sustainability

MDPI

Article

A Study on Identification of Urban Waterlogging Risk Factors Based on Satellite Image Semantic Segmentation and XGBoost

Jinping Tong [1,2], Fei Gao [1], Hui Liu [1,*], Jing Huang [2,3], Gaofeng Liu [3], Hanyue Zhang [1] and Qiong Duan [4]

[1] Business School, Changzhou University, Changzhou 213100, China
[2] Management Science Institute, Hohai University, Nanjing 210098, China
[3] Business School, Hohai University, Nanjing 210098, China
[4] Information Technology Center, Luoyang Institute of Science and Technology, Luoyang 471023, China
* Correspondence: hliu@cczu.edu.cn

Abstract: As global warming exacerbates and urbanization accelerates, extreme climatic events occur frequently. Urban waterlogging is seriously spreading in China, resulting in a high level of vulnerability in urban societies and economies. It has been urgent for regional sustainable development to effectively identify and analyze the risk factors behind urban waterlogging. A novel model incorporating satellite image semantic segmentation into extreme gradient boosting (XGBoost) is employed for identifying and forecasting the urban waterlogging risk factors. Ground object features of waterlogging points are extracted by the satellite image semantic segmentation, and XGBoost is employed to predict waterlogging points and identify the primary factors affecting urban waterlogging. This paper selects the coastal cities of Haikou, Xiamen, Shanghai, and Qingdao as research areas, and obtains data from social media. According to the comprehensive performance evaluation of the semantic segmentation and XGBoost models, the semantic segmentation model could effectively identify and extract water bodies, roads, and green spaces in satellite images, and the XGBoost model is more accurate and reliable than other common machine learning methods in prediction performance and precision. Among all waterlogging risk factors, elevation is the main factor affecting waterlogging in the research areas. For Shanghai and Qingdao, the secondary factor affecting waterlogging is roads. Water bodies are the secondary factor affecting urban waterlogging in Haikou. For Xiamen, the four indicators other than the elevation are equally significant, which could all be regarded as secondary factors affecting urban waterlogging.

Keywords: satellite images; urban waterlogging; semantic segmentation; XGBoost

Citation: Tong, J.; Gao, F.; Liu, H.; Huang, J.; Liu, G.; Zhang, H.; Duan, Q. A Study on Identification of Urban Waterlogging Risk Factors Based on Satellite Image Semantic Segmentation and XGBoost. *Sustainability* **2023**, *15*, 6434. https://doi.org/10.3390/su15086434

Academic Editors: Gang Liu, Zhisong Chen, Li Gao and Junyu Chen

Received: 5 March 2023
Revised: 5 April 2023
Accepted: 9 April 2023
Published: 10 April 2023

1. Introduction

As the impermeable surface expands significantly in urban areas due to accelerating urbanization, combined with frequent extreme rainstorms and the increasingly marked heat island effect, cities have witnessed recurring waterlogging. According to statistics, since 2010, an average of more than 180 cities in China have been affected by waterlogging every year, which has caused direct economic losses of more than CNY 100 billion [1]. Urban waterlogging has gravely threatened the life and property of urban residents and affected the safe operations and sustainable development of cities. In the past two years, China has issued the New Urbanization Implementation Plan during the 14th Five-Year Plan Period and the Implementation Opinions on Strengthening Urban Waterlogging Control, in which it is clearly stated that controlling urban waterlogging is a major project concerning both the people's livelihood and development, and it is necessary to intensify waterlogging control for remarkable results by 2025. Therefore, assessing and identifying risk factors behind urban waterlogging can lead to building an early warning mechanism for urban waterlogging and developing urban waterlogging prevention measures.

The recent studies have focused on urban waterlogging risk assessment primarily based on the mathematical statistics of historical disaster data [2], scenario analysis [3],

remote sensing and GIS analysis [4], and indicator system analysis [5,6]. Methodologies based on mathematical statistics of historical disaster data rely on historical documents, disaster databases, or field survey data, in which the problem of data sparsity may distort the assessment results. Scenario analysis-based methodologies are demanding in regard to time scale, accuracy, and simulation modeling of data [7,8]. Methodologies based on the combination of remote sensing technologies and GIS are challenged by the difficulty in accurately expanding the indicator data to space, excessively large research scale, and low mapping accuracy [9–11]. In addition, for the methodologies based on indicator system assessment, the selection of indicators and the determination of weights are subjective, which affects the accuracy of assessment [12].

