remote sensing

MDPI

Article

A Multi-Modality Fusion and Gated Multi-Filter U-Net for Water Area Segmentation in Remote Sensing

Rongfang Wang [1,*], Chenchen Zhang [1], Chao Chen [2], Hongxia Hao [3], Weibin Li [1] and Licheng Jiao [1]

[1] School of Artificial Intelligence, Xidian Univerisity, Xi'an 710071, China; 22171214748@stu.xidian.edu.cn (C.Z.); weibinli@xidian.edu.cn (W.L.); lchjiao@mail.xidian.edu.cn (L.J.)
[2] IBM Research, Silicon Valley Lab, San Jose, CA 95141, USA; chaochenxidian@gmail.com
[3] School of Computer Science and Technology, Xidian Univerisity, Xi'an 710126, China; hxhao@xidian.edu.cn
* Correspondence: rfwang@xidian.edu.cn

Abstract: Water area segmentation in remote sensing is of great importance for flood monitoring. To overcome some challenges in this task, we construct the Water Index and Polarization Information (WIPI) multi-modality dataset and propose a multi-Modality Fusion and Gated multi-Filter U-Net (MFGF-UNet) convolutional neural network. The WIPI dataset can enhance the water information while reducing the data dimensionality: specifically, the Cloud-Free Label provided in the dataset can effectively alleviate the problem of labeled sample scarcity. Since a single form or uniform kernel size cannot handle the variety of sizes and shapes of water bodies, we propose the Gated Multi-Filter Inception (GMF-Inception) module in our MFGF-UNet. Moreover, we utilize an attention mechanism by introducing a Gated Channel Transform (GCT) skip connection and integrating GCT into GMF-Inception to further improve model performance. Extensive experiments on three benchmarks, including the WIPI, Chengdu and GF2020 datasets, demonstrate that our method achieves favorable performance with lower complexity and better robustness against six competing approaches. For example, on the WIPI, Chengdu and GF2020 datasets, the proposed MFGF-UNet model achieves F1 scores of 0.9191, 0.7410 and 0.8421, respectively, with the average F1 score on the three datasets 0.0045 higher than that of the U-Net model; likewise, GFLOPS were reduced by 62% on average. The new WIPI dataset, the code and the trained models have been released on GitHub.

Keywords: water area segmentation; multi-modality fusion; multi-filter inception; attention mechanism; remote sensing

Citation: Wang, R.; Zhang, C.; Chen, C.; Hao, H.; Li, W.; Jiao, L. A Multi-Modality Fusion and Gated Multi-Filter U-Net for Water Area Segmentation in Remote Sensing. *Remote Sens.* **2024**, *16*, 419. https://doi.org/10.3390/rs16020419

Academic Editors: Chunlei Huo, Zhiqiang Zhou, Lurui Xia and Samia Ainouz

Received: 13 December 2023
Revised: 13 January 2024
Accepted: 20 January 2024
Published: 21 January 2024

1. Introduction

Detection of surface water area is essential for water resource management, flood identification, and ecological protection [1]. Satellite remote sensing images have the advantages of large coverage, low cost and a short data acquisition period and are often used in water area analysis [2–4]. The key to accurately segmenting water areas from complex ground features and creating water area maps lies in effectively highlighting water bodies.

Spectral data are the first choice for creating water area maps in cloudless conditions. They can provide multi-band remote sensing images with high resolution and less noise. The disadvantage of spectral imaging is that a spectral imager is a passive imaging instrument that needs sunlight to provide a light source, so it cannot provide satellite image all day long, and the acquired image has information loss due to the light being blocked by clouds. Synthetic Aperture Radar (SAR), which is based on the imaging of backscattered polarization information, can detect water areas without cloud interference and take images of the ground all day long. However, polarization information in SAR images cannot distinguish water, water-like surfaces and noise-like spots well [5]. These defects may limit the application of polarization information in water segmentation. Therefore, the combination of spectral data and SAR helps to improve the effect of automatic water segmentation.

In early studies, water area segmentation was limited by remote sensing technology and image quality [6]. In low-resolution remote sensing images, threshold filtering can only be carried out by the different radiation intensities of water in different bands. However, when threshold filtering is applied directly to the original imaging data, an accurate segmentation map cannot be obtained in most cases. Therefore, a water index method has been developed. Water-index-based methods make use of the intrinsic nature that water bodies have different reflectance under different wavelengths. Thereby, we can choose multiple bands from spectral imaging data and combine them to highlight water bodies and to inhibit other ground objects [7]. And then, a threshold method is used to filter the calculated water index to extract the water areas. However, for the threshold method, it is difficult to determine an appropriate threshold value that can adapt to any complex ground environment. For different water indexes from different scenarios, the threshold value is not uniform. In order to solve the above issues, some researchers have started to establish water segmentation algorithms based on machine learning to reduce the dependence on the threshold; examples include the use of Random Forest (RF) [8], Support Vector Machine (SVM) [9], Decision Tree (DT) and Deep Neural Networks (DNNs) [10,11].

Compared with other machine learning methods, DNN based on deep learning has a strong ability for automatic feature extraction and is able to learn semantic, high-level and deeper features. Therefore, deep-learning-based automatic segmentation methods have been widely applied across various fields, especially with the emergence of numerous approaches based on Convolutional Neural Networks (CNNs). Long et al. [12] proposed a Fully Convolutional Network (FCN) for natural image semantic segmentation for the first time. FCN replaces the fully connected layers of a pre-trained convolutional neural network, such as VGG16 [13], with fully convolutional layers. It allows the network to accept input images of arbitrary size and uses upsampling to restore the size of the output, thus achieving end-to-end mapping from the input image to the output segmentation map. A U-shaped convolutional Network (U-Net) is an improvement on the FCN architecture by Ronneberger et al. [14] in 2015 and has been applied to medical image segmentation. U-Net consists of an encoder (down-sampling path) and a decoder (up-sampling path) connected by skip connections. Due to its simple structure, U-Net can achieve high accuracy with limited training samples, which makes it and its series of extensions widely used in various image segmentation tasks, including remote sensing image segmentation [15,16]. DeepLab series [17–19] are another series of CNN-based models for image semantic segmentation that have been proposed and iteratively improved by the Google team. Among them, Deeplabv3+ [19] is the latest version, in which the model performs convolution operations by employing parallel atrous convolutions at various rates to capture contextual features at multiple scales. Boundary-Aware Salient object detection Network (BASNet) [20] was proposed by Qin et al. in 2019 for salient detection and water segmentation tasks. BASNet generates coarse prediction maps using the encoder–decoder network and then refines the coarse prediction maps with the ResNet34 [21] refinement module to improve the performance. For solving some difficulties existing in water areas segmentation, such as complex backgrounds, huge scales, water connectivity, and rough edges, Liu et al. [1] proposed Dense U-net Plus Network (DUPNet) in 2022, which is also based on the U-Net architecture. DUPNet integrates dense block modules from DenseNet [22] and Multi-scale Spatial Pyramid Pooling (MSPP) modules from DeeplabV3+. Dense blocks are employed as the feature extraction module for both the encoder and decoder, while MSPP is utilized for skip connections. In recent years, models based on self-attention mechanisms have made significant progress in computer vision tasks, particularly with the emergence of Vision Transformer (ViT) [23]. However, ViT-based architectures are generally parameter-heavy and require more labeled samples and computing resources. SegNeXt [24] is a popular convolutional attention network for semantic segmentation proposed by Guo et al. in 2022. SegNeXt designs a convolutional attention module that encodes contextual information more efficiently than self-attention in transformers.

Among the different methods above, improved methods based on U-Net have received more attention from researchers for water segmentation tasks. In the U-Net framework, the encoder module obtains high-level semantic information by extracting features from images layer by layer, while the decoder module collects different levels of semantic information and gradually recovers spatial information of features. Based on this simple and effective encoder–decoder network, Goutam Konapala et al. [2] successfully drew a flood inundation map. They used the inundation data of 11 global flood events as samples to conduct training on the U-shaped convolutional network. Peri Akiva et al. [25] proposed a self-supervised flood segmentation model via adversarial domain adaptation and label refinement. In this model, a refiner is added to refine the rough mask obtained by the threshold method on the improved normalized differential water index to finally improve the accuracy of flood segmentation.

Although the emergence of deep learning technology has greatly improved the accuracy and efficiency of water area segmentation, there are still the following challenges: (1) Remote sensing images usually include a wide range of ground types, so the region of non-water contains complex and diverse ground object information and occupies a relatively large proportion, while the water area is generally small. This is a typical class imbalance problem. (2) Usually, there are water bodies of different sizes and shapes in remote sensing images, including from thousands of pixels to just a few pixels. In this case, using a single form or uniform kernel size in the convolution operation cannot obtain good segmentation results. (3) Compared with natural images, remote sensing images have more interference and noise, which affect the accuracy of water segmentation. Meanwhile, in the case of high-altitude shooting, the water area under the cloud shadow cannot be effectively separated by multispectral imaging, so it is necessary to utilize multi-modality data to obtain complementary information. (4) In existing multi-modality methods for water segmentation, direct superposition of original data or features is generally adopted for the use of multi-modality data. This simple superposition method cannot effectively extract the complementary features between different modalities, but it does produce many redundant features, which affects the final segmentation results.

To overcome the aforementioned challenges, we construct the Water Index and Polarization Information (WIPI) multi-modality dataset and propose a multi-Modality Fusion and Gated multi-Filter U-Net (MFGF-UNet) convolutional neural network for water area segmentation in remote sensing. Water indexes are initially used in threshold-based unsupervised methods for water body extraction. The water index is calculated by combining imaging information from different bands utilizing the strong absorption and low radiation of water in the wavelength range from visible light to infrared. After calculation, the pixel value of the combined band shows a relatively obvious difference between water and non-water materials, which can be used to segment water bodies by setting an appropriate threshold. After years of development, various water indexes have been proposed. McFeeters et. al. [26] proposed a Normalized Difference Water Index (NDWI) and used it for water body segmentation of multispectral images. Xu et al. [27] proposed a modified NDWI (MNDWI) by replacing the near-infrared band with the short-wave infrared band. Previous studies have shown that MNDWI is more suitable for enhancing water information and can extract water bodies more accurately than NDWI. In addition to the above two water indices, many researchers have proposed water indexes for other scenarios. In this work, to enhance water information while reducing the channel dimensionality of data, we analyze the characteristics of the water index and extract 7 water indexes from 13 bands of raw multispectral images. However, multispectral images have large areas covered by clouds, which are defined as invalid areas and cannot participate in training. Therefore, based on the extracted water index, we fuse SAR polarization information for constructing the WIPI multi-modality dataset. Moreover, we propose a Multi-Model Decision Fusion (MMDF) method to automatically obtain Cloud-Free Labels to alleviate the problem of labeled sample scarcity and class imbalance in water segmentation.

Meanwhile, another promising advancement in the computer vision field is attention mechanisms, in which networks can allocate more resources to important regions in an image. Currently, attention mechanisms have been widely applied in many visual tasks, such as image classification [28,29], object detection [30,31], semantic segmentation [32,33], medical image processing [34,35], 3D vision [36] and multimodal tasks [37,38]. Gated Channel Transformation (GCT) [21] is a type of channel attention mechanism that explicitly models the relationships between channels in feature maps using learnable variables. These learnable variables can determine the competitive or cooperative behavior of neurons. Compared to Squeeze-and-Excitation (SE) networks [28], GCT greatly reduces the number of parameters and computational complexity by introducing channel normalization layers. Thereby, in our MFGF-UNet model, we replace the direct skip connection of UNet with the GCT skip connection for adaptively focusing on more meaningful channel information and enhancing inter-channel feature extraction. Furthermore, since water bodies vary significantly in size and shape, to avoid the limitations brought by the single convolution form and uniform kernel size as well as to fully utilize the multi-modal information of the constructed WIPI dataset, we design a gated multi-filter inception module between the input and the U-shaped backbone.

