

Article

GEA-MSNet: A Novel Model for Segmenting Remote Sensing Images of Lakes Based on the Global Efficient Attention Module and Multi-Scale Feature Extraction

Qiyang Li ^{1,†}, Zhi Weng ^{1,2,†} , Zhiqiang Zheng ^{1,2,*} and Lixin Wang ^{2,3,*}

¹ School of Electronic Information Engineering, Inner Mongolia University, Hohhot 010021, China; 32256006@mail.imu.edu.cn (Q.L.); wzhi@imu.edu.cn (Z.W.)

² Collaborative Innovation Center for Grassland Ecological Security, Ministry of Education of China, Hohhot 010021, China

³ School of Ecology and Environment, Inner Mongolia University, Hohhot 010021, China

* Correspondence: zqzheng@imu.edu.cn (Z.Z.); lx_wimu@imu.edu.cn (L.W.)

† These authors contributed equally to this work.

Abstract: The decrease in lake area has garnered significant attention within the global ecological community, prompting extensive research in remote sensing and computer vision to accurately segment lake areas from satellite images. However, existing image segmentation models suffer from poor generalization performance, the imprecise depiction of water body edges, and the inadequate inclusion of water body segmentation information. To address these limitations and improve the accuracy of water body segmentation in remote sensing images, we propose a novel GEA-MSNet segmentation model. Our model incorporates a global efficient attention module (GEA) and multi-scale feature fusion to enhance the precision of water body delineation. By emphasizing global semantic information, our GEA-MSNet effectively learns image features from remote sensing data, enabling the accurate detection and segmentation of water bodies. This study makes three key contributions: firstly, we introduce the GEA module within the encode framework to aggregate shallow feature semantics for the improved classification accuracy of lake pixels; secondly, we employ a multi-scale feature fusion structure during decoding to expand the acceptance domain for feature extraction while prioritizing water body features in images; thirdly, extensive experiments are conducted on both scene classification datasets and Tibetan Plateau lake datasets with ablation experiments validating the effectiveness of our proposed GEA module and multi-scale feature fusion structure. Ultimately, our GEA-MSNet model demonstrates exceptional performance across multiple datasets with an average intersection ratio union (mIoU) improved to 75.49%, recall enhanced to 83.79%, pixel accuracy (PA) reaching 90.21%, and the f1-score significantly elevated to 83.25%.

Keywords: remote sensing image; lake segmentation; neural network; attention mechanism; feature fusion



Citation: Li, Q.; Weng, Z.; Zheng, Z.; Wang, L. GEA-MSNet: A Novel Model for Segmenting Remote Sensing Images of Lakes Based on the Global Efficient Attention Module and Multi-Scale Feature Extraction. *Appl. Sci.* **2024**, *14*, 2144. <https://doi.org/10.3390/app14052144>

Academic Editor: John Dodson

Received: 23 January 2024

Revised: 25 February 2024

Accepted: 3 March 2024

Published: 4 March 2024



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

1. Introduction

Lakes are one of the carriers to measure the ecological environment of a region. The change in lake area can reflect the regional hydrology and climate changes [1]. It has long been an important demand for ecological and environmental protection workers to obtain satellite images of the lake research area through remote sensing methods and accurately segment the water body region [2].

In recent years, the extraction methods of water body in remote sensing images include band combination [1], computer vision method, threshold method, and band index method [3]. Such methods use the difference of reflected bands of water body in remote sensing images as the standard for segmentation. For example, the normalized difference water index (NDWI) is inspired by the sensitivity of plants to red bands and infrared bands,

the difference of green bands and near infrared bands are used to normalize the extraction index of water body [4]. Later, in order to adapt to the task of water body segmentation in cities, Xu [5] proposed the modified normalized difference water index (MNDWI) for investigating water extraction from remote sensing images. However, using the difference of object reflected bands as the basis for segmentation will cause the situation of “different objects with the same spectrum” and “different spectrum of the same object”. Therefore, designing a more accurate method for lake extraction is still an urgent task to be solved [6].

With the development of the neural network, semantic segmentation algorithm has become a hot research topic in the task of remote sensing image water extraction [7]. In the development of semantic segmentation algorithm, fully convolutional network (FCN) is a milestone research [8]. The proposed fully connected structure makes convolutional neural network (CNN) better applied to feature extraction of images, and classifies semantic segmentation tasks accurately to the pixel level. In 2015, U-Net algorithm was designed to achieve the semantic segmentation tasks of medical images. The pioneering idea of U-Net model is the designed encoder and decoder structure. However, the limitation of the U-Net model is that the background information and target information are transmitted at the same time after feature extraction, which hampers the model’s learning of image features [9]. In the same year, in order to solve the complexity caused by the model’s continuous stacking of convolution features, He Kaiming’s team [10] proposed the residual network framework, which uses the idea of a residual function to change the depth of simple convolution replacement, greatly reducing the calculation of the model and becoming an efficient network model. In 2016, Zhao developed pyramid scene parsing network (PSP-Net) for semantic segmentation tasks across diverse scenarios and introduced the pyramid pooling module [11] to enhance feature information from various extraction layers while expanding the network’s receptive field. In 2017, Ashish Vaswani introduced the transformer network architecture for natural language processing (NLP) field. The concept of attention emerged as a pivotal component in the development of transformer [12], and subsequently became a significant model for computer vision tasks. In 2018, Chen’s [13] team extended the DeepLabV3 model by incorporating deep separable convolution into a pyramid pooling operation. They further devised the DeepLabV3+ model to refine the boundaries of target objects in semantic segmentation, which has gained recognition as an established semantic segmentation approach. Oktay et al. [14] integrated attention-gates into the U-Net model, marking its pioneering utilization of attention mechanism to enhance learning capability.

In the field of water extraction, numerous scholars have optimized and enhanced existing models to improve the accuracy of segmentation for water extraction. For instance, Zhong et al. [15] incorporated depthwise separable convolutions in the encoder to enhance the capacity for extracting global information from remote sensing images when identifying lake water features. Li et al. [16] devised a dense network for precise water extraction and integrated a bidirectional channel attention mechanism to mitigate noise interference during lake analysis. Xin [17] proposed a multi-scale fusion attention network multi-scale successive attention fusion network to solve the problem of poor accuracy in narrow boundary in water extraction, and realized attention operations on low-level features and high-level features, respectively. Luo [18] improved the DeepLabV3+ model for small water extraction, and applied the idea of generative adversarial network (GAN) to generate an adversarial network to synthesize the original remote sensing image with false color to improve the features of small water bodies. Yu et al. [19] designed the network to improve the accuracy by linking the contextual information of water bodies. Weng et al. [20] improved the SegNet network with the idea of the residual network in the process of lake water extraction, and realized identity mapping to alleviate the training degradation problem in stacked networks. Additionally, Li et al. [21] reconstructed the DeepLabV3+ network model and optimized it using a fully connected conditional random field (CRF), resulting in a significant improvement in the accuracy of water body segmentation. Wang et al. [22] proposed a lightweight network for extracting water bodies, which not only enhanced edge segmentation accuracy but also reduced the parameters of the extraction

network. In order to enhance the detection of water bodies, Guo et al. [23] proposed a multi-scale model to water body extraction for integrating features across different scales. Wang et al. [24] employed a multi-scale dense connection module in the design of the multi-scale lake water extraction network for water extraction, effectively suppressing image noise and achieving the precise extraction of small water bodies. In recent years, the emergence of vision transformer [25] has proven to be successfully applied in various fields of computer vision tasks. In the noise-canceling transformer network (NT-Net) designed by Zhong [26], multi-level transformer modules were introduced into the NT-Net network to effectively capture global information related to water features and enhance its ability to distinguish lake boundaries. Furthermore, Zhang et al. [27] proposed a U-Net network that integrates transformer modules, utilizing CONV+Mixformer Block for feature extraction and incorporating an attention module for optimal water segmentation performance on the GID datasets. Zhao et al. [28] combined the swin-transformer algorithm for lake extraction on the Qinghai-Tibet Plateau lake datasets, replaced the convolution structure of the U-Net encoder with vision transformer (ViT) and added channel-wise and spatial-wise attention module (CBAM) attention mechanism in the decoder. The improved network achieved the best performance compared with other algorithms in the plateau lake extraction task.

