remote sensing

MDPI

Article

Floodwater Extraction from UAV Orthoimagery Based on a Transformer Model

Zhihong Wu [1], Zhe Dong [1], Kun Yang [1], Qingjie Liu [2] and Wei Wang [1,*]

[1] National Disaster Reduction Center of China, The Ministry of Emergency Management of China, Beijing 100124, China; wzh_work0530@163.com (Z.W.); dz1227@foxmail.com (Z.D.); ykvolcano@gmail.com (K.Y.)
[2] Hangzhou Innovation Institute, Beihang University, Beijing 100191, China; qingjie.liu@buaa.edu.cn
* Correspondence: wangwei@ndrcc.org.cn; Tel.: +010-52811199

Abstract: In recent years, remote sensing has experienced a significant transformation due to rapid advancements in deep learning technology, which have greatly outpaced traditional methodologies. This integration has attracted substantial interest within the academic community. To address the complex challenges of extracting data on intricate water bodies during disaster scenarios, this study developed a post-disaster floodwater body dataset and an enhanced multi-scale transformer model architecture. Through end-to-end training, the precision of the model in extracting floodwater contours has been significantly improved. Additionally, by utilizing the vast amounts of unannotated data in remote sensing through an unsupervised pre-training task, the model's backbone network has been fortified, greatly enhancing its performance in remote sensing applications. Experimental analyses have shown that the multi-scale transformer-based algorithm for floodwater contour extraction proposed in this study is not only widely applicable but also excels in delivering precise segmentation results in complex environments. This refined approach ensures that the model adeptly handles the intricacies of floodwater body delineation, providing a robust solution for accurate extraction, even in disaster-stricken areas. This innovation represents a substantial leap forward in remote sensing, offering valuable insights and tools for disaster management and environmental monitoring.

Keywords: disaster prevention and mitigation; flood disaster; disaster remote sensing; deep learning; Segformer; floodwater body dataset

Citation: Wu, Z.; Dong, Z.; Yang, K.; Liu, Q.; Wang, W. Floodwater Extraction from UAV Orthoimagery Based on a Transformer Model. *Remote Sens.* **2024**, *16*, 4052. https://doi.org/10.3390/rs16214052

Academic Editor: Gilberto Camara

Received: 12 September 2024
Revised: 22 October 2024
Accepted: 29 October 2024
Published: 31 October 2024

1. Introduction

In remote sensing imagery, water bodies such as rivers, lakes, and flood areas are crucial targets for automatic interpretation. These areas form essential components of basic geographic information and play a significant role in interpreting remote sensing data. With the rapid development of UAV technology in recent years, higher-resolution orthoimages have become widely used across various fields, particularly in disaster assessment, mitigation, and rescue efforts, where they are of great importance [1–4]. Water body data extraction technology based on UAV orthoimagery can provide critical support for disaster situation estimation and emergency rescue missions by identifying flood coverage areas and analyzing the extent of flooding.

Deep learning technology has recently flourished, surpassing traditional methods, and has been increasingly applied to remote sensing, drawing significant attention from academia and industry [5–11]. In 2014, Shelhamer E et al. introduced the groundbreaking fully convolutional network (FCN) [6], which performs end-to-end pixel-level classification and significantly enhances the performance of semantic segmentation tasks. FCN leverages the parallel computing power of GPUs to accelerate convolutional operations and can adapt to images of various sizes and ratios, thereby expanding the application of neural networks in semantic segmentation. Building on FCN, Ronneberger et al. [7] later proposed the U-Net model, which utilizes an "encoder–decoder structure" to capture image texture

and edge information and reconstructs the category information of the original image through the decoder structure. U-Net creates a richer feature representation by combining features in the channel dimension, enabling the network to better handle high-resolution images. However, to expand the model's receptive field, models like U-Net may increase the convolutional stride, leading to a reduction in feature map resolution. With the introduction of the transformer structure in computer vision, self-attention-mechanism-based modules have replaced CNN-built deep neural networks, yielding excellent results, although the network architecture has remained largely unchanged. The segmentation transformer (SETR) [12] was the first to use a visual transformer as an encoder for semantic segmentation, achieving good results but with limitations. To address these limitations, the pyramid vision transformer (PVT) [13] was introduced, but like other emerging methods such as Swin, PVT focuses on improving the encoder while overlooking enhancements to the decoder. In semantic segmentation research, the complete region of objects or entities is an important clue for determining the semantic label of local pixel points. Given that objects or entities in a scene exist at multiple scales or in different sizes and positions, it is necessary to construct multi-scale feature expression to capture the varying scales of image content [14–16]. However, due to the fixed expansion rates of the spatial pyramid pooling module, it is challenging to adapt to the scale changes of objects in the image. This sparse sampling method can result in the loss of information from adjacent pixel points, and a large expansion rate can cause a "grid effect". To address these challenges, He J et al. (2019) [17] proposed a dynamic multi-scale network capable of embedding high-level semantic information, capturing rich image content, and adaptively capturing specific scale features related to the input image.

To integrate various water body index features with LBP texture features, Zhao Haiping et al. [18] proposed a water body extraction method based on the fusion of spectral and spatial features, constructing a deep SVM network model. Chen Y et al. [19] proposed a water body extraction model (SAP-CNN) based on adaptive pooling, which extracts water bodies on the basis of superpixel segmentation, enhancing the extraction capability of water body details in urban areas. However, due to the reliance on manually set threshold values in superpixel segmentation, this method has not achieved a fully automated water body data extraction process. To enhance the fusion features' expression for water body details in high-resolution images, Chen Y et al. [20] developed a global spectral convolution module, a multi-scale convolution module, and a boundary refinement module to fuse the spectral features of water bodies at multiple levels. They proposed a fine water body extraction network model based on spectral and multi-scale spatial features using 3D convolution and verified the method's accuracy, confirming its ability to accurately extract slender rivers. Lv Yalong et al. [21] built a deep convolutional neural network model (DCNN) and used labeled samples for supervised training to obtain the spectral and spatial features of water bodies in multiple convolutional layers, enabling accurate identification of water bodies in optical remote sensing images. To mitigate the degradation problem in deep convolutional neural network models, Weng LG et al. [22] combined the residual convolution structure to propose a separable residual segmentation network, SR-SegNet (separable residual SegNet). This approach ensures that the network model training avoids gradient vanishing issues while extracting deep water body features, as the network model becomes deep and complex. This residual convolution structure is evident in many deep learning methods used in remote sensing research [23–25]. In high-resolution optical remote sensing images, some small-scale water body areas contain only tens of pixels, and multi-scale feature learning has become a key factor affecting the accuracy of water body and boundary extraction. To address this, Li Z Y et al., 2019 [26], combined the popular DeepLab-V3+ model with conditional random fields to construct a fine water body extraction method that jointly predicts multi-scale features. By inputting remote sensing images into the model after multi-scale segmentation processing, the feature output of the model is adjusted to a weighted fusion of multi-scale features, and the water body boundary details are optimized using fully connected CRF at the backend of the model. Currently, this semantic segmenta-

tion framework that integrates multiple segmentation results has proven its effectiveness in many fields and competitions. In addition to adding multi-scale convolution or pyramid structures at different positions of the network model, Duan L et al. [27] introduced a convolutional attention structure based on Li's research, proposing an erasing attention (EA) module with the ability to suppress background features. Experimental results have proven that this model can further eliminate the impact of shadows of mountains and buildings in high-resolution images, reducing the false detection in water body extraction. Feng W et al. [28] proposed an enhanced deep convolutional encoder–decoder (DCED) network, Deep U-Net, which applies superpixel segmentation and conditional random fields to enhance the connectivity and consistency of water and non-water areas.