Finally, our MFGF-UNet model is compared to six other methods on three water segmentation datasets, including our constructed WIPI dataset, the Chengdu dataset and the GF2020 dataset. These datasets contain multispectral, SAR and optical multi-modality images, cover different regions and are captured at different times. Extensive experiments of performance and complexity analysis show that our MFGF-UNet outperforms various competing methods.

To sum up, the main contributions of this paper are as follows:

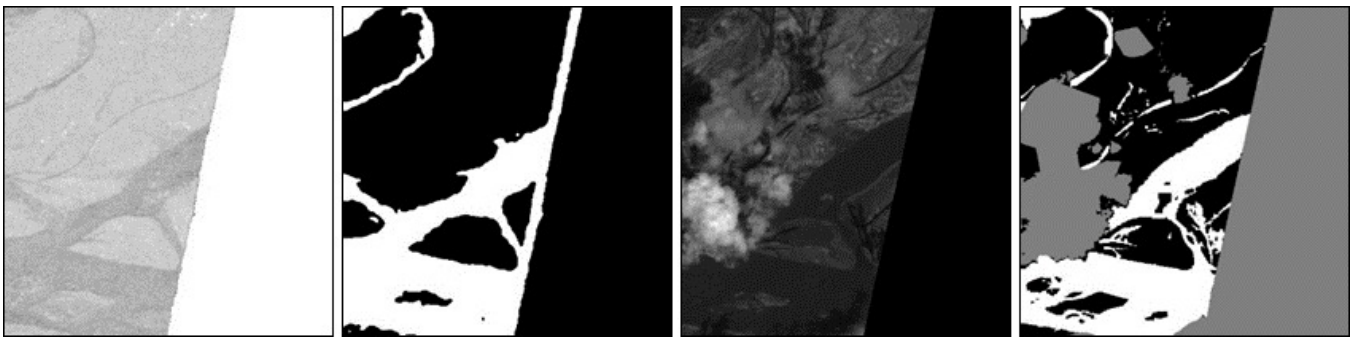
- We release the Water Index and Polarization Information multi-modality dataset for water area segmentation in remote sensing. The proposed datasets are the first ever to provide both the water index modal and the corresponding polarization information modal. Specifically, the dataset contains the Cloud-Free Labels automatically obtained by the proposed Multi-Model Decision Fusion algorithm, which can effectively alleviate the problem of labeled sample scarcity.
- We propose a multi-modality fusion and gated multi-filter U-shaped convolutional neural network that introduces the GCT skip connection for adaptively focusing on more meaningful channels and that incorporates the proposed gated multi-filter inception module for fully utilizing the multi-modality information and handling the challenge due to the variety of water body sizes and shapes.
- Extensive experiments on three benchmarks, including WIPI (water index and SAR images), Chengdu (multispectral images) and GF2020 (optical images) datasets demonstrated that our MFGF-UNet model achieves favorable performance with lower complexity and better robustness against six competing approaches.

2. WIPI Multi-Modality Dataset Construction

2.1. Original Sen1Floods11 Dataset

The original data used to construct the Water Index and Polarization Information (WIPI) multi-modality dataset are derived from the Sen1Floods11 (https://github.com/cloudtostreet/Sen1Floods11 (accessed on 12 December 2023)) [39] dataset, which consists of raw Sentinel-1 SAR images (Sen1-SAR) and raw Sentinel-2 multispectral images (Sen2-MS), where Sen1-SAR includes the two VV and VH bands, and Sen2-MS includes 13 bands (1–8, 8A, 9–12) across all spectra. The format and description of the bands for Sen1Floods11 is shown in Appendix A, Table A1. The Sen1Floods11 dataset is a surface water dataset covering 120,406 km^2 and 6 continents of the world across 11 flood events. Events are selected by the requirements that the flood event had coverage from Sentinel-1 as well as coincident Sentinel-2 imaging on the same day or within 2 days of the Sentinel-1 image. In addition to images, Sen1-SAR provides automated labels based on the Otsu threshold method (Otsu Label-SAR) that includes two classes: water (label '1') and non-water (label

'0'), and Sen2-MS provides manual labels (Manual Label-MS) that include three classes: water (label '1'), non-water (label '0') and cloud and invalid region (label '−1'). A set of sample images and masks from the Sen1Floods11 dataset is shown in Figure 1.



(**a**) (**b**) (**c**) (**d**)

Figure 1. A set of sample images and masks from the Sen1Floods11 dataset. (**a**) Sen1-SAR image (band VV), (**b**) Otsu Label-SAR, (**c**) Sen2-MS image (Band 8) and (**d**) Manual Label-MS.

2.2. Water Index Extraction

In this work, instead of directly superimposing the SAR and multispectral images, we first analyze the characteristics of the water index and extract 7 water indexes from 13 bands of raw MSI to enhance water information while reducing the channel dimensionality of the data. The seven extracted water indexes are: (1) Normalized Difference Vegetation Index (NDVI) [40], (2) Normalized Difference Moisture Index (NDMI) [41], (3) Normalized Difference Water Index (NDWI) [26], (4) Modified Normalized Difference Water Index (MNDWI) [27], (5) Automated Water Extraction Index Non-Shadow ($AWEI_{NS}$) [42], (6) Automated Water Extraction Index Shadow ($AWEI_S$) [42] and (7) Linear Discriminant Analysis Water Index (LDAWI) [43]. The calculation formula for each water index is shown in Appendix A, Table A2.

2.3. Obtaining Cloud-Free Labels

As can be seen from Figure 1, although the SAR image is not interfered with by clouds, the corresponding image only has coarse labels obtained by the unsupervised Otsu threshold method. Meanwhile, multispectral images have fine manual labels but have large areas covered by clouds, which are defined as invalid areas. Even for SAR images without cloud obstruction, these invalid regions cannot participate in training; this exacerbates the problem of labeled sample scarcity and class imbalance in water segmentation. Therefore, we propose a Multi-Model Decision Fusion (MMDF) method to automatically obtain Cloud-Free Labels based on the fine manual labels of multispectral images. The framework of MMDF is shown in Figure 2.

Since clouds and invalid regions have the same label in Manual Label-MS, first, we use the VV band of the SAR image to extract the invalid region, and then we remove the invalid region from Manual Label-MS to get the Cloud Label with only the cloud part. Specifically, this process leverages the characteristic of 'NaN' (Not a Number) values in the VV (Vertical Vertical) band and uses it to identify the locations of invalid regions. In Manual Label-MS, both clouds and invalid regions are labeled as '−1'. By computing the difference set between the '−1' region and the 'NaN' region, we can obtain the Cloud Label. On the other hand, seven extracted water indexes from multispectral images and the VV and VH from SAR images are concatenated as the input to train the semantic segmentation models. In training, Manual Label-MS is used to calculate the loss function. Two state-of-the-art models of SegNeXt [24] and DeeplabV3+ [19] are trained to get Predicted Label 1 and Predicted Label 2, respectively. Next, based on the Cloud Label, we use Otsu Label-SAR, Predicted Label 1, and Predicted Label 2 to vote on the suspected cloud region. Finally, we replace the label '−1' (cloud region) of Manual Label-MS with the voting results label

'0' (non-water) or '1' (water). In this way, the construction of the whole Water Index and Polarization Information (WIPI) dataset is completed. The WIPI dataset includes multi-modality images from nine channels (seven water indexes + VV + VH) and the Cloud-Free Label automatically obtained by our proposed MMDF algorithm. Example instances from the nine channels and the corresponding ground truths (Cloud-Free Labels) are shown in Figure 3.

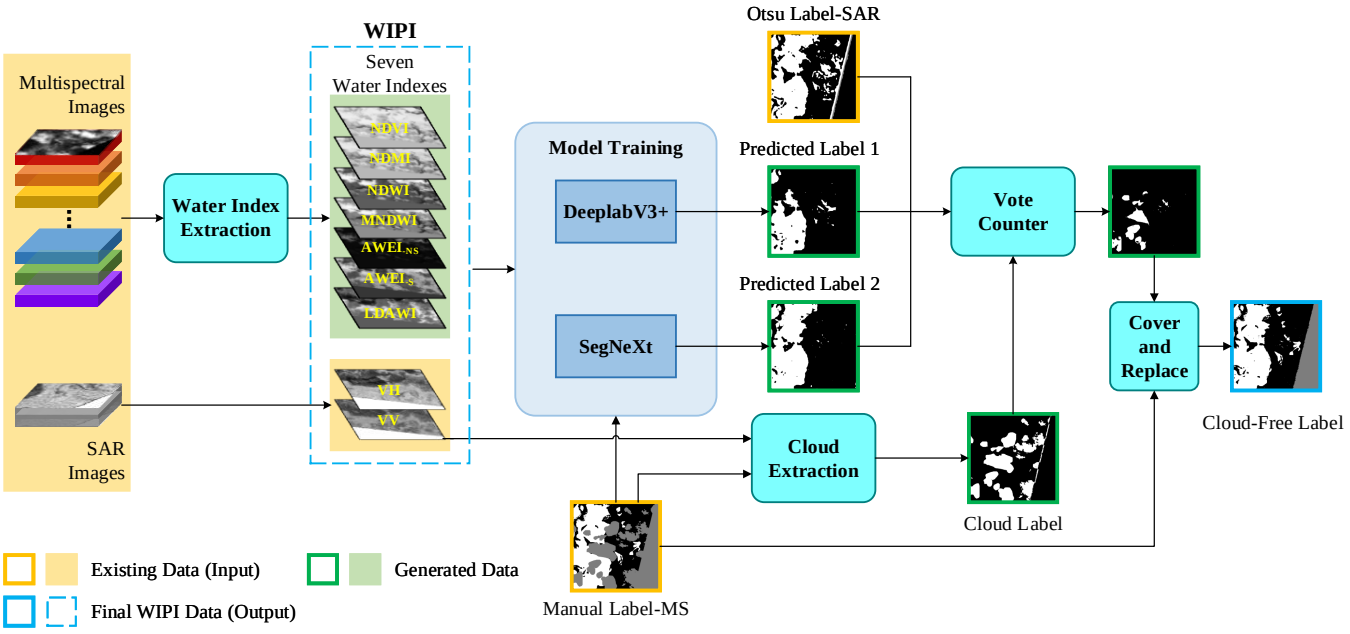


Figure 2. The framework of MMDF to automatically obtain Cloud-Free Labels.