However, there are certain limitations associated with the aforementioned methods. On the one hand, effectively utilizing attention mechanisms to enhance the model performance poses a significant challenge. Some researchers have improved model performance by directly employing existing attention mechanisms [29–31]. While the attention mechanism is widely adopted to enhance network segmentation accuracy, its effectiveness is influenced by the feature learning capability of the attention module. Therefore, the feature learning of the attention module is optimized in this study, and the global efficient attention module (GEA) is employed to enhance the extraction of water features. The GEA attention module, which we have investigated, is a novel and efficient attention mechanism that can be effectively incorporated into the algorithmic design of various image segmentation tasks to enhance the accuracy of segmentation. Moreover, the concept of efficient feature aggregation in the GEA attention module may also inspire other researchers in their pursuit of model improvement. On the other hand, the feature learning method of the water body segmentation model in up-sampling is limited, hindering the acquisition of global information and resulting in the omission of small water bodies. To address this issue, we propose a multi-scale feature fusion module to enhance the extraction of spatial semantic information for water bodies. Experimental results demonstrate that our GEA-MSNet model exhibits significant improvements in data quality and demonstrates certain generalization capabilities across different datasets. The main work of this paper is as follows:

1. The paper presents a backbone network design based on Res-UNet coding and decoding architecture. A novel GEA efficient attention module is introduced for shallow feature fusion. This module employs deep separable convolutions operations to effectively reduce the number of feature parameters. Additionally, a global response normalization layer is utilized to integrate features and enhance the learning capability of local water body information while suppressing the interference information to expedite model convergence and mitigate overfitting phenomenon. Forming an efficient feature learning attention module.
2. We introduce a multi-scale feature fusion module into the decoder component of our model to optimize the up-sampling feature transfer process. This module employs a parallel double-branch structure and combines mixed expansion convolution with ordinary convolution to enhance the receptive field of the convolution kernel, thereby integrating more contextual features for improved accuracy in lake water extraction and segmentation.
3. The model's effectiveness and results were validated through a multitude of experiments conducted on diverse datasets. To enhance the network's robustness, a GAN network was employed to generate an additional 300 image expansion datasets in the

lake scene classification datasets. Furthermore, the generalization performance of the network was assessed using an open Qinghai–Tibet Plateau lake dataset.

The remaining sections of this paper are organized as follows: The Section 2 provides an overview of relevant research, including a description of the datasets and an introduction to the theoretical framework of the model. In the Section 3, we present the experimental methodology, outline the overall model structure, and provide detailed descriptions of each designed module's specific structure. The Section 4 encompasses various experiments conducted in conjunction with different datasets and models for comparative analysis. On the Section 5, we discuss our conclusions while also conducting ablation experiments to validate each module's contribution. Finally, in the Section 6, we summarize our work comprehensively and offer prospects for future research.

2. Related Work

2.1. Datasets

The datasets used in this study comprise two parts. The first part consists of remote sensing images of lake areas from the scene classification datasets [32] proposed by Northwestern Polytechnical University in 2017. The original image has a size of 256×256 pixels and consists of RGB three-channel spectral data. The spatial resolution ranges from 30 m to 0.2 m per pixel. All images were extracted from Google Earth (Google Inc., Menlo Park, CA, USA). This part contains 700 images obtained by collecting and cropping the satellite remote sensing images from lakes worldwide. However, these datasets were originally intended for scene classification without semantic segmentation annotations. As neural network training requires substantial data support, the original sets' limited size is insufficient for model training, leading to poor performance due to inadequate data [7]. Therefore, we employed a generative adversarial network (GAN) for the generation of remote sensing lake image data to address the lack of datasets [33]. The fundamental concept behind GAN network involves designing a generative network and an adversarial network. The generator is continuously updated to effectively deceive the discriminator, while the discriminator consistently enhances its discriminatory capabilities. The application of the GAN network for generating image data has been validated in several relevant research articles [34–38]. In this study, 300 lakes are generated by the GAN network, and the remote sensing images that meet the requirements of lake segmentation are selected as the data source. The final data source contained 983 remote sensing images of lakes with a size of 256×256 . The obtained data were labeled with Labelme software (5.2.1) to mark the water body part, and the JSON format after annotation was converted into .png format. Refer to Figure 1 for the datasets and annotation of scene classification, and consult Figure 2 for the generated data and its corresponding annotation.



Figure 1. Remote sensing image and label of lake with scene classification.



Figure 2. GAN network generated data lake remote sensing image and annotation.

Additionally, we opted to validate the model's performance across various datasets by utilizing the publicly available lake datasets from the Qinghai–Tibet Plateau lake datasets, as shown in Figure 3. The Qinghai–Tibet Plateau lake datasets [24] is an open semantic segmentation datasets of lakes. The whole datasets contained 1600 lakes, and a total of 6774 remote sensing images with RGB three-channel composition were acquired, featuring a spatial resolution of 17 m and cropped to a size of 256×256 pixels per image. The datasets were labeled with a water body and background using Labelme software, which is similar to the datasets we prepared above. Therefore, these data are selected for the generalization performance test of the model.



Figure 3. Remote sensing image and annotation of the Qinghai–Tibet Plateau Lake datasets.

2.2. Encoder–Decoder Architecture

The encoding and decoding network structure is commonly employed in computer vision tasks such as object detection and semantic segmentation [39,40]. The encoder acquires semantic features by reducing image resolution and noise through the downsampling process, while the decoder restores image resolution through upsampling operations. The FCN network, based on an improved encoding and decoding structure [41], as well as the SegNet [20] network employed in the remote sensing image water body task, utilize the encoding and decoding structure to integrate feature information at various levels, ultimately generating semantic segmentation results based on predicted categories.

However, the integration of feature information in decoder sampling may potentially impact the model's learning capability in both FCN and SegNet architectures. Moreover, conventional convolutions with limited receptive fields fail to capture multi-scale contextual information, thereby diminishing the accuracy of output image features.

To address the issue of low image recovery accuracy, we propose a novel feature extraction module that is integrated throughout the model to optimize upsampling operations. This new module utilizes multi-scale feature fusion for feature recovery instead of solely focusing on image resolution restoration like traditional transpose convolution methods. By expanding convolution and enhancing global information understanding, multi-scale feature fusion mitigates the limitations imposed by limited receptive domain convolution during upsampling.

2.3. Attention Mechanism

The attention mechanism is derived from human visual features [42–45]. Several studies have demonstrated that incorporating attention modules during feature extraction can enhance network performance [46–48]. This mechanism encompasses lightweight modules that focus on feature space and channels, as well as self-attention mechanisms involving constant calculations of query (Q), key (K), and value (V) matrices [42]. For instance, in the Attention-Unet network [49], the authors introduced an Attention Gate module to the feature fusion of skip connections, resulting in improved segmentation accuracy. However, with the continuous advancement of attention mechanism, although the novel attention modules can enhance the segmentation accuracy in models, they also introduce a significant increase in parameter count. In their study, Song et al. [50] explored the integration of various attention models into the U-Net network and conducted experiments. While achieving an improvement of 4~5% in segmentation accuracy, the augmented number of parameters due to attention models emerged as a new challenge, thereby directing their future efforts towards further enhancements.

To address the issue of increasing model parameters due to attention, we propose a lightweight and efficient attention module for water body extraction models. This module utilizes a global normalized response (GRN) operation to fuse the spatial feature information and significantly reduces the parameter count through depth separable convolution. Unlike previous attention modules that directly process features, our approach enables learning of attention based on the fusion of feature heights, thereby enhancing image feature learning efficiency. In terms of performance, our proposed attention module demonstrates excellent segmentation capabilities in detecting small water bodies and lake edges in remote sensing images.

3. The Proposed Method

The present section of the paper provides a detailed exposition on the comprehensive structure and constituents of the GEA-MSNet model for remote sensing image lake extraction. Firstly, an overview of the model framework is presented, followed by individual subsections that introduce enhanced structures aimed at improving segmentation accuracy.

3.1. Overall Structure

The structure of GEA-MSNet is shown in Figure 4. GEA-MSNet is named by the added GEA attention module and is based on the improvement of the Res-UNet [51] model. This model encompasses an encoding and decoding structure based on residual networks, a GEA attention module, and a multi-scale feature fusion module. Firstly, within this model, the encoder stage employs a 50-layer residual network structure [52] for efficient feature extraction. The features of different levels are correspondingly blended, with the inclusion of shallow features in advanced ones to facilitate information restoration at a detailed level. Secondly, the GEA attention module is devised to establish connections between semantic context features while prioritizing global information. The aggregation of global feature information is increased to efficiently extract image features to optimize the learning ability of the model. Thirdly, the multi-scale feature fusion module is used in the decoder to restore the resolution of the image layer by layer. The multi-scale feature fusion module addresses the limitation of convolution operation by mitigating overfitting caused by solely focusing on local features. It enhances the recovery of water body edge information and improves the detection of small-scale spatial details, thereby alleviating these issues.