The water body extraction algorithm proposed in this study, based on a multi-scale transformer model structure, not only has broad applicability but also provides better fine segmentation effects in complex scenes. This refined approach ensures that the model adeptly handles the intricacies of floodwater body delineation, providing a robust solution for accurate extraction even in disaster-stricken areas. This innovation represents a substantial leap forward in remote sensing, offering valuable insights and tools for disaster management and environmental monitoring.

2. Materials and Methods

2.1. Research Area Overview

Zhuozhou City is located at the northern end of the Haihe River Basin, which is an important geographical area including multiple rivers. It is situated in the northwest of the North China Plain, in the central part of Hebei Province, adjacent to Beijing. Connected to the Haihe River Basin through several water systems, Zhuozhou City is an important city within the basin. Its unique geographical advantages and transportation conditions make it a vanguard for coordinated economic development and a radiation area for the Xiong'an New Area. Geographically, Zhuozhou is situated in the upper reaches of the Haihe River Basin, near rivers such as the Yongding River and the Juma River. The area's topography is characterized by a higher elevation in the west, gradually sloping down to the east, with relatively flat terrain overall, as shown in Figure 1. The city lies within the front tilting zone of the Taihang Mountains, descending from the northwest to the southeast. The region's highest point is 69.4 m above sea level, while the lowest is 19.8 m, with a ground slope of approximately 1/660. The geomorphology is shaped by the alluvial deposits of the Juma River, featuring two secondary terraces on the north and south sides with a height difference of 2–4 m.

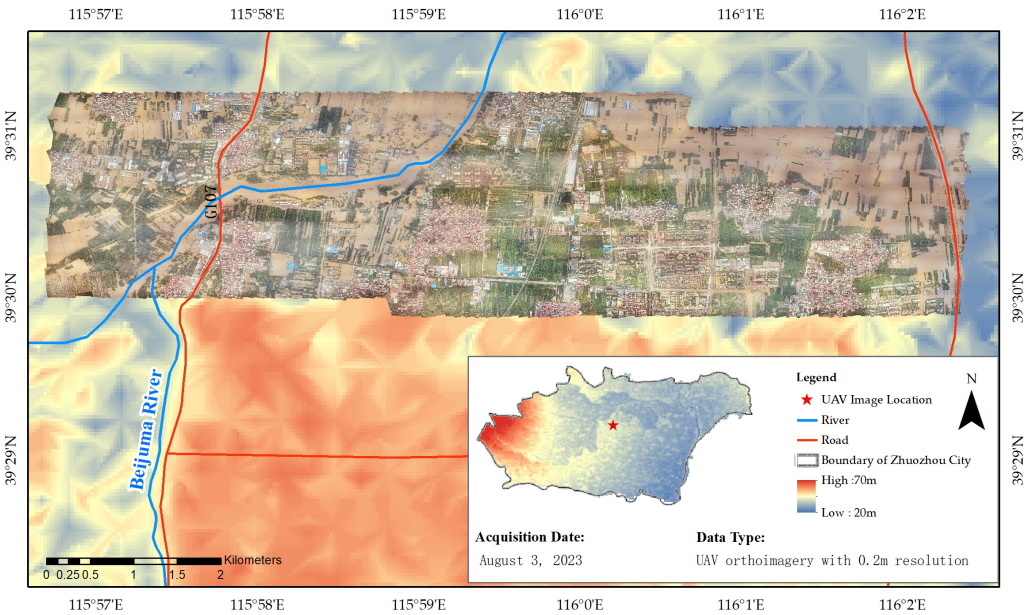

Figure 1. Position map of Zhuozhou City.

2.2. Related Methods and Water Body Extraction Overall Framework

2.2.1. Attention Mechanism

The human attention mechanism is derived from intuition; it is a means for humans to quickly filter high-value information from a vast array of data using limited attention resources. The attention mechanism in deep learning borrows from the way humans think about attention and has been widely applied in various types of deep learning tasks, such as natural language processing (NLP), image classification, and speech recognition, achieving significant results. The basic structure of the attention mechanism is shown in Figure 2 below.

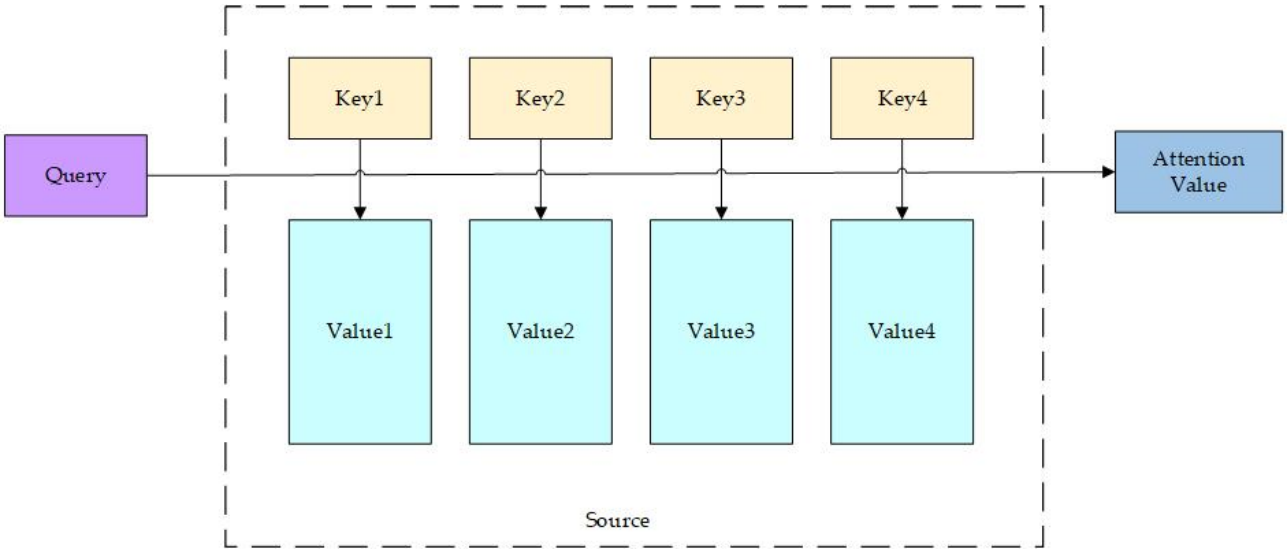

Figure 2. Basic structure diagram of attention mechanism.

It can be seen from Figure 2 that the attention mechanism envisions the constituent elements of the input (source) as a series of key–value pairs. For a given element (query) in the output (target), the mechanism calculates the similarity or correlation between the query and each key, obtaining the weight coefficients for the corresponding value of each key. Then, by performing a weighted summation of the values and their weight coefficients, the final attention value is obtained. Essentially, the attention mechanism performs a weighted summation of the value elements in the source, with the query and the key being used to calculate the weight coefficients for the corresponding values. The calculation process of the attention mechanism is shown in Figure 3.

It can be seen from Figure 3 that the calculation process is divided into three stages:

(1) Based on the query and a certain key, Key_i, calculate the similarity or correlation between the two. The similarity calculation can introduce different functions and computational mechanisms, the most common methods include the following: calculating the dot product of the two vectors, calculating the vector Cosine (Cos) similarity between the two, or introducing an additional neural network to obtain the value.

(2) Since the similarity values obtained in step (1) have different value ranges depending on the specific calculation method used, a calculation method similar to SoftMax is introduced to numerically transform the scores from the first stage. Through this step, normalization can be performed, organizing the original calculation scores into a probability distribution where the sum of all element weights is 1; at the same time, the inherent mechanism of SoftMax further highlights the weight of important elements.