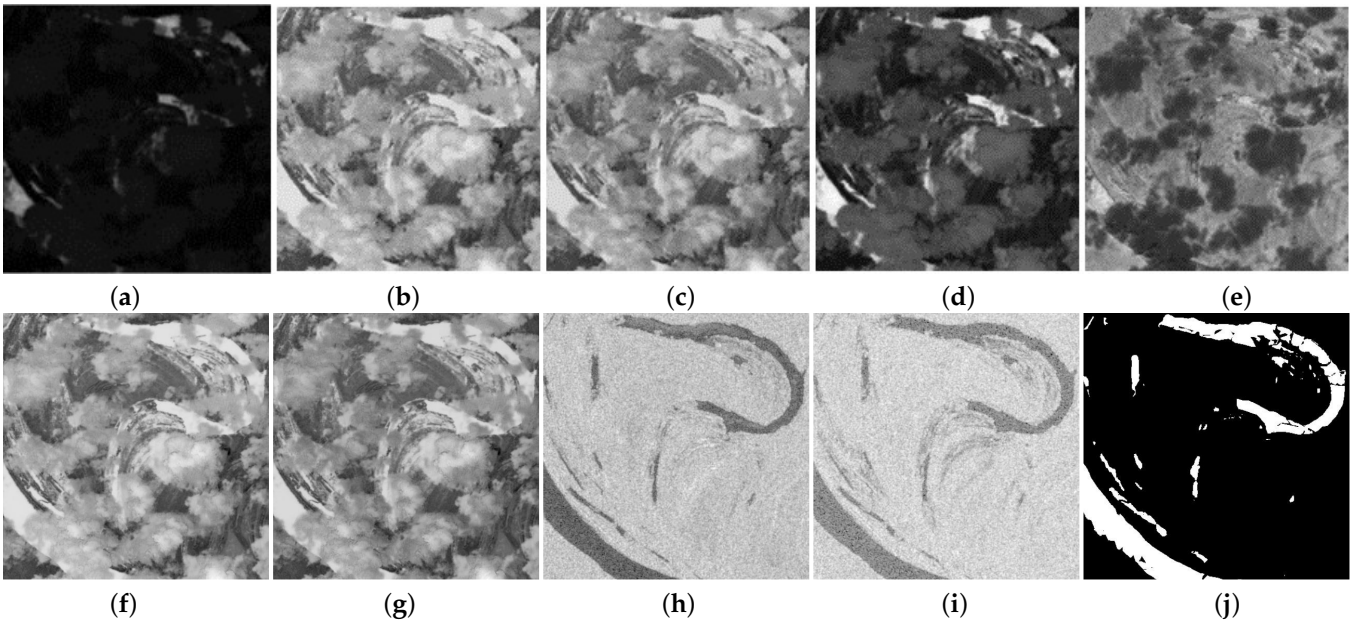


Figure 3. Example instances from the nine channels and the ground truth (Cloud-Free Label) of WIPI dataset: (**a**) NDVI, (**b**) NDMI, (**c**) NDWI, (**d**) MNDWI, (**e**) $AWEI_{NS}$, (**f**) $AWEI_{S}$, (**g**) LDAWI, (**h**) VV, (**i**) VH and (**j**) GT.

3. Methodology

The framework of our proposed multi-Modality Fusion and Gated multi-Filter U-Net (MFGF-UNet) is shown in Figure 4. MFGF-UNet is an end-to-end U-shaped network that

mainly consists of four parts: (1) WIPI multi-modality data construction, (2) U-shaped encoder–decoder, (3) skip connection with channel attention and (4) gated multi-filter inception module. The first part was introduced in detail in Section 2, and the details of the remaining three parts are described in the following.

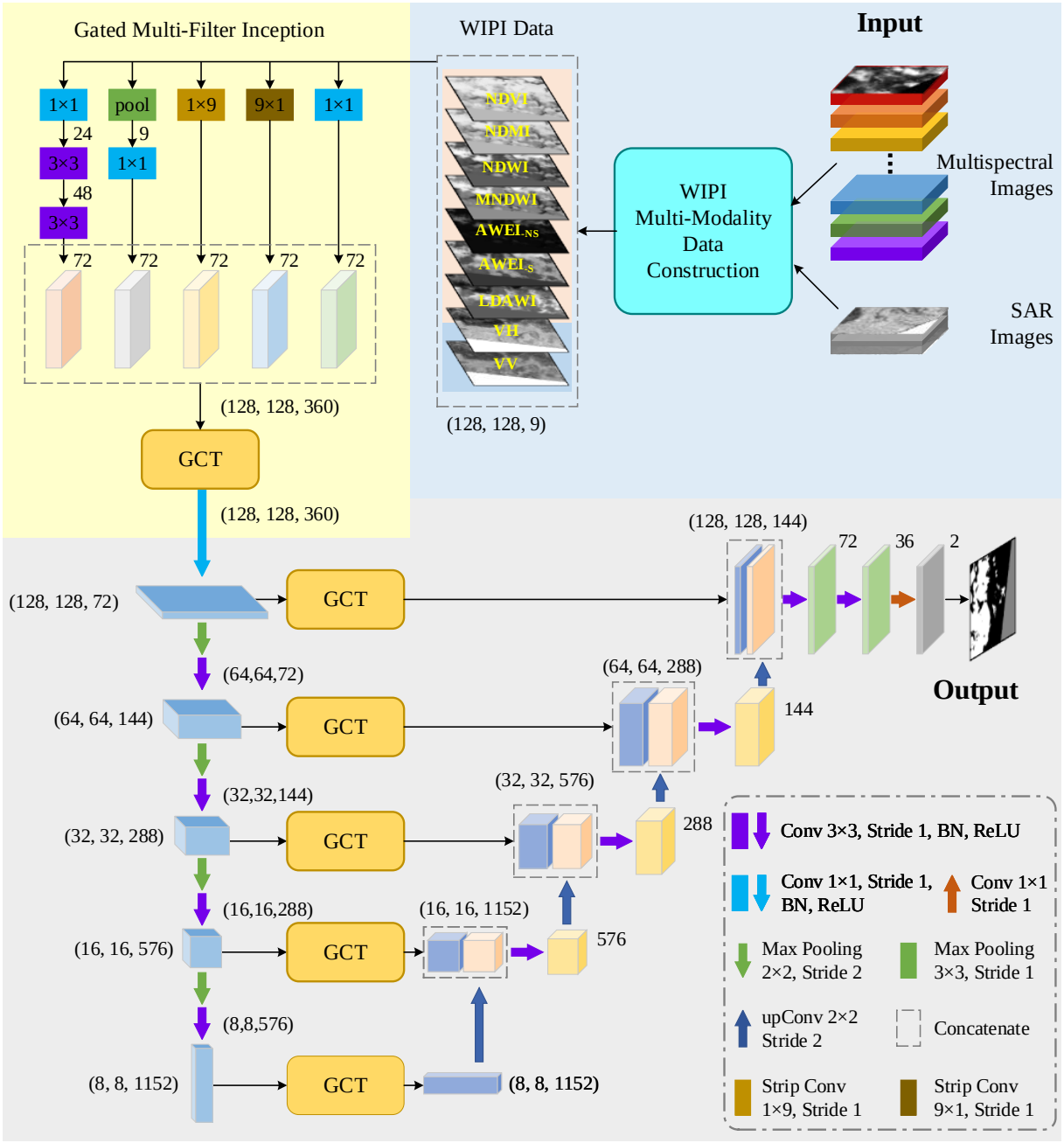


Figure 4. Illustration of the proposed MFGF-UNet framework.

3.1. U-Shaped Encoder-Decoder

Due to the existence of relatively few labeled samples for water segmentation, we adopt a simple and efficient U-shaped encoder–decoder as the base architecture. In the encoder path, we design four layers comprising max-pooling and ConvBR blocks, which repeatedly perform the following operations:

$$\tilde{x}^{(h/2,w/2,c)} = \text{maxPool}(x_l^{(h,w,c)}) \tag{1}$$

$$x_{l+1}^{(h/2,w/2,2c)} = \text{ReLU}(\text{BN}(\text{Conv}_{3\times3}(\tilde{x}^{(h/2,w/2,c)}))) \tag{2}$$

where $x_l^{(h,w,c)}$ indicates the feature map, the subscript l denotes the current layer, and the superscript (h, w, c) denotes the resolution and number of channels under the corresponding layer. The function maxPool$(\cdot)$ denotes 2 × 2 max-pooling with stride = 2 for down-sampling the feature map of the l-th layer; Conv$_{3\times3}(\cdot)$ denotes 3 × 3 convolution with stride = 1, and the number of output channels is twice that of the input channels; BN$(\cdot)$ and ReLU$(\cdot)$ denote Batch Normalization [44] and Rectified Linear Units [45], respectively.

The decoder path also consists of four layers of up-convolution and a ConvBR block, which can be expressed as:

$$\widetilde{\mathbf{y}}^{(h,w,c/2)} = \mathbf{u}\text{pConv}(\mathbf{y}_{l+1}^{(h/2,w/2,c)}) \tag{3}$$

$$\mathbf{y}_l^{(h,w,c/2)} = \mathbf{R}\text{eLU}(\mathbf{B}\text{N}(\mathbf{C}\text{Conv}_{3\times3}(\mathbf{C}\text{ont}(\hat{\mathbf{x}}_l^{(h,w,c/2)}, \widetilde{\mathbf{y}}^{(h,w,c/2)})))) \tag{4}$$

where upConv$(\cdot)$ denotes 2 × 2 up-convolution with stride = 2; it up-samples the input feature map to double its size while reducing the channel number by half. The ConvBR block of Equation (4) is similar to Equation (2) except that the output channel of $\mathbf{C}\text{Conv}_{3\times3}(\cdot)$ is half the input channel number. The output of the up-convolution $\widetilde{\mathbf{y}}^{(h,w,c/2)}$ and the skip connections $\hat{\mathbf{x}}_l^{(h,w,c/2)}$ are concatenated along the channel dimension by the $\mathbf{C}\text{ont}(\cdot)$ (concatenating operation) as the input of the ConvBR block. The last layer of the decoder is a 1 × 1 convolution layer that maps the number of feature channels to the number of classes and produces the final segmentation map.

3.2. Skip Connections with Channel Attention

As shown in Figure 4, we introduce the Gated Channel Transform (GCT) [46] module for skip connections to fuse shallow features from the encoder with deep features from the decoder. Compared to traditional skip connections, adding the GCT module can adaptively adjust the weight of each channel and improve the performance of the whole network. GCT consists of three parts: Global Context Embedding, Channel Normalization, and Gating Adaptation, where Channel Normalization is a parameter-free operation. GCT places the learnable parameters in the Global Context Embedding and Gating Adaptation operations. The former collects global information via the L2 norm and assigns a learnable weight control parameter for each channel. The latter introduces a competitive mechanism between channels via channel normalization and assigns a learnable gating parameter and bias for each channel. The last stage of GCT uses tanh to activate the attention vector and get enhanced features.

Let $x_l^{(h,w,c)}$ be the input feature of GCT of the current layer l and α, γ and β be the trainable parameters. The calculation process of GCT is as follows:

$$\mathbf{G}\text{CE} = \alpha \times L_2(\mathbf{x}_l^{(h,w,c)}) \tag{5}$$

$$\mathbf{C}\text{N} = \frac{L_2(\mathbf{G}\text{CE})}{\mathbf{G}\text{CE} \times \sqrt{c}} \tag{6}$$

$$\hat{\mathbf{x}}_l^{(h,w,c)} = tanh(\gamma \times \mathbf{C}\text{N} + \beta) \times \mathbf{x}_l^{(h,w,c)} + \mathbf{x}_l^{(h,w,c)} \tag{7}$$

where $L_2(\cdot)$ denotes the L2 norm, **G**CE denotes Global Context Embedding, **C**N denotes Channel Normalization, c is the number of channels and $\hat{x}_l^{(h,w,c)}$ denotes the output feature of GCT.