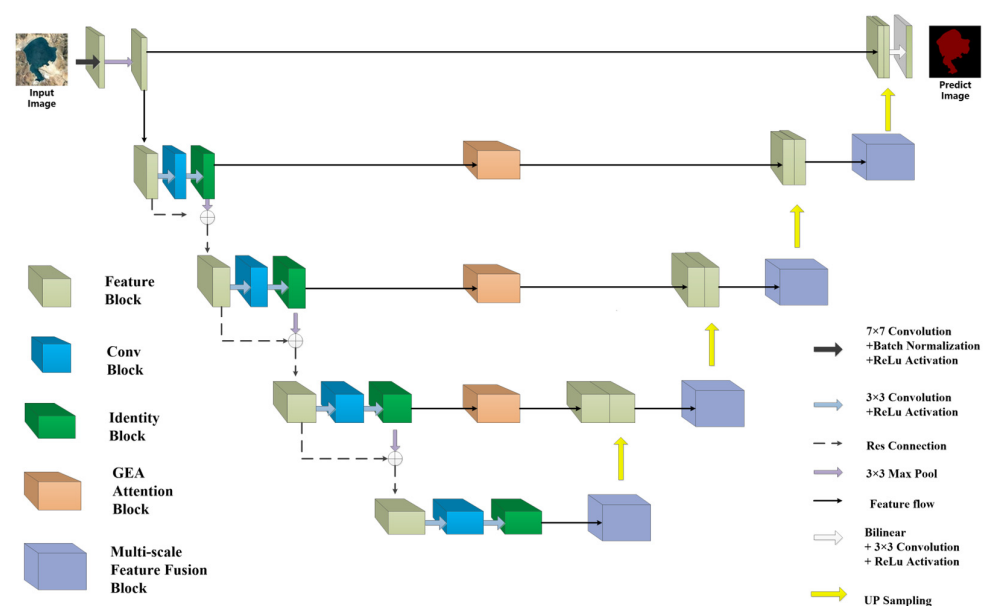


Figure 4. Overall structure of the GEA-MSNet network.

3.2. Residual Network Structure

The structure of the residual network is derived from the residual function, facilitating the deep network's acquisition of identity mapping for shallow features through residuals [51,53,54]. This effectively addresses the issues of gradient vanishing and model degradation caused by the continuous stacking of network depth. The residual network structure primarily consists of convolutional blocks and identity blocks. As depicted in Figure 5, a convolutional block comprises three 3×3 convolution modules, a normalization layer, and an activation function. Additionally, a convolutional and normalization operation is incorporated within the skip connection to facilitate the dimensionality transformation of the feature map in the lake image. The identity block is employed to enhance the network depth, reduce the image resolution while acquiring deep features, and perform the initial feature segmentation of the target area and noise region in the lake. In our model, we adopted a ResNet-50 architecture with a configuration of $3 + 6 + 4 + 3 + 1$. Each module consists of a Conv Block layer and two Identity Block layers, with the final layer being a fully connected layer that is employed to extract the features. The four output features correspond to those of the encoder in the overall GEA-MSNet framework.

3.3. The GEA Attention Module

The GEA attention module, serves as an attention mechanism to enhance the extraction of water features and refine the segmentation effect of the model. In the encoding process, downsampling is employed to reduce resolution and obtain semantic features at different levels. However, these shallow features possess rich spatial information while lacking in semantic details [55,56], leading to potential loss or the misclassification of a significant number of water bodies.

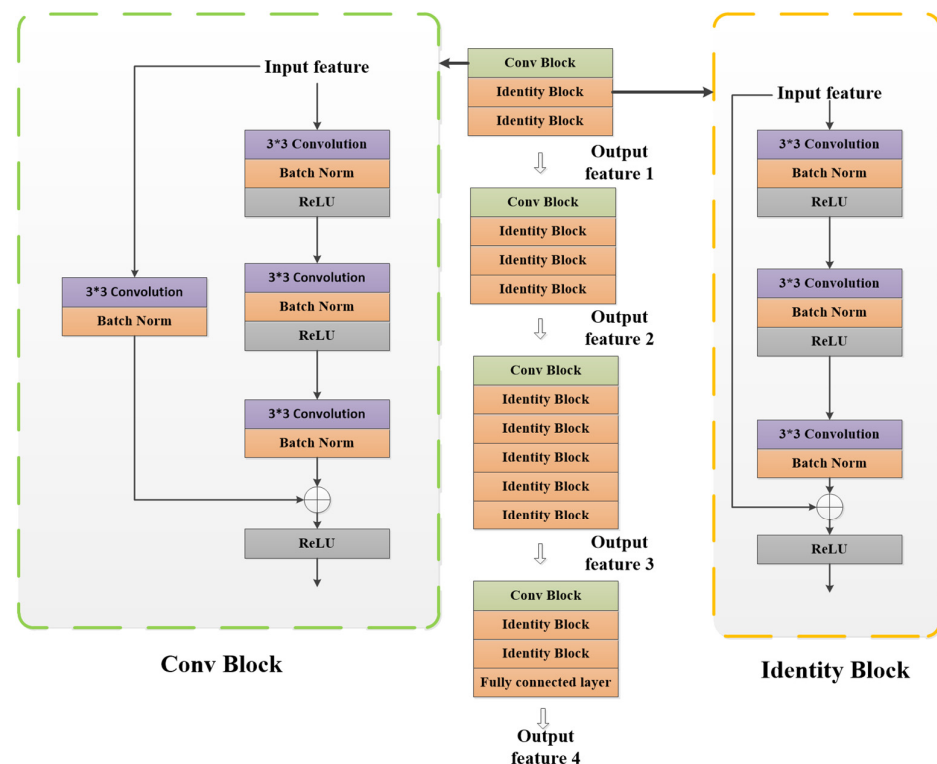


Figure 5. Residual network structure.

In the GEA module, the spatial features are first extracted by depth convolution. The depth convolution uses a convolution of 3×3 size and the number of groups is the same as the feature dimension. The proposed method performs convolution operations on each layer of channels, emphasizing the global semantic information within the channels.

Point convolution is employed using a 1×1 kernel to effectively fuse features across the channel vector. The depth separable convolutions not only captures the features of individual channels but also extracts inter-channel information within the same spatial context. After applying the Gaussian error linear unit (GELU) activation function, we employ the global response normalization (GRN) module to effectively mitigate noise and eliminate the redundant information in the image. The GRN module performs aggregation, normalization, and the calibration of global features, thereby enhancing the efficiency of feature learning. The formula is as follows:

$$X \in R^{H \times W \times C} \quad (1)$$

where R is the original input image, H and W are the height and width of the input image, respectively, C is the number of channels in the image, X represents the feature obtained from the image. The entire spatial feature is aggregated by the g function to obtain a compressed feature vector. The expression of g function is as follows:

$$g(X) = gX = \{\|X_1\| + \|X_2\| + \dots + \|X_i\|\} \in R^C \quad (2)$$

$$X \in R^{H \times W \times C} \rightarrow gX \in R^C \quad (3)$$

where g in Equations (2) and (3) represents the feature vector after the aggregation of the input X . After obtaining the feature expression X , we propose g to represent the high-level aggregation of features. Based on research [51], the L_2 norm of X can be used as a measure for feature aggregation by g , aggregates the feature of each channel into a set of vectors after generating a value, and the L_2 norm of X_i represents the statistical information scalar of the i channel. Next, the channel characteristics after aggregation will be normalized, with the following expression:

$$N(\|X_i\|) \approx \|X_i\| \in R \rightarrow \frac{\|X_i\|}{\sum_{j=1, \dots, C} \|X_j\|} \in R \quad (4)$$

where N in Equation (4) denotes the normalization operation performed on the i channel, while the fractional formula represents a means to calculate the relative significance of other channels in achieving this normalization operation by considering their influence on the current channel. Finally, it is necessary to calibrate the original input response by employing the computed feature-normalized score, and its specific expression is provided as follows:

$$X_i = \gamma \times X_i \times N(g(X)_i) + \beta + X_i \quad (5)$$

In Formula (5), to streamline the optimization process, we introduced two additional learnable parameters, γ and β , which were initialized to zero. Furthermore, a residual connection was incorporated between the input and output of the GRN layer. The resulting equation is presented as Equation (5). X_i represents the ultimate output, which is adjusted through the addition of hyperparameters γ and β to the residual structure. The inclusion of residuals and hyperparameters aims to enhance the corrective capacity for the initial input. The selection of the depth separable convolution and global response normalization (GRN) module enables the effective feature extraction and global normalization in consideration of both spatial and channel characteristics. This design not only accelerates model convergence but also facilitates efficient aggregation and the extraction of features by utilizing GRN for global feature normalization.

