing Systems 34 (NeurIPS 2021), Virtual Event, 6–14 December 2021.
28. Song, Y.; Yan, H. Image Segmentation Algorithms Overview. *arXiv* **2017**, arXiv:1707.02051.
29. Thoma, M. A Survey of Semantic Segmentation. *arXiv* **2016**, arXiv:1602.06541.
30. Cheng, J.; Li, H.; Li, D.; Hua, S.; Sheng, V.S. A Survey on Image Semantic Segmentation Using Deep Learning Techniques. *Comput. Mater. Contin.* **2023**, *74*, 1941–1957. [CrossRef]
31. Chen, X.; Ding, M.; Wang, X.; Xin, Y.; Mo, S.; Wang, Y.; Wang, J. Context Autoencoder for Self-Supervised Representation Learning. *arXiv* **2022**, arXiv:2202.03026.
32. Liu, Y.; Chen, H.; Shen, C.; He, T.; Jin, L.; Wang, L. ABCNet: Real-time Scene Text Spotting with Adaptive Bezier-Curve Network. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 3035–3042.
33. Wang, L.; Li, R.; Wang, D.; Duan, C.; Wang, T.; Meng, X. Transformer Meets Convolution: A Bilateral Awareness Network for Semantic Segmentation of Very Fine Resolution Urban Scene Images. *Remote Sens.* **2021**, *13*, 3065. [CrossRef]
34. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. *arXiv* **2021**, arXiv:2105.15203.
35. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980.
36. Cheng, B.; Misra, I.; Schwing, A.G.; Kirillov, A.; Girdhar, R. Masked-attention Mask Transformer for Universal Image Segmentation. *arXiv* **2021**, arXiv:2112.01527.
37. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *arXiv* **2021**, arXiv:2103.14030.
38. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
39. Zhu, C.; He, Y.; Savvides, M. Crafting GBD-Net for Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2109–2123.
40. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
41. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
42. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual Event, Austria, 3–7 May 2021; pp. 1–23.