3.3. Gated Multi-Filter Inception Module

For the gated multi-filter inception module upwards of the U-shaped backbone, we design five different forms of filters with different kernel sizes, including the 1 × 1 convolution, the stacking of max pooling and a 1 × 1 convolution, the stacking of a 1 × 1 and two 3 × 3 convolutions, and two strip convolutions respective of the 1 × 9 and 9 × 1 kernel size. The

receptive fields of the five filters cover multi-scales of small, medium and large, which aligns with the intuition that water bodies should be processed at various scales. The number of output channels for each filter is increased to five times that of the input channels, and then the outputs of all filters are concatenated. Instead of directly using the concatenated feature as the output of the inception module, we introduce a GCT module and 1 × 1 dimensionality reduction convolution for adaptively aggregating features by attention and gated mechanisms so that the next stage can better abstract features from different scales simultaneously.

The details of the inception module are depicted in the light-yellow background part of Figure 4; the inception module performs the following operations:

$$\widetilde{\mathrm{x}}_{in}^{(128,128,360)} = \mathrm{Cont}(F_1(x_{in}), F_2(x_{in}), F_3(x_{in}), F_4(x_{in}), F_5(x_{in})) \tag{8}$$

$$\mathrm{x}_{l=1}^{(128,128,72)} = \mathbf{R}\mathrm{eLU}(\mathbf{B}\mathrm{N}(\mathbf{C}\mathrm{onv}_{1\times1}(\mathbf{G}\mathrm{CT}(\widetilde{\mathrm{x}}_{in}^{(128,128,360)})))) \tag{9}$$

where $\mathbf{x}_{in} \in R^{128\times128\times9}$ denotes WIPI data, $F_{i=1,\ldots,5}(\cdot)$, respectively, denote the five designed filters, and $\mathbf{G}\mathrm{CT}(\cdot)$ denotes the Gated Channel Transform, for which the specific implementation is shown in Equations (5)–(7). The function $\mathbf{C}\mathrm{onv}_{1\times1}(\cdot)$ denotes a 1 × 1 convolution with stride = 1 and a number of output channels equal to one fifth that of the input channels.

4. Experiments

4.1. Dataset and Pre-Processing

WIPI multi-modality dataset: In this work, we use three datasets to comprehensively verify the performance of the proposed MFGF-UNet method. Among them, our constructed Water Index and Polarization Information (WIPI) dataset is used as the main dataset, and ablation experiments and comparison experiments are performed on WIPI. The WIPI dataset consists of seven water indexes based on multispectral images, two sets of polarization information from SAR images and the Cloud-Free Label automatically obtained by our proposed algorithm. The construction details of WIPI are given in the aforementioned Section 2. We have released the WIPI dataset (https://github.com/Dataset-RFGroup/WIPI-Dataset (accessed on 12 December 2023)).

Chengdu dataset: The Chengdu dataset (https://github.com/SCoulY/Sentinel-2-Water-Segmentation (accessed on 12 December 2023)) [47] is a remote sensing multispectral image set captured by the Sentinel-2 satellite in April 2018 covering Chengdu city in China and its surrounding area (over 15,000 km^2). The dataset consists of images with five bands sensitive to water reflection and corresponding ground-truth labels. The five bands are: bands 2–4 (Blue–Green–Red (BGR)), band 8 (near-infrared (NIR)) and band 12 (short-wave infrared (SWIR)).

GF2020 dataset: The GF2020 dataset (https://drive.google.com/file/d/1oHyzNfHe_F3MeeUQUoni9dh1LFI_N6RS/view?usp=sharing (accessed on 12 December 2023)) [48] is a high-resolution optical dataset from the GF-2 satellite for Track 5 (automatic water-body segmentation in optical satellite images) from the 2020 Gaofen Challenge. The dataset consists of images with three RGB channels and corresponding ground-truth labels.

An example instance from the WIPI dataset is shown in the aforementioned Figure 3. Figures 5 and 6 are example instances from the Chengdu and GF2020 datasets, respectively.

(**a**) (**b**) (**c**) (**d**) (**e**) (**f**)

Figure 5. Example instances of the five channels and ground truth from Chengdu dataset: (**a**) R, (**b**) G, (**c**) B, (**d**) NIR, (**e**) SWIR and (**f**) GT.

(**a**) (**b**)

Figure 6. Example instances of optical RGB image and ground truth from GF2020 dataset: (**a**) RGB and (**b**) GT.

The image size in each of the three datasets as well as the number of training, validation and test samples are shown in Table 1.

Table 1. Basic information about the three datasets.

Dataset	Image Size	Number of Images in Dataset			
		Total	Train Set	Val Set	Test Set
WIPI	(512, 512, 9)	446	221	85	140
Chengdu	(512, 512, 5)	1681	1008	336	337
GF2020	(492, 492, 3)	1000	600	200	200

"Image size" is in the form of (height, width, channel).

Pre-processing: The pre-processing steps for the three datasets mainly include two key stages: (1) normalization and standardization. We calculate the minimum and maximum values for each dataset and normalize the pixel values by min–max normalization to [0, 1]. And then, we standardize the pixel values by subtracting the mean and then dividing by the standard deviation to ensure a consistent value distribution across the entire dataset. (2) Sample Cropping: In this stage, we used a sliding approach to crop a single image into multiple patches of 128 $\times$ 128 with an overlap of 32 between adjacent patches. For the GF2020 dataset, since its height and width are 492 $\times$ 492, when cropping to 128 $\times$ 128 size, the patches at the rightmost and bottom edges have an overlap greater than 32 pixels. During testing, after the predictions are complete, all patches from a single image are reassembled to the original image size for evaluation.

4.2. Experimental Setup and Evaluation Metrics

For MFGF-UNet, we set the epochs as 200 and the batch size as 32. The initial learning rate lr = 1e-4, and then lr is decayed to 0.99 × lr every 5 epochs. A cross-entropy loss function is employed. The model is trained with the Adam optimizer with momentum 0.99 and weight decay 5e-4. All experiments are conducted on an Ubuntu system with Python 3.7, PyTorch 1.11.0, CUDA 11.3, and an NVIDIA RTX 3090 Ti GPU with 24 GB of memory. The code and trained models are available at GitHub (https://github.com/Code-RFGroup/MFGF-UNet (accessed on 12 December 2023)).

Three widely used measures, accuracy (ACC), F1 score (F1), and Mean Intersection over Union (MIoU), are adopted as criteria to quantitatively evaluate the performance of the different methods; they are calculated as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FN} + \text{FP}} \tag{10}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{11}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{12}$$

$$\text{F1} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \tag{13}$$

$$\text{MIoU} = \frac{1}{\text{k} + 1} \sum_{\text{i}=0}^{\text{k}} \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{14}$$

where k indicates the number of classes. The confusion matrix between the predicted result and a given ground truth is calculated and consists of true positives (TPs), true negatives (TNs), false positives (FPs) and false negatives (FNs). Positive and negative represent water and non-water, respectively. For ACC, MIoU and F1, a higher score means better performance. For all experiments, each algorithm runs three times independently, and the average results are given. In addition to ACC, MIoU and F1 to evaluate the performance, we measure the complexity of different models using the number of parameters (Param) and one billion floating-point operations (GFLOPs).

4.3. Ablation Studies

In order to explore the influence of different factors on MFGF-UNet's performance, we conduct ablation studies on the WIPI dataset and discuss four factors below.

AS1—Effect of U-shaped encoder–decoder (Enc-Dec): In the model with Enc-Dec, we replaced the GCT skip connection with a direct skip connection and removed the gated multi-filter inception module. Only the U-shaped encoder–decoder part is retained, and the input of the encoder is nine channels of WIPI data.

AS2—Effect of GCT skip connection (GCT): In the model with Enc-Dec and GCT, we only removed the gated multi-filter inception module.

AS3—Effect of gated multi-filter inception (Inception): In the model with Enc-Dec and Inception, we only replaced the GCT skip connection with a direct skip connection.

Table 2 shows that the above three factors influence both the segmentation results and the complexity. In AS1, which removes both improvement modules, the F1 score reaches 91.26%, which is a decrease of 0.65% compared to MFGF-UNet. This result shows that our improvements are effective. In AS2 and AS3, where the GCT or the Inception module are individually removed, there is a decrease of 0.12% or 0.48%, respectively. This indicates the proposed Inception module contributes more than the GCT module.

Table 2. Ablation studies of three factors on WIPI dataset.

Method	Enc-Dec	Inception	GCT	ACC (%)	MIoU (%)	F1 (%)	Param (M)	GFLOPs
AS1	✓			98.09 (±0.04)	83.92 (±0.31)	91.26 (±0.18) (↓0.65)	27.36	15.74
AS2	✓	✓		98.21 (±0.01)	84.82 (±0.15)	91.79 (±0.08) (↓0.12)	27.53	21.04
AS3	✓		✓	98.13 (±0.03)	84.21 (±0.28)	91.43 (±0.16) (↓0.48)	27.37	15.75
MFGF-UNet	✓	✓	✓	**98.22** (±0.01)	**85.03** (±0.18)	**91.91** (±0.11)	27.53	21.05

↓: The degree of F1 decrease is based on comparison to MFGF-UNet.

AS4—Effect of different modalities: To analyze the impact of the different modalities on MFGF-UNet's performance, we modified MFGF-UNet by using the different modality data as input while keeping the other parts the same. There are a total of four modalities: SAR, MSI, WI and WIPI, where SAR denotes the polarization information of VV and VH from SAR, MSI denotes the spectrum of 13 bands from MSI, WI denotes the seven extracted water indexes from raw MSI, and WIPI denotes the water index and polarization information, which includes VV, VH and the seven water indexes. For input images of all modalities, Cloud-Free Label is used as the ground truth to eliminate the impact of different labels. From the results given in Table 3, it can be seen that the segmentation performance of SAR is relatively poor, with ACC, MIoU and F1 scores reaching only 95.22%, 63.18% and 77.68%, respectively. The segmentation results using MSI data are far better than those of SAR. However, given that MSI data have the largest band dimension, the model requires the highest amount of parameters and calculations, in which Param and GFLOPs reached 57.43 and 43.79, respectively. It is worth noting that the results using WI extracted from MSI data are better than those using original MSI data. The improvement of WI on average ACC, MIoU and F1 are 0.24%, 1.59% and 0.97%, respectively. Additionally, using WI reduces the dimensions of the original data, which results in a significant decrease in both Param and GFLOPs. Our WIPI data not only leverage the extracted water index but also fuse the multi-modality of hyperspectral and SAR data, thereby achieving superior results, with the average ACC, MIoU and F1 reaching 98.22%, 85.03% and 91.91%, respectively.

Table 3. Ablation studies of different modalities on WIPI dataset.

Method	Modality	Channel No.	ACC (%)	MIoU (%)	F1 (%)	Param (M)	GFLOPs
MFGF-UNet	SAR	2	95.52 (±0.06)	63.18 (±0.60)	77.68 (±0.79) (↓14.23)	**1.36**	**1.04**
	MSI	13	97.55 (±0.04)	79.79 (±0.29)	88.76 (±0.18) (↓3.15)	57.43	43.79
	WI	7	97.79 (±0.05)	81.38 (±0.39)	89.73 (±0.24) (↓2.18)	16.65	12.71
	WIPI	9	**98.22** (±0.01)	**85.03** (±0.18)	**91.91** (±0.11)	27.53	21.04

↓: The degree of F1 decrease is in comparison to WIPI.