After the aforementioned module, we will perform point convolution to further enhance the extraction of feature information for the highly aggregated data. Subsequently, these features will be passed into the coordinate attention (CA) mechanism, which generates separate outputs from two directions. The spatial attention mechanism CA [57] employed here differs from self-attention in terms of feature calculation through matrices. CA focuses more on capturing spatial information and channel relationships within features [50].

The challenge in lake segmentation lies in accurately delineating water body edges and detecting small water bodies.

$$X = [x_1, x_2, \dots, x_c] \in R^{H \times W \times C} \quad (6)$$

After the previous feature aggregation, assuming that the incoming intermediate feature is denoted by R and X in Equation (6). Represents the feature tensor, this facilitates subsequent pooling operations in both the H and W directions. However, conventional global pooling compresses spatial information into channel information, thus necessitating a decomposition of the global pooling operation. The specific formula for this decomposition is provided as follows:

$$Z_c = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W x_c(i, j) \quad (7)$$

$$Z_c^h(h) = \frac{1}{W} \sum_{0 < i < W} x_c(h, i) \quad (8)$$

$$Z_c^w(w) = \frac{1}{H} \sum_{0 < j < H} x_c(j, w) \quad (9)$$

Equation (7) represents the global pooling operation of the original feature vector, whereas we employ pooling convolution with $(H, 1)$ and $(1, W)$ to perform pooling along the horizontal X direction and vertical Y direction, respectively. Subsequently, Formulas (8) and (9) yield the expressions for the obtained features. In Formula (8), Z_c^h denotes the output feature of channel c with a height h , while Z_c^w in Formula (9) represents the output feature of channel c with a height w . Following this extraction of directional features, an aggregation transformation operation is conducted as follows:

$$f = \delta(F_1[Z_c^h, Z_c^w]) \quad (10)$$

The formula incorporates the integrated features (represented by f), the nonlinear activation function (denoted by δ), and the concatenate function (indicated by F) that connects the features bidirectionally. Subsequently, the resulting output is fed into convolutional processing, which can be expressed as follows:

$$g^h = \sigma(F_h(f^h)) \quad (11)$$

$$g = \sigma(F_w(f^w)) \quad (12)$$

The features of h and w in the given formula should be separated initially, where σ represents the sigmoid function, F denotes the convolution operation, and g signifies the feature obtained in this direction. Ultimately, by connecting the residual input, we obtain the final output as follows:

$$y_c(i, j) = x_c(i, j) \times g_c^h(i) \times g_c^w(j) \quad (13)$$

The $y_c(i, j)$ in the aforementioned equation represents the ultimate output, and the aforementioned realization conducts attention learning based on both horizontal (h) and vertical (w) directions for the previously fused features. The structure diagram of GEA module is shown in Figure 6. The overarching concept of this model is to achieve feature enhancement extraction while minimizing parameter increase and computational complexity. Consequently, the GEA module consists of depth-wise separable convolution combined with global normalized response, complemented by CA feature learning. The final experimental results substantiate that this model enhances accuracy without an excessive surge in parameter quantity.

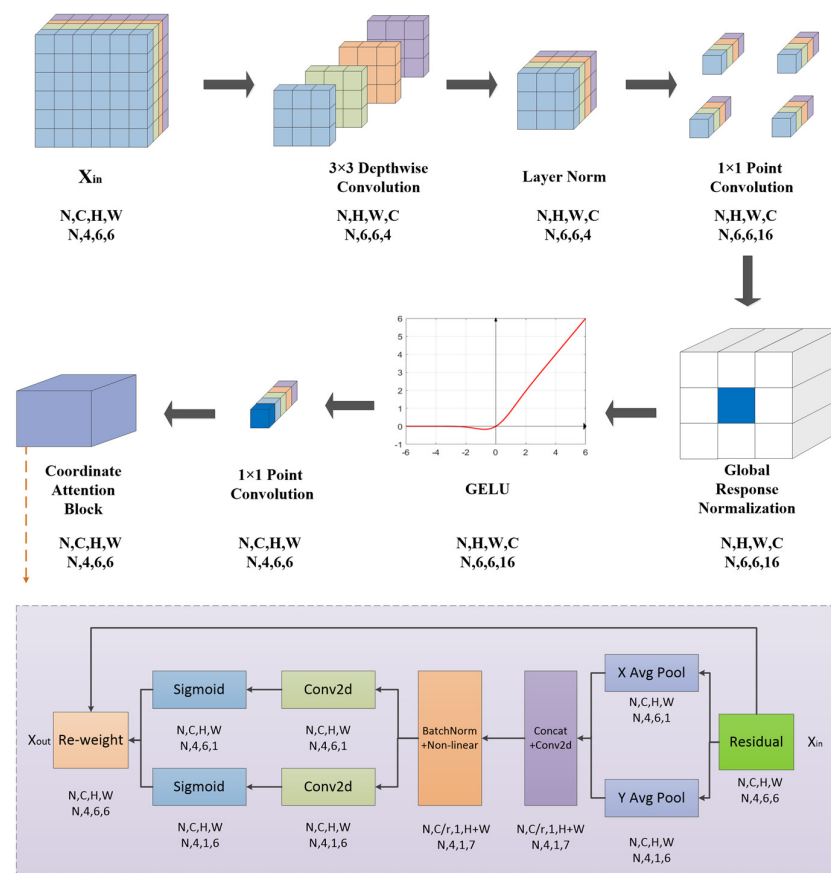


Figure 6. The GEA attention module.

3.4. Multi-Scale Feature Fusion Module

The upsampling process incorporates a multi-scale feature fusion module to restore the image resolution, which combines the mixed expansion convolution [58] and ordinary convolution in a layer-by-layer manner. This approach ensures that the resulting images have consistent resolution with the input. Unlike traditional upsampling methods that rely on transposed convolution for extracting deep information, our proposed method overcomes the limitation of focusing solely on local information by incorporating global information. The structure diagram of the multi-scale feature fusion module is illustrated in Figure 7. To enhance the receptive field of feature learning and augment the attention towards global water body information in feature mapping, a hybrid dilation convolution with a two-layer dilation ratio (2 and 3) and a convolution core size of 3×3 is employed. In contrast, the traditional 3×3 convolution is employed as a complementary approach to prevent the omission of adjacent information in feature learning by expansion convolution. This design offers a more comprehensive integration of the feature information acquired through both expansion and ordinary convolutions, thereby facilitating the detection and recovery of small water bodies. We propose a dual-branch parallel structure for extracting the semantic information from lakes. The dual-branch architecture ensures the exponential growth of learned information without compromising global or local details. Finally, normalized output is obtained by adjusting channel numbers using 1×1 convolution and incorporating BN layers.

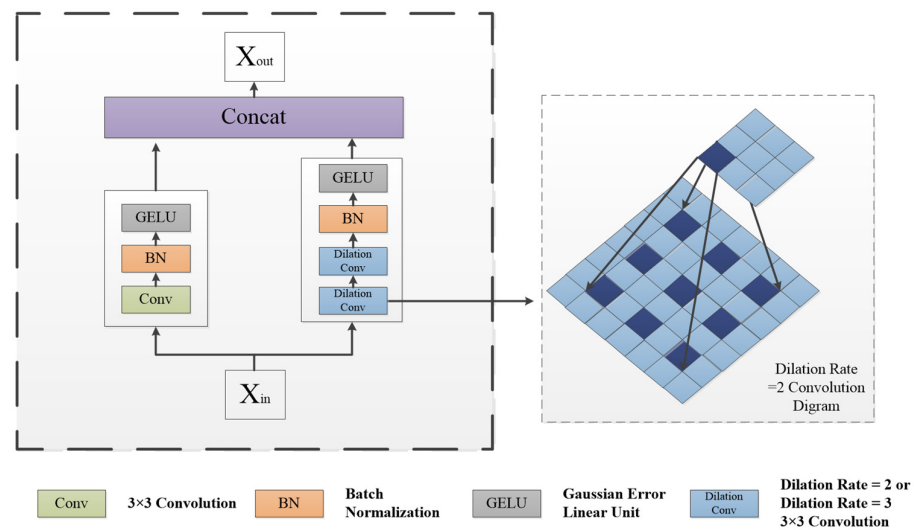


Figure 7. Multi-scale feature fusion structure.