Identifying ground object features of waterlogging points is the premise and foundation of urban waterlogging assessment. The existing literature paid little attention to it. The current recognition methods of ground object features in satellite images primarily rely on manual encoding [13] and machine learning [14]. However, for extremely complex texture features and ground object classification in satellite images, manual encoding requires designing dedicated algorithms for texture recognition, while machine learning requires automatic learning and updating of parameters according to the correspondence between texture features of known samples and ground objects, resulting in inaccurate extraction of data. With the advancement of deep learning techniques [15], classic semantic segmentation network models such as fully convolutional network(FCN) [16] and U-net [17] are applied in the recognition of ground object features in satellite images. These models can directly acquire different categories of ground objects, raster data for vectorization, and other features in images, which considerably reduces the processing cost of geographic information data. More importantly, with accuracy comparable to that of human eyes, they have become a common methodology for ground object feature recognition in satellite images. However, the U-net semantic segmentation network is more advantageous than FCN in regard to integrating more underlying features [18]. Therefore, a satellite image U-net semantic segmentation model was developed to automatically identify the categories of waterlogging points' ground objects, such as water bodies, roads, and green spaces, and extract the ground objects' features utilized by the urban waterlogging risk assessment. As an improved machine learning model, the extreme gradient boosting (XGBoost) model, which is capable of effectively eliminating the heterogeneity of source data distribution and ensuring high accuracy in prediction and fast model operations, has been applied in urban waterlogging risk assessment.

Based on the above analysis, this paper proposes an integrated model by combining satellite image semantic segmentation and XGBoost to assess satellite imagery-based urban waterlogging risk. Four coastal cities, Haikou, Xiamen, Shanghai, and Qingdao, are the research areas, and urban waterlogging sample data of these cities are obtained from social media. The semantic segmentation model is used to extract the ground object features of waterlogging points from satellite images, which are taken as the waterlogging risk factors. On this basis, the XGBoost is used to predict and analyze waterlogging points and identify the primary factor affecting urban waterlogging, thus providing a scientific basis for the effective prevention of urban waterlogging.

The remainder of this paper is organized as follows: The study area is presented in Section 2. Section 3 describes the data and methodologies adopted in this study, including the data acquisition, U-net semantic segmentation and XGBoost model. The results of the proposed methodologies' performance evaluation and influencing factors of urban waterlogging are reported and discussed in Section 4. Finally, the derived conclusions and policy implications are described in Section 5.

2. Research Areas

Due to the combined effects of geographical location, topography, and the monsoon climate, China's coastal region has become one of the disaster-prone areas most frequently and widely confronted by waterlogging in China [19]. Coastal cities share some common

features in waterlogging. First, located in the fragile and sensitive zone where sea and land interact, coastal cities have huge areas exposed to disasters. Affected by land–sea compound disasters, they are more vulnerable to major urban waterlogging in the context of climate change. Second, the monsoon climate results in uneven seasonal distribution of precipitation, and there are many typhoons and heavy rainfall in summer. Heavy rainfall occurs more frequently in coastal cities during the rainy season every year. Waterlogging will affect a wide range and last for a long time, causing serious impacts on local production and life. Third, coastal cities are key regions and strategic centers for population agglomeration, where waterlogging can cause extraordinary losses. For example, Typhoon Meranti in 2016 and Typhoon Mangkhut in 2018 both caused serious waterlogging on Xiamen Island, affecting the normal operations of the city [20]. Therefore, the study on urban waterlogging risks in coastal areas is typical and representative.

To select representative research areas among coastal cities while taking into account the data availability, in this paper, four cities from south to north along the coastline, i.e., Haikou, Xiamen, Shanghai and Qingdao, are selected as research areas for the investigation of risk factors behind waterlogging.