(3) The calculation result a_i from step (2) is the weight coefficient corresponding to value_i, and the attention value can be obtained by weighted summation.

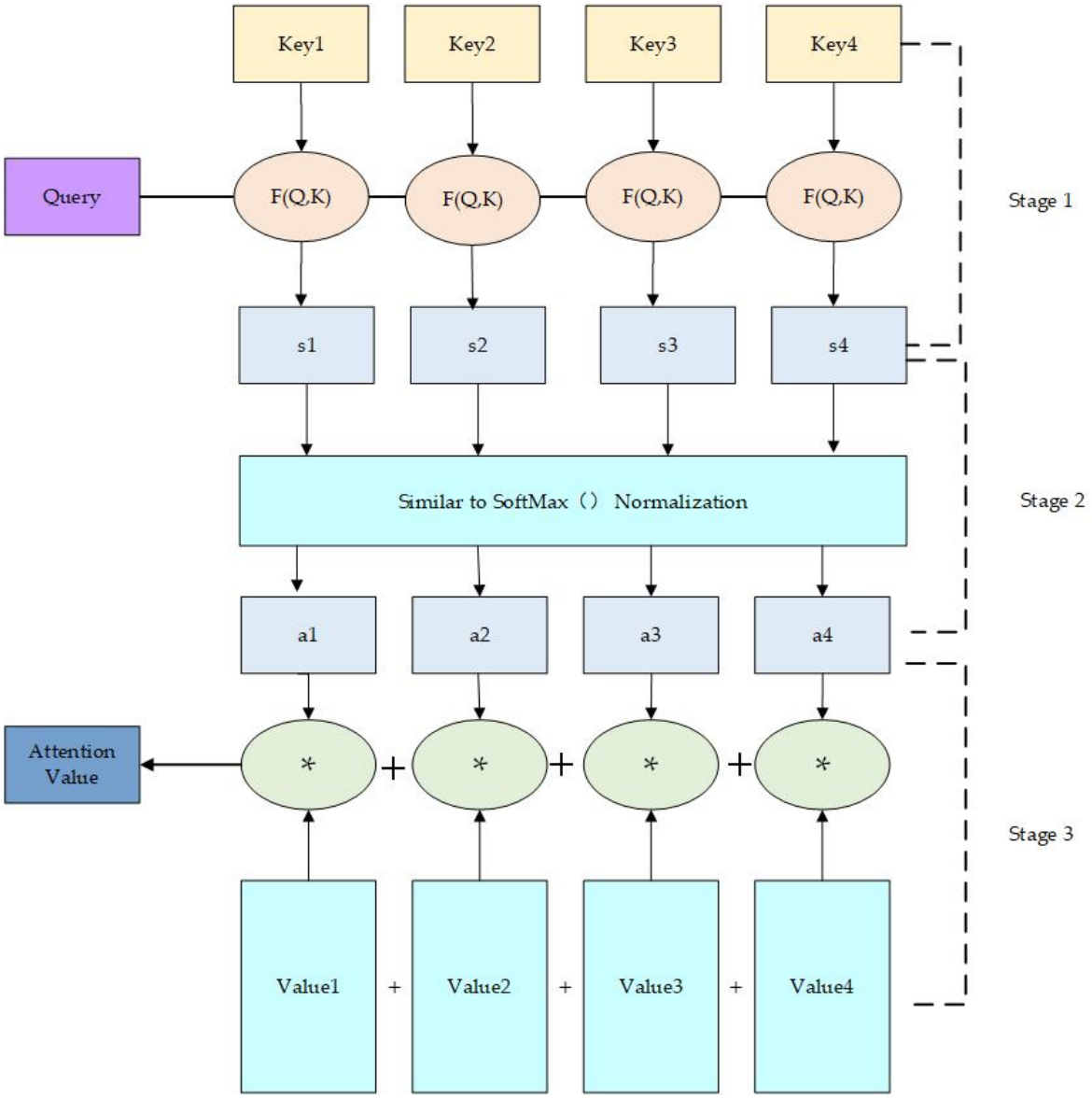

Figure 3. A schematic diagram of the calculation process of attention mechanism.

2.2.2. Water Body Extraction Overall Framework

This study proposes a method for extracting water bodies from remote sensing imagery using deep learning segmentation technology. The main technical process is illustrated in Figure 4. The framework of this method includes the following three key steps:

(1) Dataset acquisition and production: The process begins with both automatic and manual data annotation of the acquired remote sensing images to construct a comprehensive floodwater body dataset.

(2) Model training and knowledge transfer: Using the constructed dataset, the method first involves self-supervised pre-training to obtain a pre-trained model. This is followed by training with the fully annotated dataset to develop the water body extraction model.

(3) Extraction and result processing of floodwater contours: For the input remote sensing images, the method consists of five steps—data reading, preprocessing, water body extraction, morphological processing, and vector conversion.

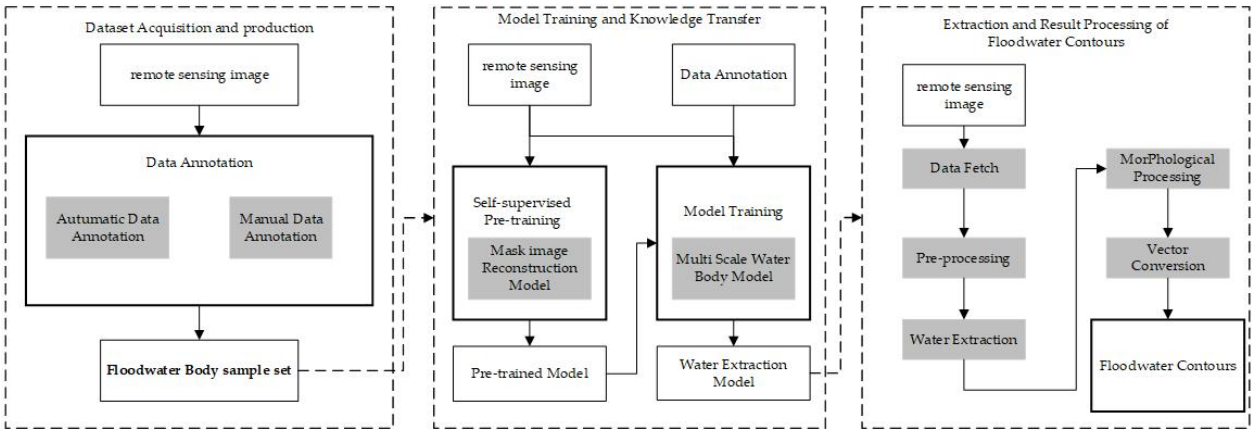

Figure 4. Floodwater extraction algorithm model structure flowchart.

2.3. Algorithm Technical Route and Process

2.3.1. Dataset Acquisition and Production

The primary goal of dataset acquisition and production is to construct a floodwater body dataset through a combination of automatic and manual data annotation for the acquired remote sensing images. To achieve high-precision floodwater body extraction results, it is essential to begin with a large-scale, accurately annotated remote sensing water body dataset and continuously enrich the data samples during use and evolution. The high-resolution GID dataset [29] and the GLH-Water dataset [5] can serve as foundational datasets in this research.