4. Experiments

4.1. Experimental Details

The experimental environment was configured on a Windows 10 system, utilizing an NVIDIA GeForce RTX 3070 GPU (NVIDIA, Santa Clara, CA, USA). The training model was developed using a PyTorch 1.11.0+cu115 framework. Each network underwent training for 100 epochs with a batch size of 4. The loss function employed was binary cross-entropy (BCE_Loss), and the optimizer used was either Adam or SGD algorithms. For Adam, the initial weight value was set to 1×10^{-2} , while for SGD it was set as 1×10^{-4} .

4.2. Evaluation Metrics

The evaluation methods employed in this study encompass commonly utilized semantic segmentation indices, namely the average intersection overlap ratio (*mIoU*), recall rate (*Recall*), pixel accuracy (*PA*), and *F1-score*. The specific formulas for these metrics are presented as follows:

$$mIoU = \frac{1}{k+1} \sum_{i=0}^k \frac{TP}{FN + TP + FP} \quad (14)$$

$$Recall = \frac{TP}{TP + FN} \quad (15)$$

$$PA = \frac{TP}{TP + FP} \quad (16)$$

$$F1 - score = \frac{2TP}{2TP + FP + FN} \quad (17)$$

The *mIoU* in Formula (14) represents the average cross-ratio between the water body and background in the predicted results and ground truth values. The recall in Formula (15) denotes the ratio of pixels correctly classified as water bodies to the total number of actual water body pixels. The *PA* in Formula (16) signifies the proportion of true water bodies among all pixels classified as water bodies. Lastly, the *F1-score* in Formula (17) incorporates both accuracy and recall measures, serving as an averaged metric for evaluating the classification model performance. In the formula, *TP* represents the number of pixels correctly classified, *FP* represents the number of pixels wrongly classified, *TN* represents the number of pixels correctly classified as background, and *FN* represents the number of pixels wrongly classified as background. In addition, the size of the parameter is also introduced into the evaluation as a reference index. Assuming a red–green–blue three-channel original lake image is used as

input to calculate the parameter quantity generated by the model, the parameter quantity of the model can reflect the efficacy of the designed efficient attention module, thereby reducing the overall network parameters while improving the accuracy.

4.3. Experimental Results

The experiments were conducted on two distinct datasets, and the selection of comparative experiments was divided into two categories. In the initial experiment, we utilize the scene classification dataset introduced in Section 2 of this paper. Firstly, we employ well-established and widely used models based on relevant research articles pertaining to semantic segmentation. These models were chosen not only as comparative benchmarks for our lake classification task but also due to their frequent appearance in other studies related to semantic segmentation, such as U-Net, FCN8, PSP-Net, and DeepLabV3+. Secondly, as our model primarily focuses on the construction of the U-Net network, we incorporated the latest lightweight attention module in semantic segmentation to compare experimental results. These include Attention-U-Net, as well as U-Net with attention modules such as CBAM (convolutional block attention module), SE (squeeze-and-excitation) module, and CA (coordinated attention) module. Additionally, During the training of the initial five models in the enumerated list, pre-training weights were employed for transfer learning. The final experimental results are presented in Table 1, demonstrating that our designed GEA-MSNet model outperforms other models tested in this study across various indicators: mIoU (75.49%), recall (83.79%), PA (90.21%), and F1-score (83.25%). These findings suggest that the GEA-MSNet model is exceptionally effective for lake segmentation tasks.

Table 1. Experimental results of different models in the lake scene classification datasets.

Method	mIoU (%)	Recall (%)	PA (%)	F1-Score (%)	Parameters (M)
U-Net	63.05	75.07	83.55	76.63	34.53
FCN8	61.35	72.57	83.02	73.16	30.12
PSP-Net	64.86	76.57	84.51	76.96	49.07
DeepLabV3+	66.24	77.27	85.28	77.89	54.71
Attention-U-Net	63.50	75.66	83.65	75.95	37.19
CBAM-U-Net	63.13	74.26	84.02	76.27	34.57
SE-U-Net	65.28	76.61	84.89	77.59	34.57
CA-U-Net	66.17	77.84	85.15	77.69	34.56
GEA-MSNet (ours)	75.49	83.79	90.21	83.25	39.60

In addition, although the number of parameters increased compared with U-Net, FCN network and U-Net adding different lightweight attention modules, but the segmentation effect produced by these models was general. Compared with DeepLabV3+ and PSP-Net models with a good segmentation effect, the number of parameters in GEA-MSNet was not only smaller than the above models but also produced a better segmentation effect. Overall, the model improved the accuracy by approximately 5% compared with other methods without too much increase in the number of parameters, which was due to the design of depthwise separable convolution and global normalization in the efficient attention module GEA, which had less parameters while fusing the spatial information and global information.

The visualization effects of different models on the lake scene classification datasets are presented in Figure 8. In this study, a subset of images containing small water bodies and complex edges from the datasets were selected for comparison to evaluate the performance of various models. Furthermore, it is observed that U-Net, FCN8, PSP-Net, and Attention-UNet models exhibit relatively low overall accuracy with indistinct segmentation boundaries in the target area. Some small water bodies are either missing or misclassified. However, despite the incorporation of various attention modules (SE, CBAM, Los Angeles, CA, USA) into U-Net to enhance the accuracy of water body edge extraction, the issue of indistinct boundaries still persists. Finally, we propose that the GEA-MSNet model effectively refines the edge features and accurately extracts micro water bodies while effectively delineating the transitional zone between land and lake surfaces. This approach mitigates misclassification caused by variations in water color and enhances the overall precision in lake extraction.