3. Data and Methodologies

The technical roadmap of the methodologies is shown in Figure 1. Firstly, Weibo posts about urban waterlogging in 2017 and 2018 are collected by calling the Sina Weibo API, and based on these posts, waterlogging points are located, and corresponding satellite images and elevation data are obtained. Secondly, satellite image semantic segmentation is used to identify ground objects and extract various features from the satellite images, which are then integrated with the extracted elevation data. Then, the XGBoost is used for training and prediction, and the primary factors affecting waterlogging in coastal cities are obtained according to their weights. Finally, performance evaluation is conducted on the model results to further verify their reliability.

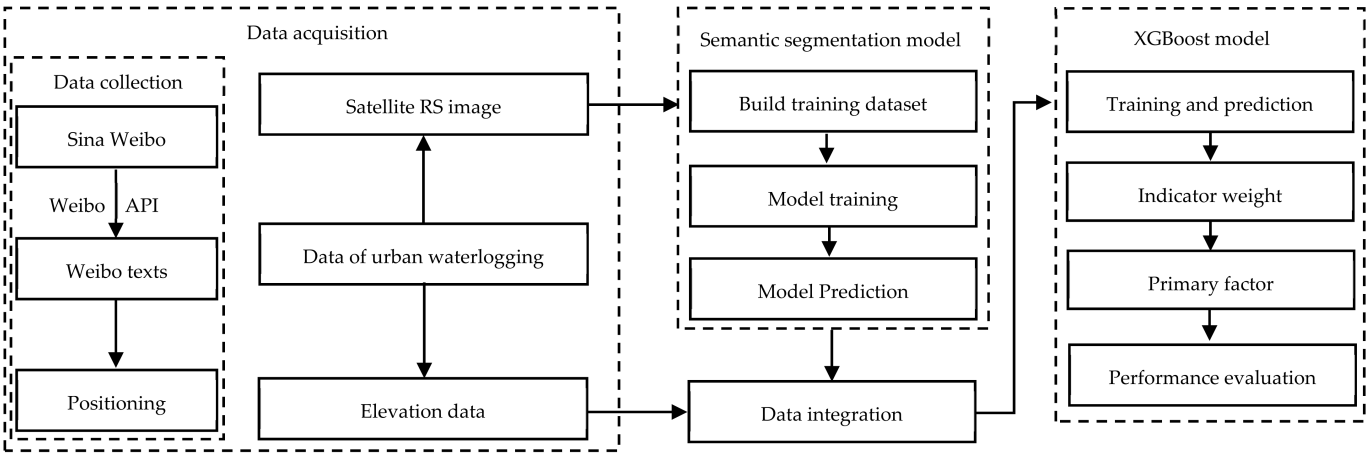


Figure 1. Technical roadmap of the methodologies.

3.1. Data Acquisition

Ground objects such as water bodies, roads, and green spaces in the satellite RS images as well as elevation data are selected in this study to investigate the risk factors behind waterlogging in coastal cities. Therefore, the research objects of this paper include the data of waterlogging points, satellite image data of waterlogging points, and elevation data.

3.1.1. Data of Waterlogging Points

The data of waterlogging points are mostly acquired from Weibo texts. Weibo posts about urban waterlogging in 2017 and 2018 are collected by calling the Sina Weibo API, based on which over 70,000 entries of Weibo texts are obtained after removing the repeating

data. To locate urban waterlogging points more accurately from the Weibo texts, a community directory of 307 cities in China is downloaded in this study, in which community names, geographical locations, floor area ratios, greening ratios, and other information is recorded (source: the residential website https://www.anjuke.com/ (accessed on 10 May 2021)). Terms about communities, roads, and directions are extracted from Sina Weibo texts and then matched with those in the community directory to determine the geographical locations of waterlogging points reported on Weibo.

In order to facilitate the acquisition of satellite images and elevation data of these waterlogging points, ArcGIS is next employed to obtain the longitudes and latitudes of waterlogging points. We utilize ArcGIS geocoding to search for the location in the map by matching the input address and feature attributes, and convert it into actual coordinates, namely longitude and latitude.

With the acquired Weibo data processed, 439 waterlogging points are collected, including 147 in Haikou, 62 in Xiamen, 186 in Shanghai, and 44 in Qingdao. A waterlogging point is taken as a positive sample and marked as "1". In the cities where positive samples are located, the function of ArcGIS to generate random points is employed to generate the same number of non-waterlogging points as negative samples, each marked as "0". The marks of positive and negative samples are then input to train the XGBoost model.