Moreover, visual comparisons between the different modalities are shown in Figure 7, with significant differences highlighted by red lines. It can be seen that if only single-modal SAR or MSI data are used, it is easy to lose the details of water bodies such as slender rivers. At the same time, using only WI data may lead to false detections with an extended range. Conversely, our WIPI can segment water bodies in a more consistent and complete manner.

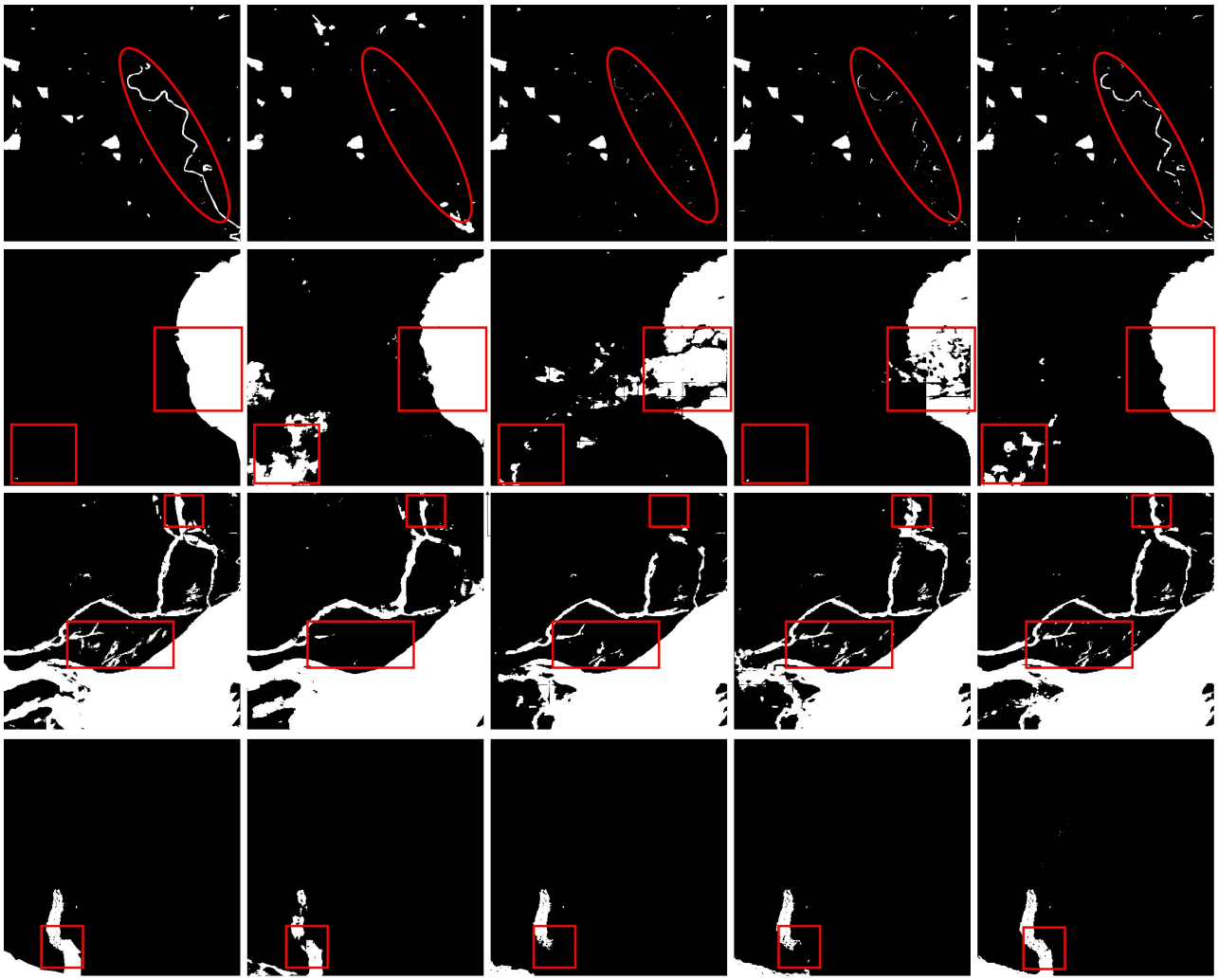


Figure 7. Visual comparisons between different modalities: (**a**) GT, (**b**) SAR, (**c**) MSI, (**d**) WI and (**e**) WIPI.

4.4. Comprehensive Comparison with Other Methods

The proposed MFGF-UNet is compared with six other state-of-the-art approaches: FCN-8s [12], U-Net [14], DeepLabV3+ [19], BASNet [20], DUPNet [1] and SegNeXt [24]. For all the other comparison methods, the parameters are set according to the corresponding original work. All methods take 128 $\times$ 128 $\times C$ as the input size, where C is the number of channels corresponding to different datasets.

To verify the effectiveness of the constructed WIPI dataset and the proposed MFGF-UNet model, we trained the seven models above on the two datasets Sen1Floods11 and WIPI. The Sen1Floods11 dataset consists of images with 15 channels (VV, VH and 13 multispectral bands) and Manual Label-MS. The WIPI dataset consists of images with nine channels (VV, VH and seven water indexes) and the Cloud-Free Label. Along with ACC, MIoU and F1 performance, we measure the complexity of different models using Param and GFLOPs. All results on the two datasets are summarized in Table 4.

Table 4. Segmentation quantitative results and complexity of seven approaches on Sen1Floods11 and WIPI datasets.

Method	Dataset	Channel No.	ACC (%)	MIoU (%)	F1 (%)	Param (M)	GFLOPs
FCN-8s [12]	Sen1Floods11	15	96.04 (±0.28)	68.12 (±1.91)	81.02 (±1.36)	46.25	2.01
	WIPI	9	96.83 (±0.06)	73.95 (±0.62)	85.02 (±0.41) (↑4.00)	46.24	**1.98**
U-Net [14]	Sen1Floods11	15	96.53 (±0.46)	72.84 (±2.47)	84.26 (±1.64)	49.89	26.91
	WIPI	9	98.09 (±0.04)	84.07 (±0.21)	91.35 (±0.12 (↑7.09))	49.89	26.69
DeepLabv3+ [19]	Sen1Floods11	15	96.43 (±0.34)	71.88 (±1.91)	83.62 (±1.28)	6.43	2.10
	WIPI	9	97.37 (±0.03)	78.22 (±0.33)	87.78 (±0.21) (↑4.16)	**6.43**	2.09
BASNet [20]	Sen1Floods11	15	97.30 (±0.06)	78.14 (±0.17)	87.73 (±0.10)	87.08	31.99
	WIPI	9	97.91 (±0.07)	82.92 (±0.45)	90.66 (±0.27) (↑2.93)	87.07	31.93
DUPNet [1]	Sen1Floods11	15	95.16 (±0.24)	66.06 (±1.18)	79.55 (±0.86)	49.87	69.20
	WIPI	9	98.01 (±0.07)	83.23 (±0.66)	90.85 (±0.40) (↑11.30)	49.86	69.17
SegNeXt [24]	Sen1Floods11	15	96.67 (±0.10)	72.15 (±1.08)	83.81 (±0.73)	27.54	2.02
	WIPI	9	97.11 (±0.17)	76.38 (±1.21)	86.60 (±0.78) (↑2.79)	27.54	2.01
MFGF-UNet (ours)	Sen1Floods11	15	96.11 (±0.53)	70.52 (±2.64)	82.68 (±1.84)	76.46	58.37
	WIPI	9	**98.22** (±0.01)	**85.03** (±0.18)	**91.91** (±0.11) (↑9.23)	27.53	21.05

↑: The degree of F1 increase is in comparison to the corresponding method on Sen1Floods11.

The experimental results in Table 4 show that when Sen1Floods11 is replaced with our WIPI, the segmentation performance of all seven approaches significantly improves, among which, the top three highest improvements are DUPNet, our MFGF-UNet and UNet, for which the average F1 scores increased by 11.30%, 9.23% and 7.09%, respectively. Meanwhile, given that the channel dimension is reduced from 15 to 7, the amount of parameters and calculations for all models have been reduced. Specifically for our MFGF-UNet, using the WIPI dataset can reduce Param and GFLOPs by 48.93 and 37.32, respectively.

To further compare the advantages of WIPI over Sen1Floods11, we visualize the ground truth of the two datasets and the predicted results of U-Net, DUPNet, SegNeXt and our MFGF-UNet in Figure 8. Among the six comparison methods, DUPNet and UNet achieve the second and third best results after our method, and SegNeXt is the latest method. It can be observed that utilizing the WIPI dataset helps to eliminate some of the interference in the Sen1Floods11 dataset, such as shadows and cloud cover, and thereby provide better segmentation results, especially for blocky, stripe-like, or patchy water bodies.

In addition to the improvement brought by the constructed WIPI dataset, the results in Table 4 also demonstrate that our method achieves the best performance against six state-of-the-art approaches, with an ACC, MIoU and F1 of 98.22%, 85.03% and 91.91%, respectively. U-Net achieves an F1 score of 91.35% on the WIPI dataset—second only to our MFGF-UNet—but Param and GFLOPs of U-Net are 22.36 and 5.64 higher, respectively, than for our method. Although Deeplabv3+ has the smallest model size, its segmentation performance on the WIPI dataset is much lower than MFGF-UNet, with a 4.13% drop in F1 score.

Figure 9 shows the segmentation results and corresponding zoom-in details on the WIPI dataset using MFGF-UNet and the six comparison methods. For the BASNet and DUPNet methods, many details, such as slender rivers, are missed, indicating that these models can only generate high-confidence predictions for large water areas. The segmentation results of SegNeXt are not precise enough because many small water bodies, particularly those exhibiting jagged edges, are missed. Compared to other models, our method produces clearer and more accurate prediction maps for large water bodies as well as for slender rivers.