(1) GID Dataset

A large-scale land cover dataset has been constructed using GF-2 satellite images. This dataset is named the Gaofen Image Dataset (GID), and it has the advantages of large coverage, wide distribution range, and high spatial resolution. The GID consists of two parts: a large-scale classification set and a fine land cover classification set. The large-scale classification set includes 150 pixel-level annotated GF-2 images, and the fine classification set is composed of 30,000 multi-scale image blocks, plus 10 pixel-level annotated GF-2 images. Training and validation data for 15 categories were collected, and the images were re-labeled based on training and validation images with five categories. Each image is sized at 7200 × 6800 pixels, with a spatial resolution of 4 m.

(2) GLH-Water Dataset

Very high resolution (VHR) satellite imagery for global surface water detection can directly serve key applications such as refining flood mapping and water resource assessment. Although some progress has been made in the detection of surface water from local areas and low-resolution satellite imagery, high-resolution datasets suitable for global surface water mapping and analysis are still to be explored. To encourage the undertaking of this task and facilitate the implementation of related applications, the GLH-Water dataset [5] has been proposed. This dataset consists of 250 satellite images and manually annotated surface water labels, distributed globally and encompassing various types of water bodies (e.g., rivers, lakes, and ponds in forests, irrigated fields, bare areas, and urban areas). Each image is sized at 12,800 × 12,800 pixels with a spatial resolution of 0.3 m.

(3) Floodwater Body Dataset

To address the complexity of floodwater scenarios further, drone imagery from various regions and water bodies has been collected. These images are annotated with a graffiti-style mark (i.e., an arbitrary straight line drawn through the center of the water area). Based on the annotation information, the images are cropped to create samples. Samples containing the annotation information are placed into the water body subset, while the remaining samples are categorized into the non-water body subset, thereby constructing a

floodwater body sample set from UAV orthoimages. This sample set is expanded through manual data annotation, and in subsequent use, the model's output is used as automatic data annotation, which is then manually corrected to further expand the sample set. Post-disaster UAV orthoimages from various flood events across China have been collected, including the Poyang Lake breach in Jiangxi Province, the Flash Flood Disaster in Luonan of Shaanxi Province, the Dehui flood in Changchun City, the Suizhou flood in Hubei Province, the "720" catastrophic flood disaster in Zhengzhou of Henan Province, and Super Typhoon "Hinnamnor" in Zhejiang Province. Through manual annotation, a large number of learning samples have been created to participate in model training. Due to the high complexity of water bodies in flooded areas, discerning boundaries is often difficult, particularly in rural farmland, where the distinction becomes particularly challenging.

To enhance the floodwater body model's adaptability and improve the extraction of floodwater contours, this study has classified and annotated floodwater bodies, treating water bodies, flooded farmland, and receded farmland as separate samples for model training. Additionally, considering the complexity of weather conditions at disaster sites and the reflective properties of water bodies, a reflective water body dataset was constructed by annotating reflective water body contours from GF-2 satellite imagery This dataset was incorporated into the model training to reduce errors caused by reflection. The dataset constructed in this study categorizes water bodies into the following classifications:

(1) River, pond (Figure 5a).
(2) Submerged urban roads—cannot include buildings; parked vehicles on the street can be ignored (Figure 5b).
(3) Submerged agricultural land—large areas of forests and vegetation should not be included in the water body area (Figure 5c).
(4) Reflective water bodies from satellite images (Figure 5d).

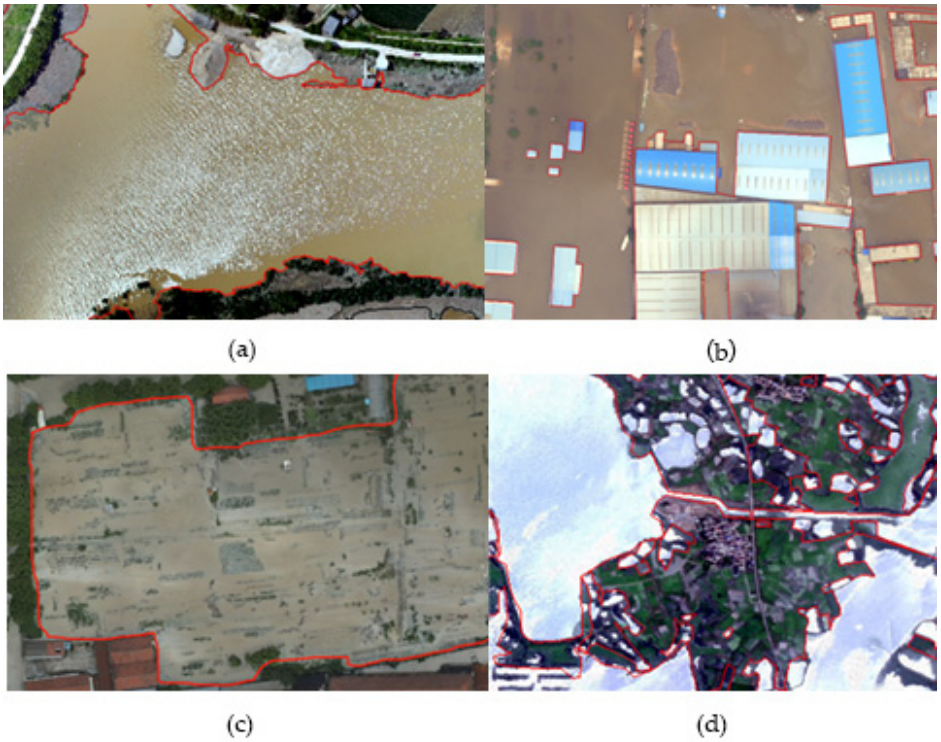

Figure 5. Independently constructed dataset: (**a**) river, pond; (**b**) submerged urban roads; (**c**) submerged agricultural land; (**d**) reflective water bodies in satellite imagery.

2.3.2. Model Training and Knowledge Transfer

Using the publicly available high-resolution GLH-Water water body dataset and the independently built floodwater body dataset, the study first conducts self-supervised pre-

training to obtain a pre-trained model. Subsequently, the complete annotated dataset is used for model training to develop the water body extraction model.

(1) Model Pre-Training and Knowledge Transfer Based on Deep Learning for Remote Sensing Data

Remote sensing images differ significantly from natural images in terms of imaging equipment, environment, and scene content, leading to substantial domain differences. Most existing remote sensing semantic segmentation models use a backbone model pre-trained on ImageNet to initialize the feature extraction network. However, this approach results in weak feature expression, which severely impacts the accuracy of floodwater body extraction. To address this issue, this module proposes a model pre-training method specifically designed for large-scale remote sensing images. Given the abundance of unlabeled data in remote sensing scenes, the model's main network is pre-trained using an unsupervised large-scale pre-training task, which effectively enhances the model's performance in remote sensing applications.

With the continuous development of remote sensing technology, an increasing volume of rich remote sensing images is being obtained. However, due to the high cost of manual annotation and the need for specialized knowledge, it is impractical to use a supervised approach for pre-training remote sensing models. Considering the success of self-supervised learning on natural images, this method adopts a self-supervised approach for pre-training the model on a large-scale, unlabeled remote sensing image dataset. This process yields a pre-trained model specifically tailored for the remote sensing domain. Given the complex backgrounds and varied target shapes in UAV images, this subsystem employs a generative model based on masked image reconstruction for model pre-training. After the input image is fed into the network, it is randomly blocked and masked, followed by reconstruction in the output. By performing the unsupervised task of masked reconstruction, the performance of the model's main network in remote sensing scenarios is significantly improved. The structure of the pre-training model is illustrated in Figure 6.