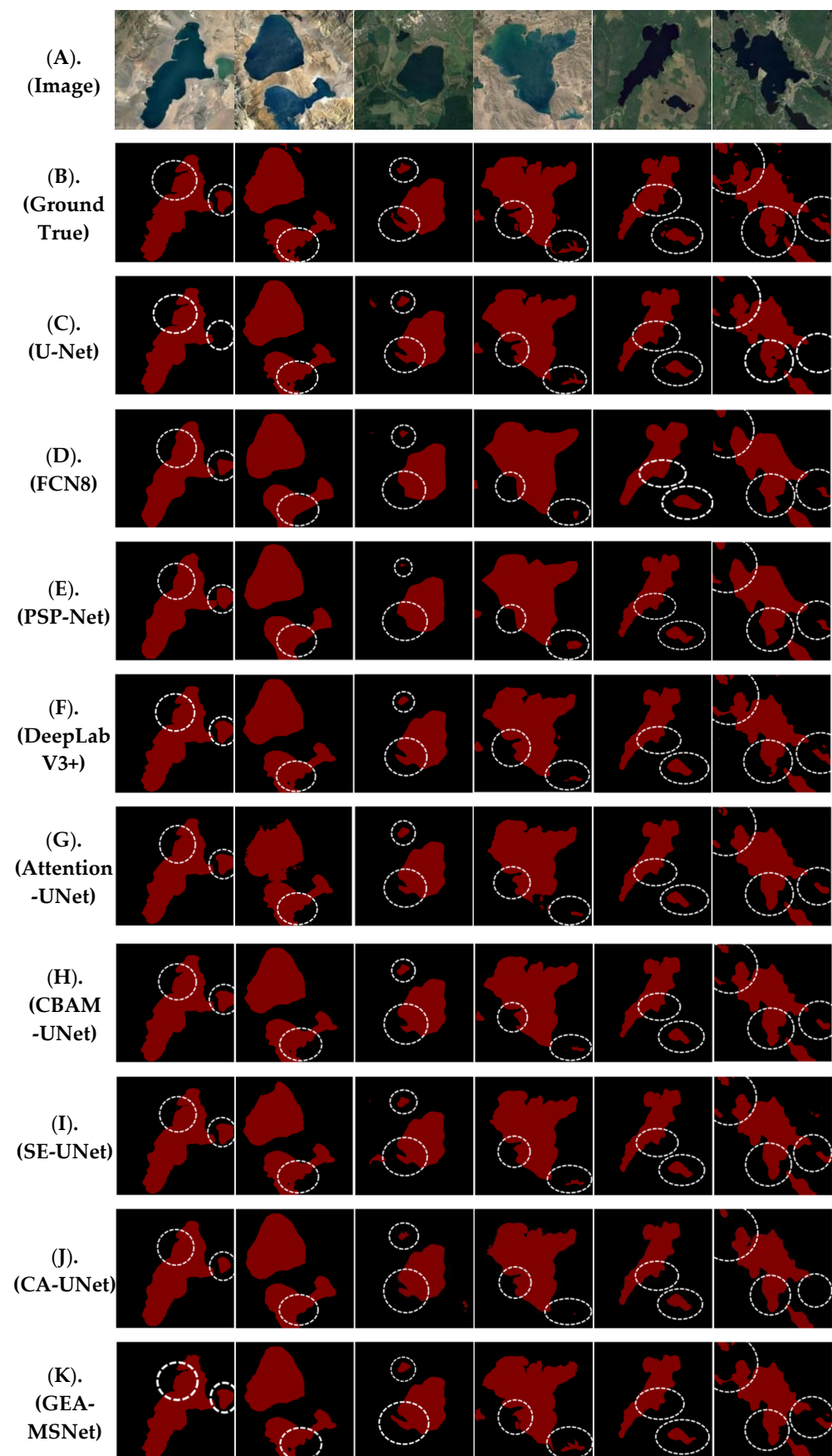


Figure 8. Visualization results showing different models on a lake scene classification datasets: (A) image; (B) basic fact; (C) U-Net results; (D) FCN8 results; (E) PSP-Net results; (F) DeepLabV3+

results; (G) Attention-UNet results; (H) CBAM-UNet results; (I) SE-UNet results; (J) CA-UNet results; and (K) GEA-MSNet results. The black area in the prediction results is the background, and the red area is the target water body. The white circle area represents the focus area where the prediction occurs.

Furthermore, to ensure the model's generalization performance, the generalization of the model was tested using the lake dataset from the Tibetan Plateau. For this test, we selected certain models from the initial experiment and conducted additional comparative experiments by including both classical models and attention models with higher accuracy identified in Experiment 1. Although the datasets exhibit a high overall image brightness, which may lead to confusion between clouds and lakes, their annotations are accurate, making them an ideal choice for evaluating the segmentation performance of our model. The experimental data demonstrate that our model performance remains optimal. However, owing to the extensive dataset and precise annotations incorporated in this dataset, the experiments exhibit an elevated level of segmentation accuracy. The segmentation outcomes for the Tibetan Plateau lake datasets are presented in Table 2.

Table 2. The experimental results of various models applied to the datasets of lakes in the Tibetan Plateau.

Method	mIoU (%)	Recall (%)	PA (%)	F1-Score (%)
U-net	95.51	97.67	97.79	96.76
DeepLabV3+	96.79	98.40	98.42	97.64
CBAM-U-Net	95.57	97.65	97.83	97.40
GEA-MSNet (Ours)	97.57	98.86	98.82	98.56

According to Table 2, we conducted a comparative analysis of the evaluation metrics for different network architectures using the Qinghai–Tibet Plateau lake datasets. It is noteworthy that the GEA-MSNet model exhibited a superior semantic segmentation accuracy on these datasets, with an increase of 0.24% in pixel accuracy compared to the top-performing DeepLabV3+ model, as well as improvements of 0.46% in mIoU and 0.2% in Recall. According to the experimental results, even if other remote sensing lake segmentation datasets were replaced, our model also achieved the best segmentation accuracy, and the model has good generalization performance and visualization results can be observed in Figure 9.

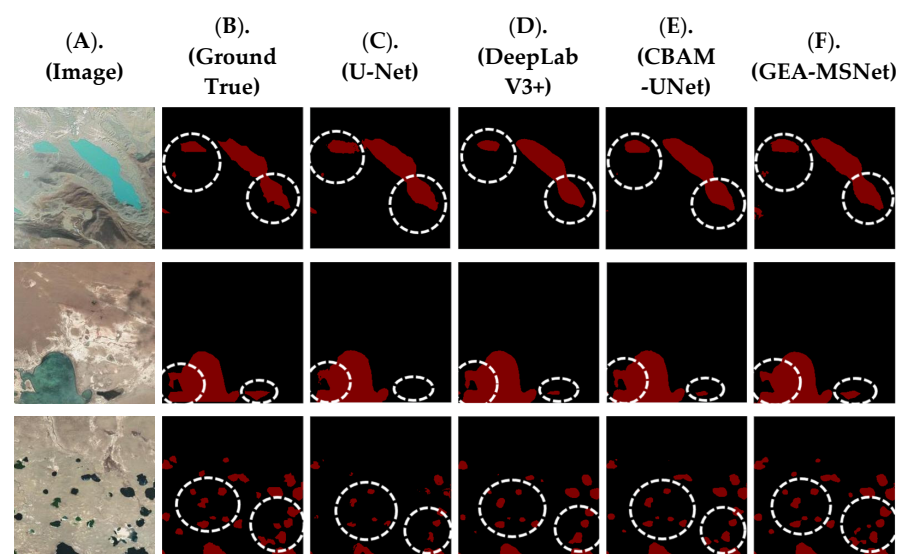


Figure 9. Displaying different models on Qinghai–Tibet Plateau lake datasets: (A) image; (B) ground true; (C) U-net; (D) DeepLabV3+; (E) CBAM-UNet; and (F) GEA-MSNet. The black area in the prediction results is the background, and the red area is the target water body. The white circle area represents the focal area where the prediction occurs.

For the visualization results in Figure 9, we specifically selected the DeepLabV3+ model and CBAM-UNet model. Regarding the lake edge segmentation, our model demonstrates a remarkable segmentation effect with significantly improved edge blurring compared to other models. In terms of small lake extraction, although our model does not achieve the complete extraction of all small water targets, its overall performance surpasses that of other models. Consequently, our model achieves commendable outcomes in the task of water segmentation.

5. Discussion

In the research of semantic segmentation algorithms, ablation experiments serve as an effective approach to validate the significance of each module [59]. In this study, we employed Res50-U-Net as the fundamental network architecture for conducting experiments. Subsequently, diverse combination models were devised for both the GEA module and the multi-scale feature fusion module individually. The specific results are presented in Table 3.

Table 3. The results of our ablation experiment, where ✓ indicates the selected modules in our study.

	Module Choose				Indicators		
	U-net	Res-50	GEA	Multi-Scale	mIoU	Recall	PA
Ablation Study Module	✓				63.05	75.07	83.55
	✓	✓			65.01	76.41	84.73
	✓	✓	✓		71.62	81.24	88.29
	✓	✓		✓	73.08	81.98	89.10
GEA-MSNet	✓	✓	✓	✓	75.49	83.79	90.21

According to the experimental results presented in Table 3, it is observed that the original U-Net model exhibits the poorest performance on the datasets, while incorporating the Res-50 structure into the U-Net model leads to improvements across all evaluation metrics. Notably, when solely employing the GEA module and Res-50 structure, a significant enhancement in model performance is evident. Furthermore, leveraging both the multi-scale feature fusion module and Res-50 structure demonstrates a substantial impact on enhancing model performance with an increase of 10.03%, 6.92%, and 5.55% for each respective evaluation indicator. However, the incorporation of the GEA module and the multi-scale feature fusion structure significantly enhances the model's segmentation performance, yielding an optimal mIoU of 75.49%, recall of 83.79%, and a pixel accuracy of 90.21%.

In summary, our model achieves excellent results in both dataset performance and model generalization testing. Compared to other models, the algorithm we designed can accurately distinguish the details of lake edges in scene classification datasets during experiment 1, while other models produce fuzzy predictions for lake edge detection. Our model outperforms others in detecting small lakes in remote sensing images during experiment 2 where other models have difficulty identifying these water bodies. Furthermore, each component of our model contributes to its overall effectiveness as demonstrated by ablation experiments indicating that it performs well for semantic segmentation tasks involving lakes.