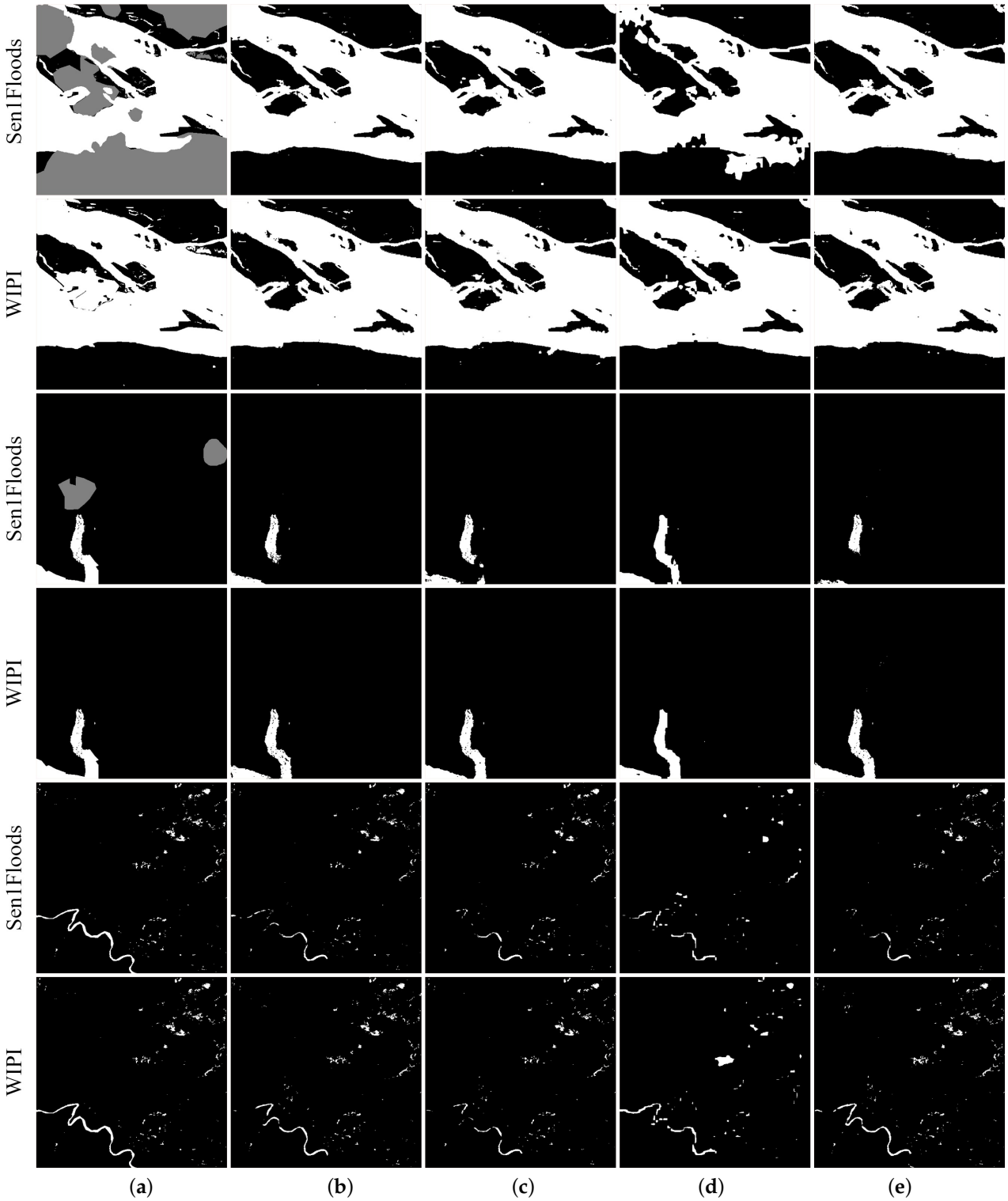


Figure 8. Visual comparisons of segmentation results from four approaches on Sen1Floods11 and WIPI datasets: (**a**) GT, (**b**) U-Net, (**c**) DUPNet, (**d**) SegNeXt and (**e**) MFGF-UNet.

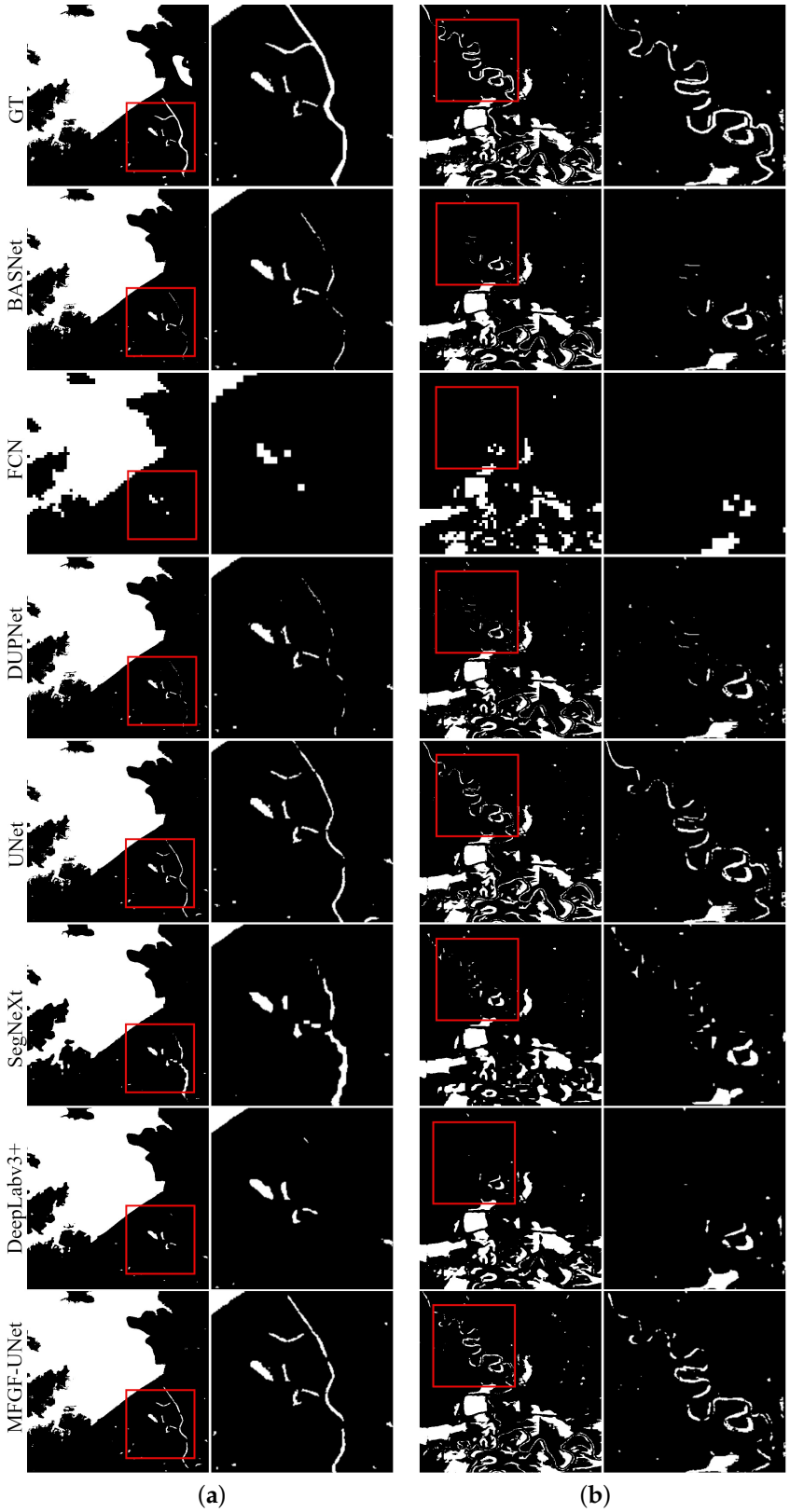


Figure 9. Visual comparisons of segmentation results and corresponding zoom-in details from seven approaches on WIPI dataset: (**a**) Test Sample A and (**b**) Test Sample B.

4.5. Evaluation of Model Robustness

Compared with the WIPI dataset, the Chengdu and GF2020 datasets have different modalities: the former provides five-channel multispectral images, and the latter provides

high-resolution optical images. Moreover, these two datasets were captured from different sensors at different times while covering different ground ranges. Therefore, we conducted experiments of all methods on Chengdu and GF2020 to evaluate the robustness of the proposed MFGF-UNet model. The experimental results are shown in Table 5. It can be observed that on both the Chengdu and GF2020 datasets, MFGF-UNet outperforms the FCN, UNet, Deeplabv3+ and SegNeXt models in performance. Although on the Chengdu dataset, MFGF-UNet has a 0.66% lower F1 score than BASNet and on the GF2020 dataset, it has 0.91% and 1.02% lower F1 scores than BASNet and DUPNet, respectively, MFGF-UNet exhibits obvious lower complexity than BASNet and DUPNet. It is noteworthy that DUPNet only performs well on the GF2020 dataset, and the results on the Chengdu dataset are not very good.

Table 5. Segmentation quantitative results and complexities of seven approaches on Chengdu and GF2020 datasets.

Method	Dataset	ACC (%)	MIoU (%)	F1 (%)	Params (M)	GFLOPs
FCN-8s [12]	Chengdu	98.56 (±0.01)	45.78 (±0.76)	62.71 (±0.64)	46.25	1.96
	GF2020	93.12 (±0.01)	70.83 (±0.22)	82.93 (±0.15)	46.25	1.95
U-Net [14]	Chengdu	98.91 (±0.01)	58.02 (±0.61)	73.42 (±0.49)	49.88	26.54
	GF2020	93.69 (±0.11)	72.72 (±0.77)	84.10 (±0.51)	49.88	26.46
DeepLabv3+ [19]	Chengdu	98.86 (±0.01)	55.96 (±0.50)	71.76 (±0.41)	6.43	2.09
	GF2020	92.56 (±0.16)	68.86 (±0.90)	81.55 (±0.63)	6.43	2.09
BASNet [20]	Chengdu	98.90 (±0.01)	**59.69** (±0.12)	**74.75** (±0.09)	87.07	31.89
	GF2020	93.45 (±0.16)	74.11 (±0.35)	85.13 (±0.23)	87.07	31.87
DUPNet [1]	Chengdu	98.87 (±0.01)	56.87 (±0.72)	72.50 (±0.58)	49.86	69.16
	GF2020	**93.79** (±0.06)	**74.27** (±0.09)	**85.23** (±0.06)	49.86	69.15
SegNeXt [24]	Chengdu	98.84 (±0.01)	55.29 (±0.13)	71.21 (±0.11)	27.54	2.01
	GF2020	93.06 (±0.09)	72.03 (±0.26)	83.74 (±0.17)	27.54	2.01
MFGF-UNet	Chengdu	**98.91** (±0.01)	58.85 (±0.10)	74.09 (±0.08)	8.50	6.49
	GF2020	93.63 (±0.09)	72.73 (±0.45)	84.21 (±0.30)	3.06	2.36

Figures 10 and 11 show the segmentation results and corresponding zoom-in details for the Chengdu and GF2020 datasets using MFGF-UNet and six comparison methods. The Chengdu dataset comprises a greater number of slender water areas, while the GF2020 dataset includes more extensive and continuous water areas. The DUPNet method achieved the highest F1 score on the GF2020 dataset but exhibited relatively lower performance on the Chengdu dataset. This result can be reasonably explained by the fact that, as seen in the figures, DUPNet is not very good at detecting small water bodies but has better performance on large water areas. The visualized results further illustrate that our proposed MFGF-UNet excels at the segmentation of small water bodies and simultaneously proves effective at segmenting large water areas.

The comprehensive comparison of the segmentation performance and complexity of the seven models on the three datasets is shown in Figure 12, where the x-axis represents GFLOPs, the y-axis represents F1, and the radius of the circle represents the parameter size of each model. Being closer to the upper left corner and having a smaller radius indicate better effectiveness and efficiency of the algorithm. The above results and analysis illustrate that our MFGF-UNet has achieved favorable performance with lower complexity and better robustness on the three datasets.