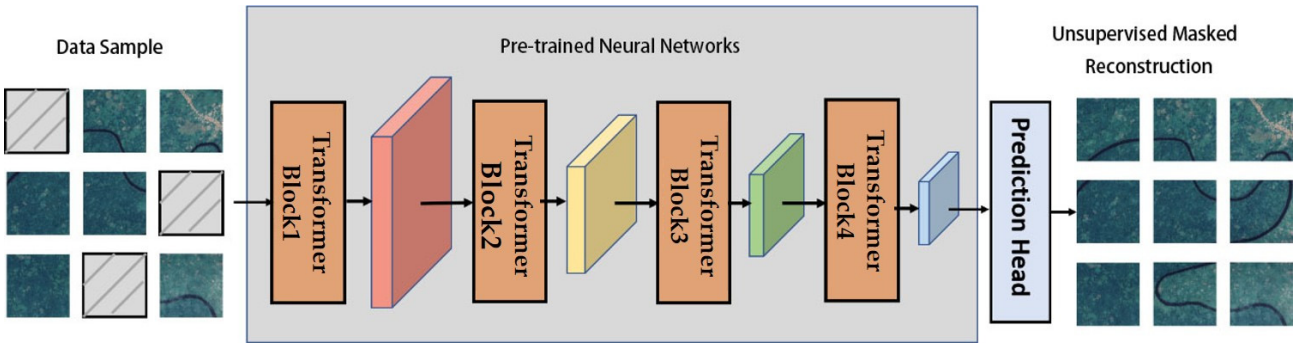

Figure 6. Unsupervised masked reconstruction pre-training task for remote sensing.

(2) Water Body Extraction Based on Deep Learning Semantic Segmentation Technology

The transformer model, a type of neural network based on the self-attention mechanism, has gained significant popularity in natural language processing (NLP) applications. Building upon the standard transformer, the multi-scale transformer model adds the capability to handle different scales or resolutions. In the field of semantic segmentation technology based on deep learning, the transformer structure leverages its global attention mechanism and superior feature representation to effectively capture the spatial contextual information in images. Compared to traditional convolutional neural networks (CNNs), the transformer can process various parts of an image more flexibly, without being constrained by a fixed receptive field. As a result, the transformer structure has demonstrated superior performance to CNNs in semantic segmentation tasks, particularly in scenarios where global information is essential for accurately classifying each pixel.

In response to the business needs for disaster prevention and mitigation, and to address the challenges of water body extraction from high-resolution remote sensing imagery, we have conducted research on water body extraction using deep learning technology. We have adopted an improved multi-scale transformer model structure to achieve the segmentation of water areas, as illustrated in Figure 7. By combining feature maps of different resolutions, the multi-scale transformer model can capture both local details and global context within the image. This ability to integrate information at various scales enables the model to deliver more accurate and robust segmentation results, particularly when dealing with images that contain complex structures and multi-scale objects.

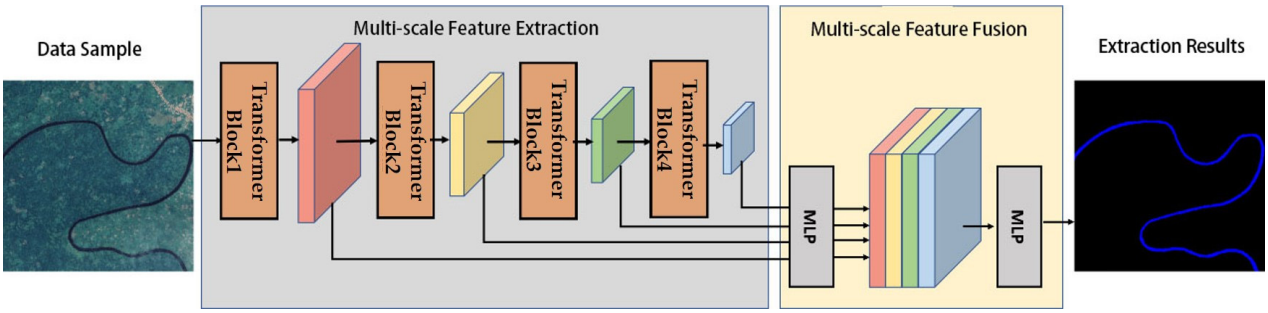

Figure 7. Multi-scale transformer water body segmentation model structure.

The framework primarily involves two key steps:

(1) Using a pre-trained transformer model as the backbone network to extract hierarchical features of water bodies in remote sensing images. The transformer model offers stronger feature expression capabilities, enabling the network to extract multi-scale water body features effectively.

(2) Aligning and splicing the multi-scale water body features to achieve feature fusion. Once the model is trained in an end-to-end manner, it is employed for floodwater body extraction, yielding the water body segmentation results.

(3) The transformer block structure is designed as depicted in Figure 8. Each block comprises multiple identical transformer layers, and each transformer layer consists of three main components: layer norm, self-attention, and MLP (multi-layer perceptron). Multi-layer perceptron (MLP) [30,31] is a fundamental type of artificial neural network, composed of multiple layers including an input layer, one or more hidden layers, and an output layer. It is a type of feedforward neural network. Each layer consists of multiple neurons, and each neuron is connected to all neurons in the previous layer, transmitting and processing information through weights and activation functions. The self-attention operation, in particular, is crucial, as it is the primary reason that explains why the transformer structure possesses such strong representation capabilities.

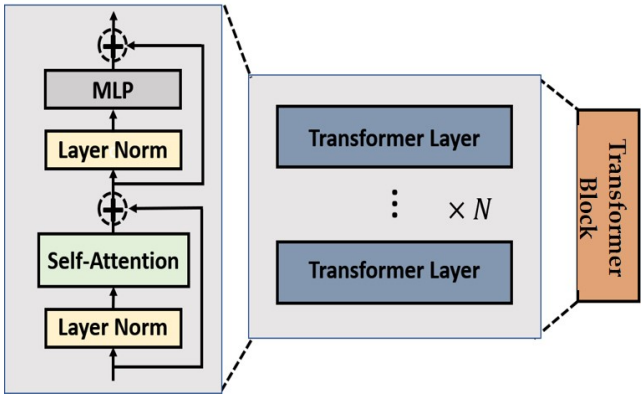

Figure 8. Transformer block structure.

3. Results

3.1. Application in the Haihe River '23·7' Basin Catastrophic Flood

The current research process of water body extraction based on deep learning algorithms can generally be divided into four stages: image (and label) input, feature extraction, semantic segmentation, and post-processing (as shown in Figure 9). Due to the flexibility of designing convolutional layer structures, deep learning algorithms can use remote sensing images of different types and channel numbers as inputs, but usually require normalization of the images before training. Subsequently, after learning from the sample data using the network model, the neural network can perform pixel-level segmentation based on the learned water body features to obtain preliminary segmentation results. To further overcome issues such as segmentation voids and fragments caused by other objects in high-resolution images, it is usually necessary to combine image processing and other means to post-process the extraction results, thereby improving the completeness and accuracy of water body extraction.

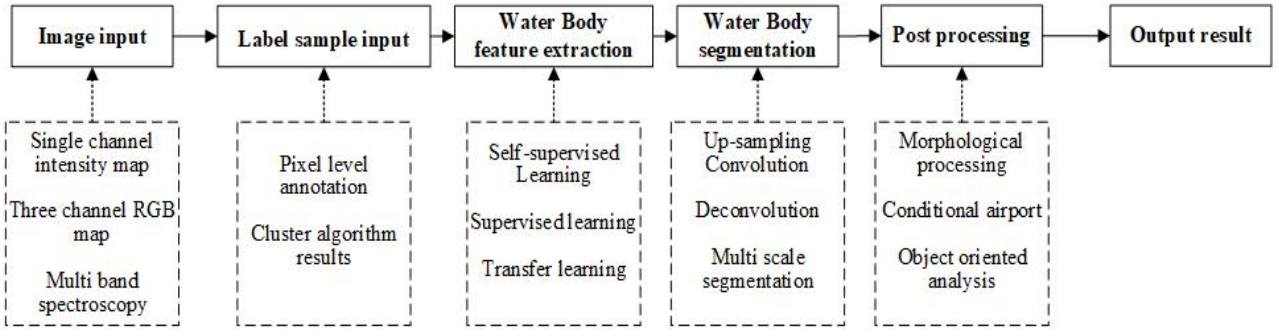

Figure 9. Water body extraction process flowchart.