3.1.2. Satellite Images Data

In this study, ground objects in satellite images are the primary means to identify the factors affecting the waterlogging in coastal cities. Therefore, the ground objects in the satellite images of the sample points, such as roads, water bodies, and green spaces, are recognized and extracted. Then the number of pixels of each object category is taken as an indicator to measure the waterlogging risk factors and construct a feature dataset of satellite images. By using the geocoding and reverse geocoding of Tianditu, a national platform for common geospatial information services, satellite images with a resolution of 1024*1024 are captured with Tianditu according to the determined coordinates of positive and negative sample points (each sample point is taken as the center for image capturing). In some cases, as the waterlogging point in road waterlogging events cannot be determined, the geometric center of the road is taken as the waterlogging point [21,22].

3.1.3. Elevation Data

Elevation is the most direct manifestation of floods. The frequency of waterlogging in a region generally increases with decreases in elevation. Therefore, in this study, elevation is selected as one of the factors affecting urban waterlogging, and most of the elevation data are downloaded from the Geospatial Data Cloud (http://www.gscloud.cn/search (accessed on 1 June 2021)). Specifically, after downloading the tif file of elevation from the Cloud (with each sample point, either positive or negative, taken as the center), the elevation data of each center point are extracted from the tif file using a program to obtain the elevation value of each sample point.

As the coastal cities differed from each other with regard to the overall elevation, relative elevation is adopted in this study as an extra risk factor for the urban waterlogging to make the elevation data comparable among the coastal cities. The relative elevation is obtained by subtracting the average elevation of the four vertices from the elevation value of the center point calculated above.

3.2. Research Methods

3.2.1. U-Net Semantic Segmentation Model

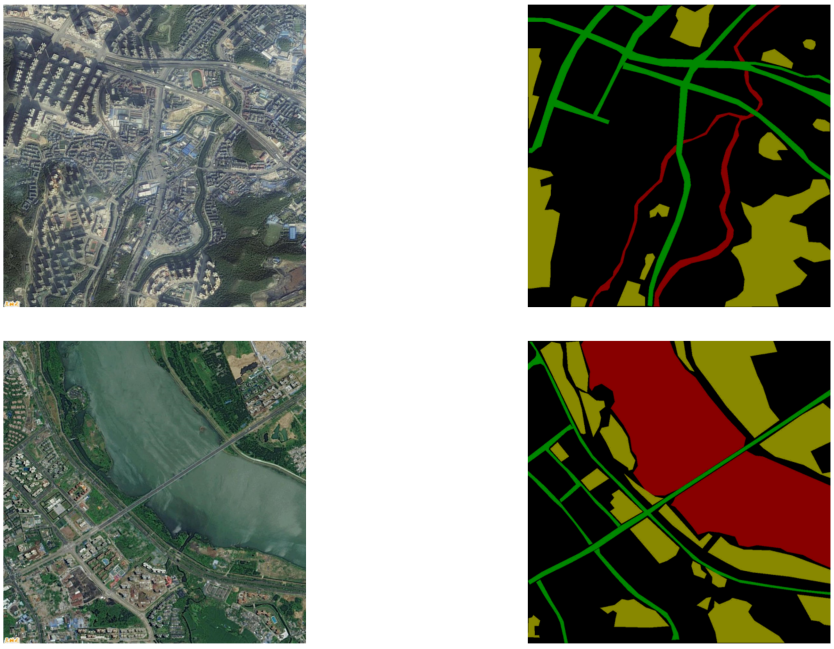
The model structure with semantic segmentation plays a key role in ground object classification algorithms, and the semantic segmentation model that incorporates the underlying features has great advantages, so a U-shaped network structure has been constructed on this basis so that the convolution results of each layer of the model are involved in the final feature fusion. This U-shaped semantic segmentation network is

known as the U-net [17]. In this study, an improved U-net semantic segmentation model is employed to classify ground objects such as water bodies, roads, and green spaces in satellite images. For each pixel in the input satellite image, the model determined the category to which the pixel belonged, and finally outputted the prediction result. Specific steps are as follows:

(1) Construct a training dataset

In the training for semantic segmentation of satellite images, the performance of the model is inextricably linked to the quality of the training data. Generally speaking, the training data fall into two categories, supervised and unsupervised learning, according to whether the data have been manually annotated or not. The semantic segmentation model used in this paper belonged to the supervised learning category, so a satellite image dataset with manual annotation has to be constructed for the training of the semantic segmentation model.