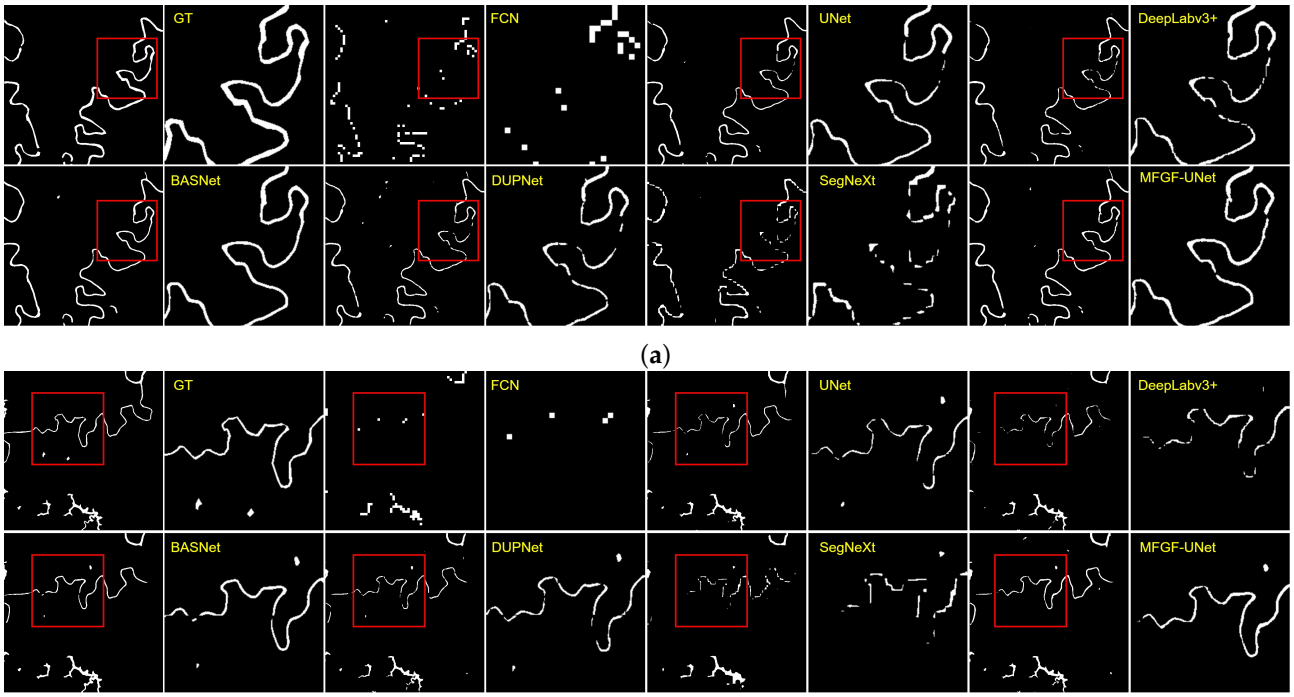


Figure 10. Visual comparisons of segmentation results and corresponding zoom-in detail from seven approaches on Chengdu dataset: (**a**) Test Sample A and (**b**) Test Sample B.

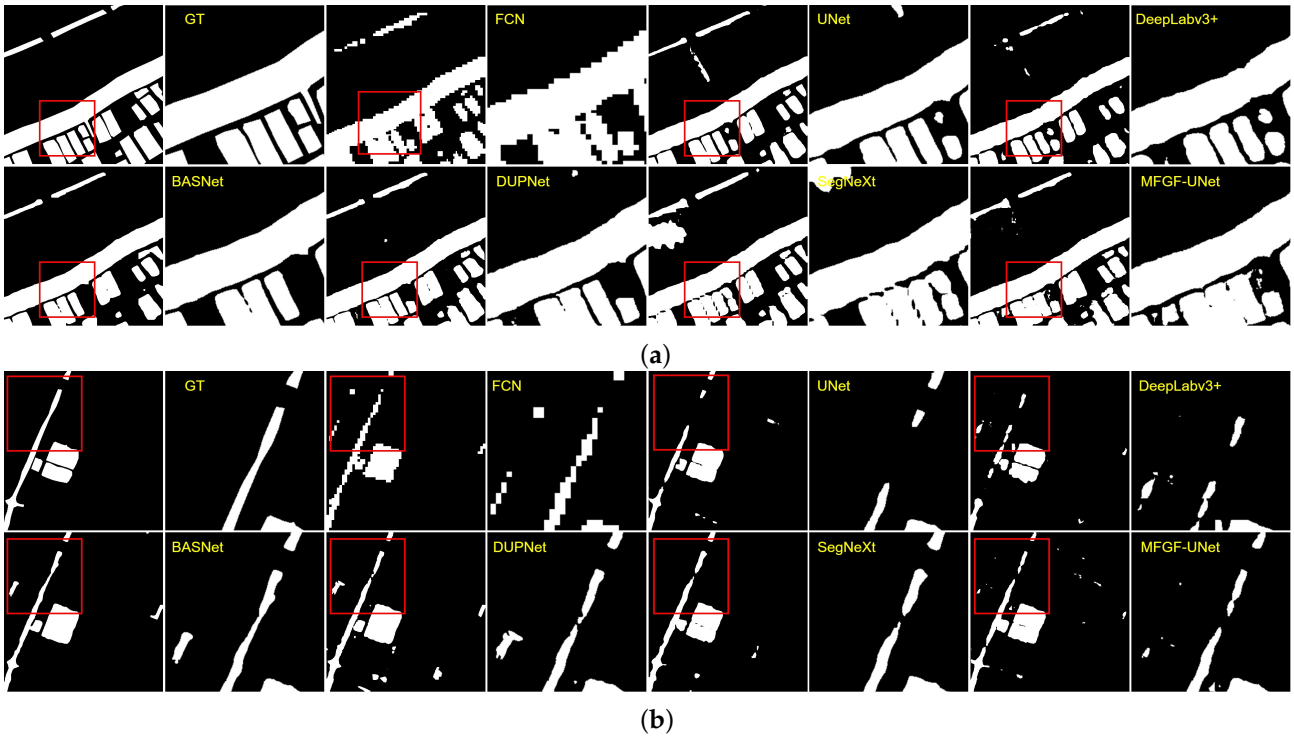


Figure 11. Visual comparisons of segmentation results and corresponding zoom-in detail from seven approaches on GF2020 dataset: (**a**) Test Sample A and (**b**) Test Sample B.

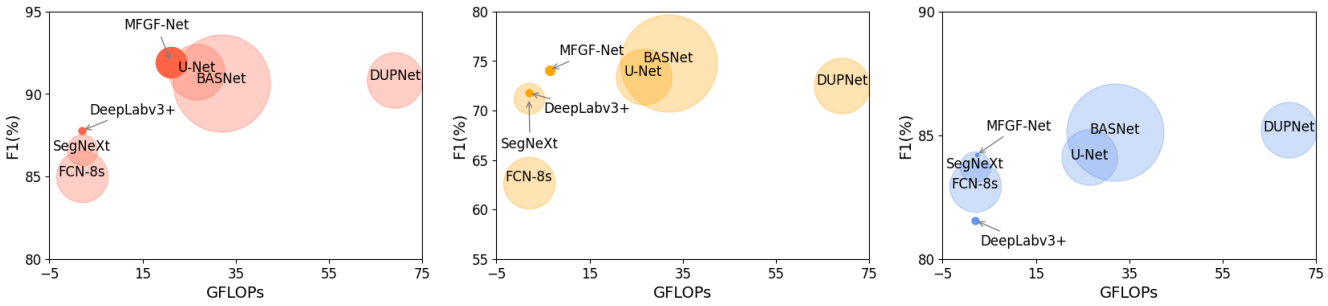


Figure 12. Param and GFLOPs vs. F1 of seven approaches on WIPI, Chengdu and GF2020 datasets: (**a**) WIPI, (**b**) Chengdu and (**c**) GF2020.

5. Discussion

As shown in the experiments in this work, the use of different modalities or different ways of using the modalities has a great impact on water segmentation. For example, in Table 3, for the same ground region and the same model, the F1 from lowest to highest is SAR, MSI, WI and WIPI at 77.68%, 88.76%, 89.73% and 91.91%, respectively. Since the polarization information of SAR images cannot distinguish water, water-like surfaces and noise-like spots well, its F1 is the lowest. MSI provides 13 band of information with high resolution and less noise, so its F1 is 11.08% higher than that of SAR. As expected, multi-modal WIPI achieves the best results. For all seven methods, the average F1 on WIPI is 89.17%, as seen in Table 4. In Table 5, the average F1 scores on Chengdu (multispectral images with five bands) and GF2020 (high-resolution optical images) are 71.49% and 83.84%, respectively. Even if the differences in the regions and the sizes of the datasets are put aside, it also can be found that our constructed WIPI dataset performs best. It is worth noting that the F1 of MSI is 0.97% lower than that of WI, as seen in Table 3. At the same time, as seen in Table 4, the average F1 on Sen1Floods11 is 83.24%, which is 5.93% lower than that of WIPI. Even though the raw data for MSI and WI are both multispectral images with 13 bands and the raw data for Sen1Floods11 and WIPI are the same multispectral and SAR images, how different modalities are analytically utilized changes the performance of the model. The above results also verify the effectiveness of the water index we extracted.

Another reason for the lower performance on Sen1Floods11 is that Manual Label-MS contains a large number of invalid areas, resulting in multi-modality results that are lower than those of single-modality using the Cloud-Free Label: for examples, for the MFGF-UNet model, the F1 scores of single-modal MSI and WI are 88.76% and 89.73%; however, the F1 score of multi-modal Sen1Floods11 is only 82.68%.

Utilization of the water index not only enhances the water information but also reduces the data dimensionality: thus decreasing the overall complexity of the model. For example, on the WIPI (nine channels), Chengdu (five channels) and GF2020 (three channels) datasets, the proposed MFGF-UNet model achieves F1 scores of 91.91%, 74.10% and 84.21% with GFLOPs of 21.05, 6.49 and 2.36, respectively. As the input data dimensions are reduced, our model complexity is significantly reduced. This trend emanates from the meticulous design of our encoder, wherein the output feature channel count of the first network layer is eight times that of the input feature channels, and the subsequent network layers have an output channel count twice those of the input channels. It is evident that our model can dynamically adjust with a high degree of adaptability to accommodate varying scales of input data. Compared with the U-Net, which has better performance, F1 scores reach 91.35%, 73.42% and 84.10% with GFLOPs of 26.69, 26.54 and 26.46, respectively. As the input data's dimensions are reduced, the complexity of the U-Net model also decreases (note that the unit of measurement is one billion). Even while the GFLOPs are only 24% and 9% of those of U-Net, our MFGF-UNet achieves slightly better segmentation performance on the Chengdu and GF2020 datasets, respectively. These results illustrate that, besides

the design of the architecture, focusing on the data level can also reduce the computational cost of the model.

6. Conclusions

In this work, we proposed an effective and efficient MFGF-UNet model for water area segmentation in remote sensing. We first analyzed the task's current challenges and then designed corresponding strategies to address these issues. Concretely, we constructed a WIPI multi-modality dataset to overcome the shortcomings of single-modal data and to enhance water information while reducing the channel dimensionality of the data, the Cloud-Free Label provided in the dataset can also effectively alleviate the problem of labeled sample scarcity. For the issue that a single form or uniform kernel size cannot handle the variety of sizes and shapes of water bodies, we proposed the gated multi-filter inception (GMF-Inception) module. Moreover, we utilized the attention mechanism by introducing the GCT skip connection and integrating GCT into GMF-Inception to further improve model performance. Extensive experiments on three various benchmarks, including the WIPI (water index and SAR images), Chengdu (multispectral images) and GF2020 (optical images) datasets demonstrated that our MFGF-UNet model achieves favorable performance with lower complexity and better robustness against six competing approaches.

Based on these improvements, in the future, we or other researchers can conduct in-depth research from the following aspects: Firstly, consider the use of convolutional kernels with different scales or shapes for feature extraction based on the size or shape of the target objects. Secondly, despite the automatic feature extraction capability of convolutional neural networks, effectively leveraging prior knowledge for data pre-processing and early fusion can help reduce redundancy and misleading information: thereby improving the robustness of the model. Meanwhile, we hope the release of a new dataset can help researchers to develop and verify their frameworks for water area segmentation and facilitate more breakthroughs in this field.