Figure 10 shows the structures of four different deep learning models used in water body extraction research. Figure 10a shows the convolutional neural network model, represented by VGG16 (or VGG19), which is most commonly used in remote sensing image classification and object detection. It uses stacked multi-layer convolutions and pooling structures to obtain the texture and semantic features of the original image, with the number of feature maps at each layer determined by the number of convolutional kernels, and each feature map can respond differently to different types of ground objects. Figure 10b shows a deep residual convolutional network composed of multiple stacked residual convolutional blocks, which is often used as the feature encoding structure of a U-shaped neural network and is frequently used in remote sensing image segmentation research. Figure 10c,d show two different multi-scale feature extraction model structures. Figure 10c uses downsampling to process the original image at multiple scales in the input layer, and Figure 10d uses pooling/convolution operations with different pooling rates or dilation rates in the middle convolutional layer to process the feature maps at multiple scales, obtaining local and global texture and semantic features of the water body.

In July 2023, Typhoon "Dusuri" made landfall in Fujian Province and continued its path northward. After making landfall, the residual circulation of the typhoon lingered over the North China region for an extended period, resulting in extreme rainfall across many areas and causing the catastrophic '23·7' flood in the Haihe River Basin.

This study focuses on Zhuozhou City, located in the Haihe River Basin in Hebei Province, as the research area. Zhuozhou City is situated in the northern part of the Hebei Plain, characterized by relatively low terrain. The city experienced severe urban waterlogging due to the combined effects of intense rainfall and upstream river flooding.

For this study, orthoimagery data captured by UAV with a resolution of 0.2 m was utilized, taken on 3 August 2023, in the northern suburbs of Zhuozhou City, was used to automatically extract the boundaries of floodwater bodies. The extraction process is illustrated in Figure 11.

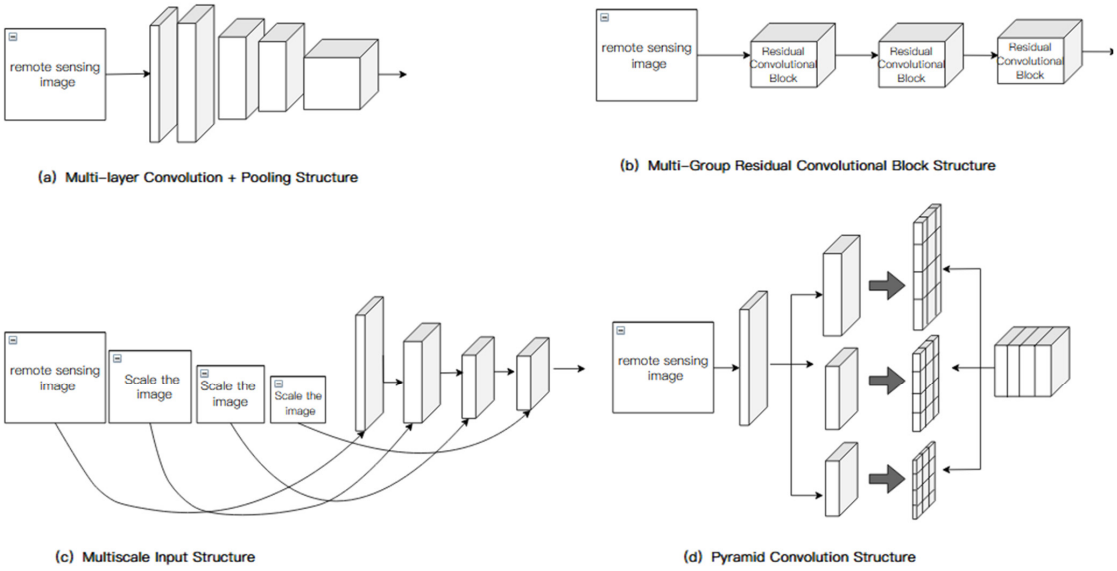

Figure 10. Method and structure of model features for remote sensing images.

Figure 11. Results diagram of flood water vector: (**a**,**c**) in fields and (**b**,**d**) in residential area.

On the morning of 3 August 2023, two drone monitoring flights were conducted over the severely affected northern and central suburban areas of Zhuozhou City, covering an area of 8 square kilometers. By comparing the pre-disaster remote sensing data of the flight area (Figure 12) and excluding data overlapping with rivers, it was determined that approximately 6.5 square kilometers of the area was flooded, as shown in Figure 13. The inundated regions were primarily farmland, and 64 km of roads, including sections of provincial highways and expressways, such as G107 and G4, were submerged. Analysis revealed that the northern part of Zhuozhou City, characterized by lower terrain, experienced severe flooding; meanwhile, the southern part, with its higher elevation, likely only faced localized flooding on the fringes of the urban area, leaving the main urban areas relatively unaffected.

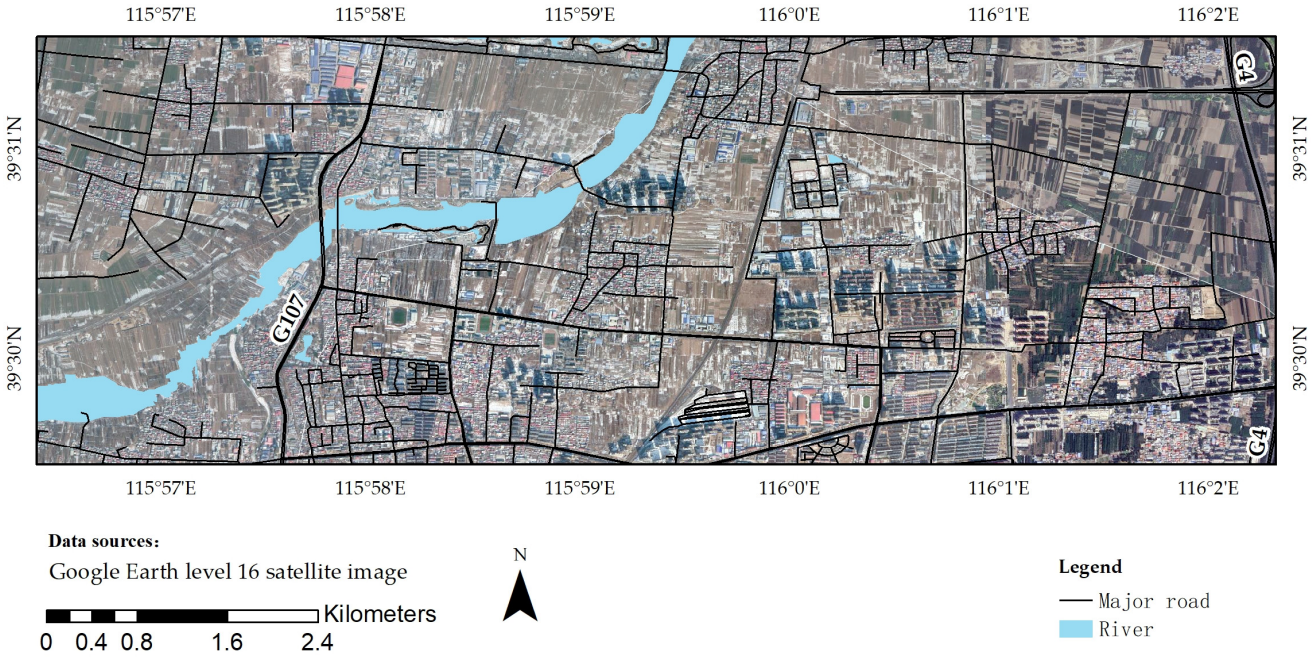

Figure 12. Pre-disaster basic situation map of Zhuozhou City.