The annotation tool LabelMe is utilized in this paper to mark and annotate ground objects such as water bodies, roads, and green spaces in satellite images, and the final results are 200 satellite images and 200 annotated images. Data pair examples of the training dataset are shown in Figure 2, and each data pair shows the satellite image on the left and the manually annotated image on the right.



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

Figure 2. Examples of training dataset: (**a**) A satellite remote sensing image; (**b**) A manual tag map.

(2) Model training

The dataset used to train the model in this paper is a high-resolution satellite image dataset created by us, which had to be preprocessed before the model could be trained. The dataset is first normalized by subtracting the mean from each image and removing the variance to ensure that images with too much data variation could have the same scale of distribution. Then, in the course of training, the image data are processed in a random manner, including image flipping, panning, and zooming, with the aim of improving the robustness of the model and reducing overfitting of the model. Finally, the larger the batch size is, the more representative of the overall distribution characteristics of the dataset. However, due to the limitations of computer capabilities, the entire dataset could not be loaded at once in a semantic segmentation task. In view of this, the batch size (batch) of the input data is set to 8 with due regard to the hardware performance of the server.

The learning rate is not only the iteration step size of a deep learning model, but also one of the most important hyperparameters when training a model. If the learning rate is set too small, the model usually converges too slowly and tends to converge to a local optimum, while on the contrary, if the learning rate is too large, the model tends not to converge. Therefore, the strategy for setting the learning rate is of paramount importance in training a model. Compared to nonrandom algorithms, stochastic gradient descent (SGD) utilizes information more effectively, especially when it is redundant, and has better performance in early iterations. Moreover, SGD has an advantage over nonrandom algorithms in computational complexity with large samples [23]. In this paper, the stochastic gradient descent (SGD) algorithm is chosen as the strategy for parameter updating, and the initial learning rate for training is set as 0.0001. In a semantic segmentation model, a loss function is usually used to measure the training effect of the model and to perform gradient optimization. The smaller the value of the loss function, the higher the accuracy of the model and the better the training effect, so the selection of the loss function is of prime importance to the model training. A properly selected loss function will lead to a steady improvement in the predicted results of the model. In this paper, cross-entropy loss and Dice loss are selected as the loss functions for investigation. The Dice coefficient is a statistic used to gauge the similarity of two samples, indicating the degree of overlap between the predicted results and the true results. Its possible values are in the interval (0, 1), and the larger the value, the better. As such, Dice Loss = 1-Dice is taken as the loss function for semantic segmentation. The calculation formula is as follows:

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

where X denotes the predicted results and Y denotes the true results.

Finally, the final output of the model is usually a convolution result with variable value, while the result obtained from the semantic segmentation task is a probability value representing the category to which the pixel belonged, so the output of the model has to be normalized to a number between 0 and 1. In addition, the values of each pixel on all channels should sum to 1, representing the sum of the probabilities of each pixel's overall output categories equaled 1. To this end, the output result of the softmax function normalization model is selected in this paper:

$$y_i = \frac{x_i}{\sum_{k=1}^{C} x_i} \tag{2}$$

where x_i denotes the output value of pixel, C denotes the number of channels, and y_i denotes the predicted probability.

(3) Model prediction

Since the original U-net semantic segmentation model is mainly used for the segmentation of single-channel biomedical images, VGG-16 [24] is taken by the improved model in this paper as the encoding network, with its network structure detailed in Table 1. In this study, Conv1 to Conv5 are selected as the encoding structure of U-net, where each Conv comprised two 3 × 3 convolutions and two ReLU activation functions. Each Conv is followed by a max pooling layer, the main purpose of which is to extract important features from the input of the upper layer and reduce the number of feature parameters, and the operation principle is to select the extreme values on a fixed region from the input features of the upper layer. The max pooling selected in this paper is 2 × 2 max pooling with a step size of 2, i.e., the maximum in a 2 × 2 region is selected from the input vector matrix, and the final output width and height are half the input features.

Table 1. Network structure and parameters of VGG-16 model.