Author Contributions: Methodology and investigation, R.W. and C.Z.; resources, software, and writing—original draft preparation, C.Z.; writing—review and editing, R.W.; validation, C.C.; visualization, H.H.; project administration, L.J.; funding acquisition, W.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (No. 62176196 and 62077038), the Key Research and Development Program of Shaanxi (No. 2023-YBNY-218 and 2023-YBNY-284), the Open Project Program of the State Key Laboratory of Multimodal Artificial Intelligence Systems (No. 202200021), the Shaanxi Provincial Water Conservancy Fund Project (No. SLKJ2024-06), the Research Project of Shaanxi Coal Geology Group Co., LTD. (No. SMDZ-2023CX-14).

Data Availability Statement: The WIPI Dataset used in this article can be downloaded at https://github.com/Dataset-RFGroup/WIPI-Dataset (accessed on 12 December 2023).

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

Appendix A

Appendix A.1. Description of Sen1Floods11 Dataset

The Sen1Floods11 dataset consists of raw Sentinel-1 SAR images (Sen1-SAR) and raw Sentinel-2 multispectral images (Sen2-MS), where Sen1-SAR includes the two VV and VH bands, and Sen2-MS includes 13 bands (1–8, 8A, 9–12) across all spectra. The values, formats and band descriptions for Sen1Floods11 are shown in Appendix A, Table A1.

Table A1. Sen1Floods11 dataset description.

Modality	Value	Format	Bands
SAR	Unit:dB	GeoTIFF 512 × 512 2 bands Float32	0:VV 1:VH
MSI	Unit: TOA reflectance (scaled by 10,000)	GeoTIFF 512 × 512 13 bands UInt16	0: B1 (Coastal) 1: B2 (Blue) 2: B3 (Green) 3: B4 (Red) 4: B5 (RedEdge-1) 5: B6 (RedEdge-2) 6: B7 (RedEdge-3) 7: B8 (NIR) 8: B8A (Narrow NIR) 9: B9 (Water Vapor) 10: B10 (Cirrus) 11: B11 (SWIR-1) 12: B12 (SWIR-2)

Appendix A.2. Description of the Seven Water Indexes Used

In this work, we analyze the characteristics of the water index and extract seven different water indexes from 13 bands of raw MSI data for constructing our WIPI dataset. The formulas and descriptions of the seven extracted water indexes are shown in Table A2.

Table A2. The formulas and descriptions of the seven water indexes.

Index	Name	Formula	Description
1	Normalized Difference Vegetation Index (NDVI) [40]	(B8 − B4)/(B8 + B4)	High values indicate dense tree canopies, while low or negative values indicate urban areas or water bodies.
2	Normalized Difference Moisture Index (NDMI) [41]	(B8A − B11)/(B8A + B11)	Vegetation with higher values is relatively moist.
3	Normalized Difference Water Index (NDWI) [26]	(B3 − B8)/(B3 + B8)	Highlights water bodies, suppresses vegetation information; susceptible to the influence of object shadows.
4	Modified Normalized Difference Water Index (MNDWI) [27]	(B3 − B11)/(B3 + B11)	Eliminates the influence of buildings and land, highlights water bodies; easily affected by shadows of objects.
5	Automated Water Extraction Index (Non-Shadow, $AWEI_{NS}$) [42]	(4B3 − 4B11)/(0.25B8 + 2.75B12)	Removes black buildings that are easily misclassified as water bodies.
6	Automated Water Extraction Index (Shadow, $AWEI_S$) [42]	B2 + 2.5B3 − 1.5 (B8 + B11) − 0.25B12	Suitable for scenes with a high amount of shadow.
7	Linear Discriminant Analysis Water Index (LDAWI) [43]	1.7204 + 171B3 + 3B4 − 70B8 − 45B11 − 71B12	Suitable for scenes where there is a large difference in spectral distribution between non-water and water bodies.

References

1. Liu, Z.; Chen, X.; Zhou, S.; Yu, H.; Guo, J.; Liu, Y. DUPnet: Water Body Segmentation with Dense Block and Multi-Scale Spatial Pyramid Pooling for Remote Sensing Images. *Remote Sens.* **2022**, *14*, 5567. [CrossRef]
2. Konapala, G.; Kumar, S.V.; Ahmad, S.K. Exploring Sentinel-1 and Sentinel-2 diversity for flood inundation mapping using deep learning. *ISPRS J. Photogramm. Remote Sens.* **2021**, *180*, 163–173. [CrossRef]
3. Li, Y.; Dang, B.; Zhang, Y.; Du, Z. Water body classification from high-resolution optical remote sensing imagery: Achievements and perspectives. *ISPRS J. Photogramm. Remote Sens.* **2022**, *187*, 306–327. [CrossRef]
4. Liu, J.; Wang, Y. Water Body Extraction in Remote Sensing Imagery Using Domain Adaptation-Based Network Embedding Selective Self-Attention and Multi-Scale Feature Fusion. *Remote Sens.* **2022**, *14*, 3538. [CrossRef]
5. Shen, X.; Wang, D.; Mao, K.; Anagnostou, E.; Hong, Y. Inundation extent mapping by synthetic aperture radar: A review. *Remote Sens.* **2019**, *11*, 879. [CrossRef]
6. Longfei, S.; Zhengxuan, L.; Fei, G.; Min, Y. A review of remote sensing image water extraction. *Remote Sens. Nat. Resour.* **2021**, *33*, 9–11.
7. Cao, M.; Mao, K.; Shen, X.; Xu, T.; Yan, Y.; Yuan, Z. Monitoring the spatial and temporal variations in the water surface and floating algal bloom areas in Dongting Lake using a long-term MODIS image time series. *Remote Sens.* **2020**, *12*, 3622. [CrossRef]
8. Shetty, S.; Gupta, P.K.; Belgiu, M.; Srivastav, S. Assessing the effect of training sampling design on the performance of machine learning classifiers for land cover mapping using multi-temporal remote sensing data and google earth engine. *Remote Sens.* **2021**, *13*, 1433. [CrossRef]
9. Razaque, A.; Ben Haj Frej, M.; Almi'ani, M.; Alotaibi, M.; Alotaibi, B. Improved support vector machine enabled radial basis function and linear variants for remote sensing image classification. *Sensors* **2021**, *21*, 4431. [CrossRef]
10. Li, A.; Fan, M.; Qin, G.; Xu, Y.; Wang, H. Comparative analysis of machine learning algorithms in automatic identification and extraction of water boundaries. *Appl. Sci.* **2021**, *11*, 10062. [CrossRef]

11. Acharya, T.D.; Subedi, A.; Lee, D.H. Evaluation of machine learning algorithms for surface water extraction in a Landsat 8 scene of Nepal. *Sensors* **2019**, *19*, 2769. [CrossRef] [PubMed]
12. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
13. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
14. 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: Cham, Switzerland, 2015; pp. 234–241.
15. Cui, M.; Li, K.; Chen, J.; Yu, W. CM-Unet: A novel remote sensing image segmentation method based on improved U-Net. *IEEE Access* **2023**, *11*, 56994–57005. [CrossRef]
16. 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]
17. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic image segmentation with deep convolutional Nets and fully connected CRFs. *arXiv* **2014**, arXiv:1412.7062.
18. 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]
19. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
20. Qin, X.; Zhang, Z.; Huang, C.; Gao, C.; Dehghan, M.; Jagersand, M. Basnet: Boundary-aware salient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7479–7489.
21. 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, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
22. 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, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
23. 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.
24. Guo, M.H.; Lu, C.Z.; Hou, Q.; Liu, Z.; Cheng, M.M.; Hu, S.M. Segnext: Rethinking convolutional attention design for semantic segmentation. *arXiv* **2022**, arXiv:2209.08575.
25. Akiva, P.; Purri, M.; Dana, K.; Tellman, B.; Anderson, T. H_2O-Net: Self-supervised flood segmentation via adversarial domain adaptation and label refinement. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2021; pp. 111–122.
26. 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]
27. Xu, H. Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery. *Int. J. Remote Sens.* **2006**, *27*, 3025–3033. [CrossRef]
28. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
29. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
30. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 764–773.
31. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part I 16; Springer: Cham, Switzerland, 2020; pp. 213–229.
32. Yuan, Y.; Huang, L.; Guo, J.; Zhang, C.; Chen, X.; Wang, J. Ocnet: Object context network for scene parsing. *arXiv* **2018**, arXiv:1809.00916.
33. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3146–3154.
34. Oktay, O.; Schlemper, J.; Folgoc, L.L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N.Y.; Kainz, B.; et al. Attention U-Net: Learning where to look for the pancreas. *arXiv* **2018**, arXiv:1804.03999.
35. Guan, Q.; Huang, Y.; Zhong, Z.; Zheng, Z.; Zheng, L.; Yang, Y. Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification. *arXiv* **2018**, arXiv:1801.09927.
36. Xie, S.; Liu, S.; Chen, Z.; Tu, Z. Attentional shapecontextnet for point cloud recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4606–4615.
37. Su, W.; Zhu, X.; Cao, Y.; Li, B.; Lu, L.; Wei, F.; Dai, J. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv* **2019**, arXiv:1908.08530.

38. Xu, T.; Zhang, P.; Huang, Q.; Zhang, H.; Gan, Z.; Huang, X.; He, X. AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1316–1324.
39. Bonafilia, D.; Tellman, B.; Anderson, T.; Issenberg, E. Sen1Floods11: A georeferenced dataset to train and test deep learning flood algorithms for sentinel-1. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 13–19 June 2020; pp. 210–211.
40. Rouse, J.W.; Haas, R.H.; Schell, J.A.; Deering, D.W. Monitoring vegetation systems in the Great Plains with ERTS. *NASA Spec. Publ.* **1974**, *351*, 309.
41. Jackson, R.D. Remote sensing of biotic and abiotic plant stress. *Annu. Rev. Phytopathol.* **1986**, *24*, 265–287. [CrossRef]
42. Feyisa, G.L.; Meilby, H.; Fensholt, R.; Proud, S.R. Automated Water Extraction Index: A new technique for surface water mapping using Landsat imagery. *Remote Sens. Environ.* **2014**, *140*, 23–35. [CrossRef]
43. Fisher, A.; Flood, N.; Danaher, T. Comparing Landsat water index methods for automated water classification in eastern Australia. *Remote Sens. Environ.* **2016**, *175*, 167–182. [CrossRef]
44. 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.
45. Glorot, X.; Bordes, A.; Bengio, Y. Deep sparse rectifier neural networks. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, Ft. Lauderdale, FL, USA, 11–13 April 2011; JMLR Workshop and Conference Proceedings; pp. 315–323.
46. Yang, Z.; Zhu, L.; Wu, Y.; Yang, Y. Gated channel transformation for visual recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11794–11803.
47. Yuan, K.; Zhuang, X.; Schaefer, G.; Feng, J.; Guan, L.; Fang, H. Deep-learning-based multispectral satellite image segmentation for water body detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 7422–7434. [CrossRef]
48. Sun, X.; Wang, P.; Yan, Z.; Diao, W.; Lu, X.; Yang, Z.; Zhang, Y.; Xiang, D.; Yan, C.; Guo, J.; et al. Automated high-resolution earth observation image interpretation: Outcome of the 2020 Gaofen challenge. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 8922–8940. [CrossRef]