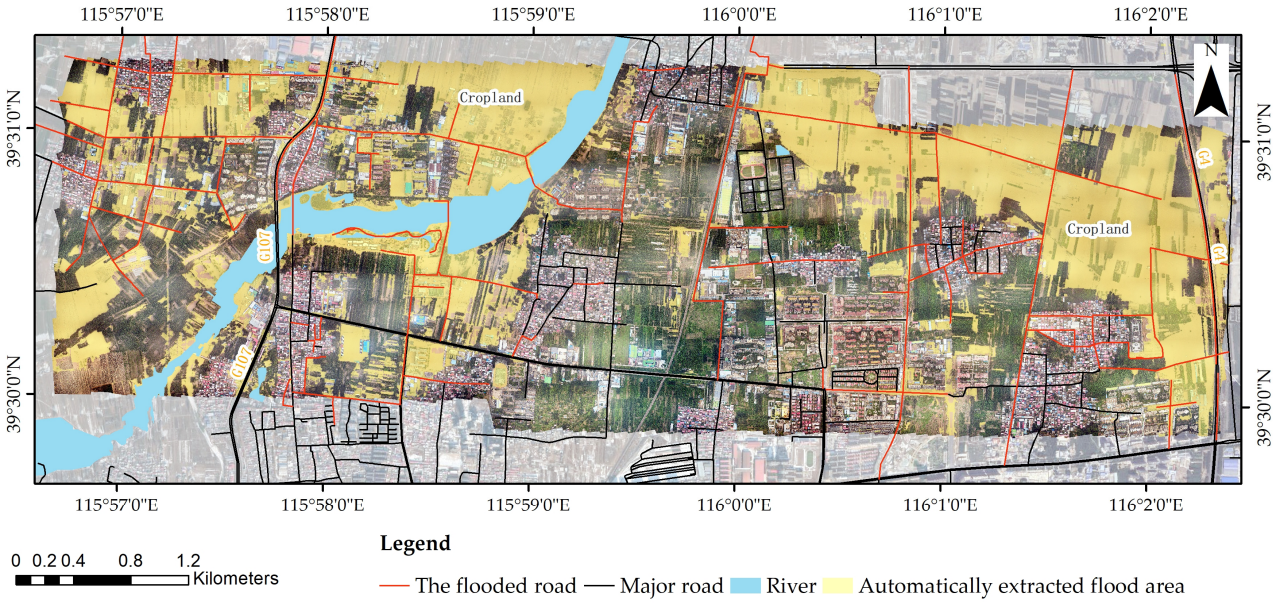

Figure 13. Post-disaster comprehensive analysis map of Zhuozhou City.

3.2. Accuracy Assessment

The model's accuracy is evaluated using the Paddle framework for training, with the pre-trained multi-scale transformer model utilized for parameter initialization, and the results are as shown in Table 1. All components of the network, except for the main network part, are initialized randomly with a Gaussian distribution (mean = 0, variance = 0.01). The AdamW optimizer [32] is employed with an initial learning rate of 0.00006 and a weight decay parameter of 0.01. Data augmentation techniques, including random horizontal flipping, random scaling, and random pixel perturbation, are applied, with a training batch size of 8. After extensive experimentation and training, we have determined the optimal values for our model's hyperparameters: a learning rate of 0.00006 and the AdamW optimizer is configured with parameters beta1 = 0.9, beta2 = 0.99, and weight decay = 0.01. The model was trained for a total of 80,000 iterations.

Table 1. Technical indicators on GLH-Water dataset [1].

Method	IoU (%)	F1-Score (%)
MECNet	44.67	61.75
MSResNe	69.76	82.18
MSCENet	74.81	85.58
FCN8s (General)	73.66	84.83
PSPNet (General)	75.19	85.84
DeepLab v3	79.8	88.76
HRNet-48	78.6	88.01
STDC-1446	75.82	86.25
MagNet (High-Res)	62.77	-
FCtL (High-Res)	74.92	85.66
ISDNet (High-Res)	53.04	-
PCL	82.26	90.27
Our Method	92.25	91.91

(1) Intersection over Union (IoU)

Intersection over Union (IoU) measures the overlap between the predicted water body area and the annotated true area. Introduced by Paul Jaccard in the early 20th century, IoU is defined as the ratio of the intersection area to the union area. To compute IoU, the predicted area and the true area must overlap. The intersection refers to the overlapping region of the detected and true areas, while the union represents the total area covered by both. The IoU calculation method is described in Equation (1).

$$\mathrm{IoU} = \frac{\mathrm{A} \cap \mathrm{B}}{\mathrm{A} \cup \mathrm{B}} \tag{1}$$

(2) Evaluation Metrics

To assess precision and recall in water body extraction, the number of correctly predicted pixels (true positives—TPs) must be determined. The formulas for calculating precision and recall are as follows:

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}(\mathrm{False\ Positives})} \tag{2}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}(\mathrm{False\ Negative})} \tag{3}$$

$$\mathrm{F1} = \frac{2\mathrm{PR}}{\mathrm{P} + \mathrm{R}} \tag{4}$$

where TP + FP represents the total number of predicted water body pixels, and TP + FN denotes the total number of true water body pixels. P and R are the abbreviations for precision and recall, respectively.

The model is trained for a total of 80,000 iterations using 4 NVIDIA GTX 3090 graphics cards (24 GB memory each) with CUDA acceleration. Comparative experiments with existing water body extraction methods demonstrate that the proposed algorithm outperforms others, achieving higher Intersection over Union (IoU) and F1-score metrics.

4. Discussion

Due to significant variations in the color, shape, and other textural and geometric features of water bodies across different regions and basin environments, pixel-level classification and regional segmentation of water bodies present substantial challenges. Current multi-scale convolutional neural networks, while capable of extracting more detailed features, often result in complex models with high computational demands. Consequently, even minor improvements in algorithm accuracy frequently require more data and increased computational resources. The challenges in extracting flood water bodies from UAV orthoimagery include the following:

(1) Diverse water body types: Remote sensing images feature various water bodies, such as rivers, lakes, and wetlands, each display different spectral characteristics due to terrain, sediment content, and microorganism density. This variability complicates accurate model identification.
(2) Seasonal variations: Water bodies exhibit different textures and spectral properties due to seasonal climate changes, which can be mistaken for shadows cast by vegetation or buildings. The lack of post-disaster training samples further hampers the accuracy of water body boundary extraction.
(3) Limited high-resolution datasets: High-resolution remote sensing datasets with pixel-level annotations are scarce, and the available samples of different water body types are often imbalanced. This scarcity makes it challenging to train robust and generalizable extraction algorithms.
(4) Computational efficiency: The extensive spatial distribution of water bodies in remote sensing images requires algorithms that are both resource-efficient and capable of rapid computation.