Name	Output Channel	Operation
Conv1	64 64	3 × 3 conv + ReLU 3 × 3 conv + ReLU
Pooling1	64	2 × 2 max pooling, stride = 2
Conv2	128 128	3 × 3 conv + ReLU 3 × 3 conv + ReLU
Pooling2	128	2 × 2 max pooling, stride = 2
Conv3	256 256 256	3 × 3 conv + ReLU 3 × 3 conv + ReLU 3 × 3 conv + ReLU
Pooling3	256	2 × 2 max pooling, stride = 2
Conv4	512 512 512	3 × 3 conv + ReLU 3 × 3 conv + ReLU 3 × 3 conv + ReLU
Pooling4	512	2 × 2 max pooling, stride = 2
Conv5	512 512 512	3 × 3 conv + ReLU 3 × 3 conv + ReLU 3 × 3 conv + ReLU
Pooling5	512	2 × 2 max pooling, stride = 2
Conv6	4096	7 × 7 conv + ReLU
Classifier	4096 1000	fully connection + ReLU fully connection

In the decoding structure, to facilitate the construction of the network and for better versatility, the U-net used in this paper superimposes the outputs of Conv1-Conv4 onto 2 times upsampled results of the output features of the decoder, thus obtaining a feature layer with the height and width the same as those of the input image. The detailed structure of U-net is shown in Figure 3.

3.2.2. Extreme Gradient Boosting (XGBoost) Model

The extreme gradient boosting (XGBoost) model is a decision tree-based integrated machine learning algorithm proposed by CHEN [25], which is based on classification and regression trees (CART) to classify and predict datasets. XGBoost is employed in this study to train and predict integrated datasets, and its prediction process is as follows:

Step 1: Construct a dataset containing n samples and m features, $|D| = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, and the predicted output of the integrated model is expressed as:

$$\hat{y}_i = \varnothing(x_i) = \sum_{k=1}^{K} f_k(x_i) \tag{3}$$

where x_i denotes the ith sample, $\hat{y}_i$ denotes the prediction value of the ith sample x_i, f_k denotes the Kth regression tree, and K denotes the number of regression trees. Equation (2) indicates that given an input x_i, the output value is the sum of the predicted values of K regression trees (i.e., the weights of the leaf nodes divided according to the decision rules of corresponding regression trees).

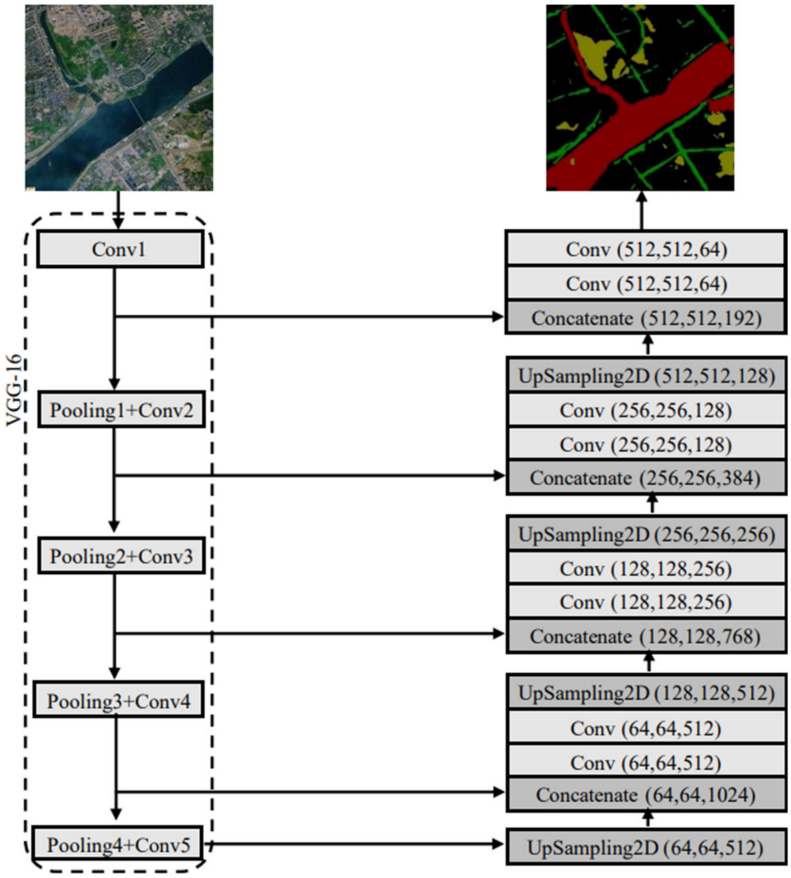


Figure 3. Model Structure of U-net.