6. Conclusions

In this paper, we propose a novel semantic segmentation algorithm model called GEA-MSNet for the water body semantic segmentation of remote sensing images. The proposed model improves in three aspects: firstly, we introduce an innovative high-efficiency attention module called GEA to enhance the learning capability of shallow features and aggregate spatial and channel information features while maintaining a reasonable parameter count. Secondly, a residual network architecture is incorporated into the encoder component to extract the semantic features of water bodies during image downsampling, thereby augmenting both the accuracy and generalization capabilities of the model for water body extraction. Thirdly, in the decoder section, a multi-scale feature fusion module is devised to effectively leverage

the semantic information of feature maps and enhance the precision of small water body extraction as well as lake edge segmentation during upsampling.

The GEA-MSNet model outperforms other models in the task of the semantic segmentation of water bodies in remote sensing images. In terms of visualization results, the GEA-MSNet model demonstrates superior performance in accurately segmenting lake edges and small water bodies, while maintaining a minimal increase in parameter count. Furthermore, we conducted ablation experiments to demonstrate the individual contributions of each component in our model. However, despite achieving the highest accuracy in the lake semantic segmentation task, GEA-MSNet still exhibits issues related to target segmentation errors and a loss of water body information. Addressing these challenges will be the focus of our future research endeavors. Additionally, there exist numerous potential combinations of enhanced modules that warrant further investigation into their specific effects and performance. Our subsequent strategy entails exploring diverse combinations to enhance the model's performance in water body segmentation. The research focus of this paper solely pertains to the segmentation of lake areas in remote sensing images. The algorithm's performance in other image segmentation tasks remains unknown, and its suitability for multi-target segmentation has not been explored. The generalization of our study was limited to the use of only two public datasets for experimentation; however, future research could expand this scope by incorporating additional datasets.

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

Funding: This work is funded by National Natural Science Foundation of China (No. 32160279), Science and Technology Major Project of Inner Mongolia (No. 2021ZD0011), Science and Technology Plan Project of Inner Mongolia, China (No. 2022YFHH0017).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The present study incorporates several publicly available datasets. The datasets were provided by National Cryosphere Desert Data Center. (<http://www.ncdc.ac.cn>, (accessed on 16 July 2023)).

Acknowledgments: The author expresses gratitude to Rui Meng and Fei Chen for their invaluable assistance in this article. Furthermore, the contributions of the diligent editors and insightful reviewers are also sincerely acknowledged.

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

References

1. Liu, H.; Hu, H.; Liu, X.; Jiang, H.; Liu, W.; Yin, X. A Comparison of Different Water Indices and Band Downscaling Methods for Water Bodies Mapping from Sentinel-2 Imagery at 10-M Resolution. *Water* **2022**, *14*, 2696. [\[CrossRef\]](#)
2. Dörnhöfer, K.; Oppelt, N. Remote sensing for lake research and monitoring—Recent advances. *Ecol. Indic.* **2016**, *64*, 105–122. [\[CrossRef\]](#)
3. Qiao, C.; Luo, J.; Sheng, Y.; Shen, Z.; Zhu, Z.; Ming, D. An Adaptive Water Extraction Method from Remote Sensing Image Based on NDWI. *J. Indian Soc. Remote Sens.* **2011**, *40*, 421–433. [\[CrossRef\]](#)
4. McFeeters, S.K. The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features. *Int. J. Remote Sens.* **1996**, *17*, 1425–1432. [\[CrossRef\]](#)
5. Xu, H. Modification of normalized difference water index(NDWI) to enhance open features in remotely sensed imagery. *Int. J. Remote Sens.* **2006**, *27*, 3025–3033. [\[CrossRef\]](#)
6. Zhang, G.; Yao, T.; Chen, W.; Zheng, G.; Shum, C.K.; Yang, K.; Piao, S.; Sheng, Y.; Yi, S.; Li, J.; et al. Regional differences of lake evolution across China during 1960s–2015 and its natural and anthropogenic causes. *Remote Sens. Environ.* **2019**, *221*, 386–404. [\[CrossRef\]](#)
7. Yu, H.; Yang, Z.; Tan, L.; Wang, Y.; Sun, W.; Sun, M.; Tang, Y. Methods and datasets on semantic segmentation: A review. *Neurocomputing* **2018**, *304*, 82–103. [\[CrossRef\]](#)

8. Shelhamer, E.; Long, J.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 640–651. [\[CrossRef\]](#) [\[PubMed\]](#)
9. Zunair, H.; Hamza, A.B. Sharp U-Net: Depthwise Convolutional Network for Biomedical Image Segmentation. *Comput. Biol. Med.* **2021**, *136*, 104699. [\[CrossRef\]](#) [\[PubMed\]](#)
10. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
11. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid Scene Parsing Network. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6230–6239.
12. Vaswani, A.; Shazeer, N.M.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All you Need. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*; NeurIPS Proceedings: New Orleans, LA, USA, 2017.
13. 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, Munich, Germany, 8–14 September 2018.
14. Oktay, O.; Schlemper, J.; Folgoc, L.L.; Lee, M.J.; Heinrich, M.P.; Misawa, K.; Mori, K.; McDonagh, S.G.; Hammerla, N.Y.; Kainz, B.; et al. Attention U-Net: Learning Where to Look for the Pancreas. *arXiv* **2018**, arXiv:1804.03999.
15. Zhong, H.-F.; Sun, H.-M.; Han, D.-N.; Li, Z.-H.; Jia, R.-S. Lake water body extraction of optical remote sensing images based on semantic segmentation. *Appl. Intell.* **2022**, *52*, 17974–17989. [\[CrossRef\]](#)
16. Li, M.; Wu, P.; Wang, B.; Park, H.; Hui, Y.; Yanlan, W. A Deep Learning Method of Water Body Extraction From High Resolution Remote Sensing Images With Multisensors. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 3120–3132. [\[CrossRef\]](#)
17. Lyu, X.; Jiang, W.; Li, X.; Fang, Y.; Xu, Z.; Wang, X. MSAFNet: Multiscale Successive Attention Fusion Network for Water Body Extraction of Remote Sensing Images. *Remote Sens.* **2023**, *15*, 3121. [\[CrossRef\]](#)
18. Luo, Y.; Feng, A.; Li, H.; Li, D.; Wu, X.; Liao, J.; Zhang, C.; Zheng, X.; Pu, H. New deep learning method for efficient extraction of small water from remote sensing images. *PLoS ONE* **2022**, *17*, e0272317. [\[CrossRef\]](#) [\[PubMed\]](#)
19. Yu, J.; Cai, Y.; Lyu, X.; Xu, Z.; Wang, X.; Fang, Y.; Jiang, W.; Li, X. Boundary-Guided Semantic Context Network for Water Body Extraction from Remote Sensing Images. *Remote Sens.* **2023**, *15*, 4325. [\[CrossRef\]](#)
20. Weng, L.; Xu, Y.; Xia, M.; Zhang, Y.; Liu, J.; Xu, Y. Water Areas Segmentation from Remote Sensing Images Using a Separable Residual SegNet Network. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 256. [\[CrossRef\]](#)
21. Li, Z.; Wang, R.; Zhang, W.; Hu, F.; Meng, L. Multiscale Features Supported DeepLabV3+ Optimization Scheme for Accurate Water Semantic Segmentation. *IEEE Access* **2019**, *7*, 155787–155804. [\[CrossRef\]](#)
22. Wang, Y.; Li, S.; Lin, Y.; Wang, M. Lightweight Deep Neural Network Method for Water Body Extraction from High-Resolution Remote Sensing Images with Multisensors. *Sensors* **2021**, *21*, 7397. [\[CrossRef\]](#) [\[PubMed\]](#)
23. Guo, H.; He, G.; Jiang, W.; Yin, R.; Yan, L.; Leng, W. A Multi-Scale Water Extraction Convolutional Neural Network (MWEN) Method for GaoFen-1 Remote Sensing Images. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 189. [\[CrossRef\]](#)
24. Wang, Z.; Gao, X.; Zhang, Y.; Zhao, G. MSLWENet: A Novel Deep Learning Network for Lake Water Body Extraction of Google Remote Sensing Images. *Remote Sens.* **2020**, *12*, 4140. [\[CrossRef\]](#)
25. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.
26. Zhong, H.-F.; Sun, Q.; Sun, H.-M.; Jia, R.-S. NT-Net: A Semantic Segmentation Network for Extracting Lake Water Bodies From Optical Remote Sensing Images Based on Transformer. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5627513. [\[CrossRef\]](#)
27. Zhang, Y.; Lu, H.; Ma, G.; Zhao, H.; Xie, D.; Geng, S.; Tian, W.; Sian, K.T.C.L.K. MU-Net: Embedding MixFormer into Unet to Extract Water Bodies from Remote Sensing Images. *Remote Sens.* **2023**, *15*, 3559. [\[CrossRef\]](#)
28. Zhao, X.; Wang, H.; Liu, L.; Zhang, Y.; Liu, J.; Qu, T.; Tian, H.; Lu, Y. A Method for Extracting Lake Water Using ViTenc-UNet: Taking Typical Lakes on the Qinghai-Tibet Plateau as Examples. *Remote Sens.* **2023**, *15*, 4047. [\[CrossRef\]](#)
29. Chen, C.; Wang, Y.; Yang, S.; Ji, X.; Wang, G. A K-Net-based hybrid semantic segmentation method for extracting lake water bodies. *Eng. Appl. Artif. Intell.* **2023**, *126*, 106904. [\[CrossRef\]](#)
30. Wang, H.; Shen, Y.; Liang, L.; Yuan, Y.; Yan, Y.; Liu, G.; Lakshmana, K. River Extraction from Remote Sensing Images in Cold and Arid Regions Based on Attention Mechanism. *Wirel. Commun. Mob. Comput.* **2022**, *2022*, 9410381. [\[CrossRef\]](#)
31. Huang, M.; Cheng, C.; De Luca, G.; Xue, X. Remote Sensing Data Detection Based on Multiscale Fusion and Attention Mechanism. *Mob. Inf. Syst.* **2021**, *2021*, 6466051. [\[CrossRef\]](#)
32. Cheng, G.; Han, J.; Lu, X. Remote Sensing Image Scene Classification: Benchmark and State of the Art. *Proc. IEEE* **2017**, *105*, 1865–1883. [\[CrossRef\]](#)
33. Gu, S.; Zhang, R.; Luo, H.; Li, M.; Feng, H.; Tang, X. Improved SinGAN Integrated with an Attentional Mechanism for Remote Sensing Image Classification. *Remote Sens.* **2021**, *13*, 1713. [\[CrossRef\]](#)
34. Toda, R.; Teramoto, A.; Kondo, M.; Imaizumi, K.; Saito, K.; Fujita, H. Lung cancer CT image generation from a free-form sketch using style-based pix2pix for data augmentation. *Sci. Rep.* **2022**, *12*, 12867. [\[CrossRef\]](#)
35. Kuntalp, M.; Düzyel, O. A new method for GAN-based data augmentation for classes with distinct clusters. *Expert Syst. Appl.* **2024**, *235*, 121199. [\[CrossRef\]](#)
36. Mariani, G.; Scheidegger, F.; Istrate, R.; Bekas, C.; Malossi, A.C.I. BAGAN: Data Augmentation with Balancing GAN. *arXiv* **2018**, arXiv:1803.09655.