5. Conclusions

This study addresses the critical technical challenges in extracting complex water bodies during disaster scenarios, aiming to support disaster-prevention and -mitigation efforts. A comprehensive dataset comprising post-disaster imagery of flood-affected water bodies has been compiled, utilizing high-resolution orthoimagery data captured by UAVs. The study further investigated water body extraction techniques using deep learning and implemented an enhanced multi-scale transformer model structure for precise water area segmentation. This advanced model optimizes feature extraction and significantly improves accuracy through end-to-end training. The application of this model enhances response speed during disasters, reduces reliance on manual interpretation, lowers costs, and increases efficiency. Rapid and accurate water body extraction is vital for disaster assessment and planning rescue operations.

In practical applications, the study utilized post-disaster images from the "Haihe 23·7 River Basin Catastrophic Flood" as experimental data. The results demonstrated that the model effectively delineated flood zones, providing timely and accurate water distribution information to disaster management personnel. However, in challenging areas like those with rugged terrain or dense vegetation, the model exhibited some misclassification, which could impact accuracy. To enhance the model's robustness and precision, human intervention remains necessary. Professionals can refine and optimize the model's outputs based on actual conditions, ensuring more accurate water body extraction results that meet disaster management needs. With ongoing technological advancements and model optimization, it

is anticipated that manual intervention will be further reduced, and automation levels will increase in the future.

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

Funding: This research was funded by the National Key Research and Development Program of China (No. 2022YFB3903404), and the National Key Research and Development Program of China (No. 2023YFC3006504).

Data Availability Statement: Data are unavailable due to ethical restrictions.

Acknowledgments: We would like to thank the editor and anonymous reviewers for their constructive comments and suggestions for improving this manuscript.

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

References

1. Nagasawa, R.; Mas, E.; Moya, L.; Koshimura, S. Model-based analysis of multi-UAV path planning for surveying postdisaster building damage. *Sci. Rep.* **2021**, *11*, 18588. [CrossRef] [PubMed]
2. Garnica-Peña, R.J.; Alcántara-Ayala, I. The use of UAVs for landslide disaster risk research and disaster risk management: A literature review. *J. Mt. Sci.* **2021**, *18*, 1–17. [CrossRef]
3. Dong, Z.; Zhang, M.; Li, L.; Liu, Q.; Wen, Q.; Wang, W.; Luo, W.; Wu, Z.; Tang, T.; Ji, W. A multiscale building detection method based on boundary preservation for remote sensing images: Taking the Yangbi M6.4 earthquake as an example. *Nat. Hazards Res.* **2022**, *2*, 121–131. [CrossRef]
4. Busetti, A.; Leone, C.; Corradetti, A.; Fracaros, S.; Spadotto, S.; Rai, P.; Zini, L.; Calligaris, C. Coastal Storm-Induced Sinkholes: Insights from Unmanned Aerial Vehicle Monitoring. *Remote Sens.* **2024**, *16*, 3681. [CrossRef]
5. Li, Y.; Dang, B.; Li, W.; Zhang, Y. GLH-Water: A Large-Scale Dataset for Global Surface Water Detection in Large-Size Very-High-Resolution Satellite Imagery. *arXiv* **2023**, arXiv:2303.09310. [CrossRef]
6. Shelhamer, E.; Long, J.; Darrell, T. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 640–651. [CrossRef]
7. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015, Proceedings of the 18th International Conference, Munich, Germany, 5–9 October 2015*; Springer International Publishing: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
8. Liang-Chieh, C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. *arXiv* **2014**, arXiv:1412.7062.
9. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021.
10. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv* **2017**, arXiv:1706.05587.
11. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 833–851.
12. Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P.H.; et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021.
13. Wang, W.; Xie, E.; Li, X.; Fan, D.P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021.
14. Zhao, H.; Qi, X.; Shen, X.; Shi, J.; Jia, J. ICNet for Real-Time Semantic Segmentation on High-Resolution Images. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; Springer: Berlin/Heidelberg, Germany, 2018; pp. 405–420.
15. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
16. 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.* **2018**, *40*, 834–848. [CrossRef]

17. He, J.; Deng, Z.; Qiao, Y. Dynamic Multi-Scale Filters for Semantic Segmentation. In Proceedings of the International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3562–3572.
18. Zhao, H.P. Research on Water Body Recognition Combining Spectral and Deep Learning Features Under Big Data. Master's Thesis, Dalian Jiaotong University, Dalian, China, 2018.
19. Chen, Y.; Tang, L.L.; Kan, Z.H.; Bilal, M.; Li, Q.Q. A novel water body extraction neural network (WBENN) for optical high resolution multispectral imagery. *J. Hydrol.* **2020**, *588*, 125092. [CrossRef]
20. Chen, Y.; Fan, R.S.; Yang, X.C.; Wang, J.X.; Latif, A. 2018. Extraction of urban water bodies from high-resolution remote-sensing imagery using deep learning. *Water* **2018**, *10*, 585. [CrossRef]
21. Lv, Y.L.; Tian, S.W.; Yu, L.; Zhang, R. Water body recognition based on CNN_SVM with joint spectral features. *Comput. Eng. Des.* **2019**, *40*, 243–247.
22. Weng, L.G.; Xu, Y.M.; Xia, M.; Zhang, Y.H. Water areas segmentation from remote sensing images using a separable residual SegNet network. *Int. J. Geo-Inf.* **2020**, *9*, 256. [CrossRef]
23. Zhang, Z.X.; Liu, Q.J.; Wang, Y.H. Road Extraction by Deep Residual U-Net. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 749–753. [CrossRef]
24. Qiu, C.P.; Mou, L.C.; Schmitt, M.; Zhu, X.X. Fusing multiseasonal Sentinel-2 imagery for urban land cover classification with multibranch residual convolutional neural networks. *IEEE Geosci. Remote Sens. Lett.* **2020**, *17*, 1787–1791. [CrossRef]
25. Cai, Y.H.; Guo, Y.J.; Lang, S.N.; Liu, J.Q.; Hu, S.B. Classification of hyperspectral images by spectral-spatial dense-residual network. *J. Appl. Remote Sens.* **2020**, *14*, 036513. [CrossRef]
26. Li, Z.Y.; Wang, R.; Zhang, W.; Hu, F.M.; Meng, L.K. Multiscale features supported DeepLabV3+ optimization scheme for accurate water semantic segmentation. *IEEE Access* **2019**, *7*, 155787–155804. [CrossRef]
27. Duan, L.H.; Hu, X.Y. Multiscale refinement network for water body segmentation in high resolution satellite imagery. *IEEE Geosci. Remote Sens. Lett.* **2020**, *17*, 686–690. [CrossRef]
28. Feng, W.Q.; Sui, H.G.; Huang, W.M.; Xu, C.; An, K.Q. Water body extraction from very high-resolution remote sensing imagery using deep U-Net and a super pixel-based conditional random field model. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 618–622. [CrossRef]
29. Tong, X.-Y.; Xia, G.-S.; Lu, Q.; Shen, H.; Li, S.; You, S.; Zhang, L. Land-cover classification with high-resolution remote sensing images using transferable deep models. *Remote Sens. Environ.* **2020**, *237*, 111322. [CrossRef]
30. Minsky, M.; Papert, S. *Perceptrons: An Introduction to Computational Geometry*; MIT Press: Cambridge, MA, USA, 1969.
31. Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. Learning representations by back-propagating errors. *Nature* **1986**, *323*, 533–536. [CrossRef]
32. Gugger, S.; Howard, J. AdamW and Super-Convergence Is Now the Fastest Way to Train Neural Nets. fast.ai blog. 2018. Available online: https://www.fast.ai/posts/2018-07-02-adam-weight-decay.html (accessed on 1 April 2023).