Step 2: Define the objective function. The objective function of XGBoost is composed of a loss function and a regularization term, and defining the objective function is to define the loss function and the regularization term. The loss function is used to fit the training data, while the regularization term is used to control the model complexity, and the equation is as follows:

$$L(\varnothing) = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k} \Omega(f_k) \tag{4}$$

where $L(\varnothing)$ denotes the objective function, $L(y_i, \hat{y}_i)$ denotes the loss function, $\Omega(f_k)$ is the regularization term, and y_i denotes the true value of the sample.

Step 3: Optimize the objective function. A forward distribution algorithm is used to optimize the objective function. Supposing $\hat{y}_i^{(t)}$ is the predicted value of the ith sample after tth iteration (the tth tree), then:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{5}$$

where $\hat{y}_i^{(t)}$ denotes the predicted value of the ith sample after tth iterations, $\hat{y}_i^{(t-1)}$ denotes the predicted value of the ith sample after $t-1$ iteration, and $f_t(x_i)$ denotes the predicted value of the tth tree.

Therefore, the objective function can be expressed as:

$$L(\varnothing) = \sum_{i=1}^{n} L\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \sum_{k} \Omega(f_k) \tag{6}$$

where $L(\varnothing)$ denotes the objective function, $L\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right)$ denotes the loss function, $\Omega(f_k)$ is the regularization term, and y_i denotes the true value of the sample.

Step 4: Optimize the loss function in the objective function. The second-order Taylor expansion is used to expand the loss function to approximate to the true value. Its equation is as follows:

$$L\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) \approx L\left(y_i, \hat{y}_i^{(t-1)}\right) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \tag{7}$$

where $L\left(y_i, \hat{y}_i^{(t-1)}\right)$ denotes the loss value of the ith sample from the preceding $t-1$ trees, $g_i f_t(x_i)$ is the first-order partial derivative of $\hat{y}_i^{(t-1)}$, and $\frac{1}{2} h_i f_t^2(x_i)$ is the second-order partial derivative of $\hat{y}_i^{(t-1)}$.

Step 5: Obtain the final objective function, which is as follows:

$$L\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) \approx L\left(y_i, \hat{y}_i^{(t-1)}\right) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \tag{8}$$

3.2.3. Performance Evaluation Metrics

(1) Confusion matrix:

A confusion matrix is an important tool for evaluating the performance of a classification model, with each column representing the instances in a predicted class while each row representing the instances in an actual class. The four metrics used in the analysis are True Positive (TP), False Positive (FP), False Negative (FN) and True Negative (TN). The confusion matrix is manifested by Table 2.

Table 2. Manifestation of the confusion matrix.

Confusion Matrix		**Actual Classes**	
		Positive	**Negative**
Predicted Classes	Positive	TP	FP
	Negative	FN	TN

(2) Performance metrics for the semantic segmentation model:

The semantic segmentation model is mainly evaluated using two metrics, Mean Pixel Accuracy (MPA) and Mean Intersection over Union (mIoU), whose expressions are as follows:

$$MPA = \frac{1}{k+1} \sum_{i=0}^{k} \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

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

where $k+1$ denotes $k+1$ categories, and TP, TN, FP and FN denote correctly identified positive sample, correctly identified negative sample, incorrectly identified positive sample, and incorrectly identified negative sample, respectively.