37. Ding, H.; Jiang, X.; Shuai, B.; Liu, A.Q.; Wang, G. Semantic Segmentation With Context Encoding and Multi-Path Decoding. *IEEE Trans. Image Process.* **2020**, *29*, 3520–3533. [\[CrossRef\]](#)
38. Dieste, Á.G.; Argüello, F.; Heras, D.B. ResBaGAN: A Residual Balancing GAN with Data Augmentation for Forest Mapping. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 6428–6447. [\[CrossRef\]](#)
39. Kriegeskorte, N.; Douglas, P.K. Interpreting encoding and decoding models. *Curr. Opin. Neurobiol.* **2019**, *55*, 167–179. [\[CrossRef\]](#)
40. Xu, H.; Huang, Y.; Hancock, E.R.; Wang, S.; Xuan, Q.; Zhou, W. Pooling Attention-based Encoder–Decoder Network for semantic segmentation. *Comput. Electr. Eng.* **2021**, *93*, 107260. [\[CrossRef\]](#)
41. Xing, Y.; Zhong, L.; Zhong, X. An Encoder-Decoder Network Based FCN Architecture for Semantic Segmentation. *Wirel. Commun. Mob. Comput.* **2020**, *2020*, 8861886. [\[CrossRef\]](#)
42. Chen, B.; Li, P.; Sun, C.; Wang, D.; Yang, G.; Lu, H. Multi attention module for visual tracking. *Pattern Recognit.* **2019**, *87*, 80–93. [\[CrossRef\]](#)
43. Hou, Y.; Luo, Z.; Deng, J.; Gao, Y.; Huang, K.; Li, W. Attention meets involution in visual tracking. *J. Vis. Commun. Image Represent.* **2023**, *90*, 103746. [\[CrossRef\]](#)
44. Wang, J.; Meng, C.; Deng, C.; Wang, Y. Learning attention modules for visual tracking. *Signal Image Video Process.* **2022**, *16*, 2149–2156. [\[CrossRef\]](#)
45. Niu, Z.; Zhong, G.; Yu, H. A review on the attention mechanism of deep learning. *Neurocomputing* **2021**, *452*, 48–62. [\[CrossRef\]](#)
46. Ghaffarian, S.; Valente, J.; van der Voort, M.; Tekinerdogan, B. Effect of Attention Mechanism in Deep Learning-Based Remote Sensing Image Processing: A Systematic Literature Review. *Remote. Sens.* **2021**, *13*, 2965. [\[CrossRef\]](#)
47. de Santana Correia, A.; Colombini, E.L. Attention, please! A survey of neural attention models in deep learning. *Artif. Intell. Rev.* **2022**, *55*, 6037–6124. [\[CrossRef\]](#)
48. Brauwiers, G.; Frasinicar, F. A General Survey on Attention Mechanisms in Deep Learning. *IEEE Trans. Knowl. Data Eng.* **2023**, *35*, 3279–3298. [\[CrossRef\]](#)
49. Das, M.N.; Das, D.S. Attention-UNet Architectures with Pretrained Backbones for multi-class Cardiac MR Image Segmentation. *Curr. Probl. Cardiol.* **2023**, *49*, 102129. [\[CrossRef\]](#)
50. Song, H.; Wu, H.; Huang, J.; Zhong, H.; He, M.; Su, M.; Yu, G.; Wang, M.; Zhang, J. HA-Unet: A Modified Unet Based on Hybrid Attention for Urban Water Extraction in SAR Images. *Electronics* **2022**, *11*, 3787. [\[CrossRef\]](#)
51. Zhang, K.; Sun, M.; Han, T.X.; Yuan, X.; Guo, L.; Liu, T. Residual Networks of Residual Networks: Multilevel Residual Networks. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *28*, 1303–1314. [\[CrossRef\]](#)
52. Huang, K.; Li, S.; Deng, W.; Yu, Z.; Ma, L. Structure inference of networked system with the synergy of deep residual network and fully connected layer network. *Neural Netw.* **2022**, *145*, 288–299. [\[CrossRef\]](#) [\[PubMed\]](#)
53. Alaeddine, H.; Jihene, M. Wide deep residual networks in networks. *Multimed. Tools Appl.* **2023**, *82*, 7889–7899. [\[CrossRef\]](#)
54. Alaeddine, H.; Jihene, M. Deep Residual Network in Network. *Comput. Intell. Neurosci.* **2021**, *2021*, 6659083. [\[CrossRef\]](#) [\[PubMed\]](#)
55. Sediqi, K.M.; Lee, H.J. A Novel Upsampling and Context Convolution for Image Semantic Segmentation. *Sensors* **2021**, *21*, 2170. [\[CrossRef\]](#) [\[PubMed\]](#)
56. Wang, P.; Chen, P.; Yuan, Y.; Liu, D.; Huang, Z.; Hou, X.; Cottrell, G. Understanding Convolution for Semantic Segmentation. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018; pp. 1451–1460.
57. Hou, Q.; Zhou, D.; Feng, J. Coordinate Attention for Efficient Mobile Network Design. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 13708–13717.
58. Woo, S.; Debnath, S.; Hu, R.; Chen, X.; Liu, Z.; Kweon, I.-S.; Xie, S. ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 16133–16142.
59. Peng, S.; Jiang, W.B.; Pi, H.; Bao, H.; Zhou, X. Deep Snake for Real-Time Instance Segmentation. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 8530–8539.

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.