(3) Performance metrics for the XGBoost model;

The XGBoost model is mainly evaluated using 4 metrics, accuracy (ACC), precision (P), recall (R) and F-score (F1), all of which could be obtained by calculation from the confusion matrix. F-score (F1) indicates the harmonic mean of precision and recall values [26]. It can be calculated by the following formula:

$$ACC = (TP + TN)/(TP + FN + FP + TN) \tag{11}$$

$$P = TP/(TP + FP) \tag{12}$$

$$R = TP/(TP + FN) \tag{13}$$

$$F1 = 2(P \times R)/(P + R) \tag{14}$$

In addition, the receiver operating characteristic (ROC) curve is plotted to measure the area under the ROC curve (AUC), which is then used to determine the accuracy of the classification results of the binary classification model [27]. AUC < 0.6 indicates the model has a poor predictive ability; 0.6 < AUC < 0.7 indicates the model has a moderate predictive ability; 0.7 < AUC < 0.8 indicates the model's predictive ability is good; and AUC > 0.8 indicates the model's predictive ability is excellent.

4. Results

4.1. Performance Evaluation for the Semantic Segmentation Model

In designing the experiments, the program is designed using the PyTorch framework with the learning rate set to the initial 0.0001, and the cross-entropy loss and Dice loss are employed as the loss functions. Moreover, in order to improve the robustness of the model and reduce the overfitting, the image data is processed in a random manner, including image flipping, panning, and zooming, thus realizing data augmentation on the dataset. Finally, the training dataset is randomly divided into a training dataset and a validation dataset in the ratio of 9:1. The role of the validation dataset is to verify the predictive power of the model and is not involved in the model training. After training, the training effect and results of the model are judged by the prediction accuracy of the semantic segmentation model on the training and validation datasets. The results are given in Table 3.

Table 3. Performance metrics for semantic segmentation model based on validation dataset.

Metric	Water Bodies	Roads	Green Spaces	Others	Overall
mPA	87.94	66.21	82.92	90.49	81.89
mIoU	82.72	45.43	71.29	82.83	70.57

As can be seen from Table 3, the model achieves an overall mPA of 81.89% and an mIoU of 70.57%, indicating the model could effectively recognize ground objects in satellite images. However, the mPA and mIoU values for the category of roads are somewhat low. The reason for this would be that the roads in the satellite images are so obscured by the nearby buildings, trees, and their shadows in the sun that the model is unable to clearly identify the obscured roads, resulting in low values.

Figure 4 shows the performance of the semantic segmentation model extracting ground objects from the satellite images, indicating the model could accurately extract and distinguish among different feature categories such as water bodies, roads, and green spaces in the satellite images. The images are rendered very clearly.

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

Figure 4. Example of semantic segmentation: (**a**) Satellite image to be predicted; (**b**) Annotated image; (**c**) Prediction result.

4.2. Performance Evaluation for the XGBoost Model

Five-fold cross-verification [28] is employed in this paper to evaluate the model performance during the experiment, and the average of the five results is used as the metric for performance evaluation, as listed in Table 4.

Table 4. Five-fold cross-validation for performance evaluation of XGBoost model based on test dataset.

Metric	ACC	P	R	F1	AUC
1	0.82	0.80	0.88	0.84	0.89
2	0.83	0.82	0.83	0.82	0.91
3	0.79	0.75	0.88	0.81	0.85
4	0.79	0.71	0.89	0.79	0.87
5	0.83	0.83	0.86	0.84	0.90
Average	0.81	0.78	0.87	0.82	0.88

As can be seen from Table 4, the mean accuracy of the five cross-validations is 0.81, indicating that 81% on average of the samples in the test dataset could be correctly identified. The mean precision of the five cross-validations is 0.78, indicating that only a few non-waterlogging points from the positive samples are incorrectly identified as waterlogging points, which showed that the XGBoost has a good fitting over urban waterlogging points. The mean F1 of the five cross-validations is 0.82, showing that the model performed well. The mean AUC of the five cross-validations is 0.88, indicating the XGBoost has perfect prediction accuracy. Random Forest (RF), Logistic Regression (LG) and Support Vector Machine (SVM) are the common machine learning models in urban waterlogging. In order to verify the effectiveness and reliability of the proposed model, the AUC of XGboost, RF, LG, and SVW is shown in Table 5.

Table 5. Prediction accuracy of the XGBoost, RF, LG, and SVM model based on the test.

Model	XGBoost	RF	LG	SVM
AUC	0.88	0.83	0.86	0.84

As can be seen from Table 5, the prediction accuracy of the XGBoost is better than RF, LG, and SVM.

Based on the results above, we conclude that the XGBoost can be further used in the significance analysis and study of urban waterlogging factors.

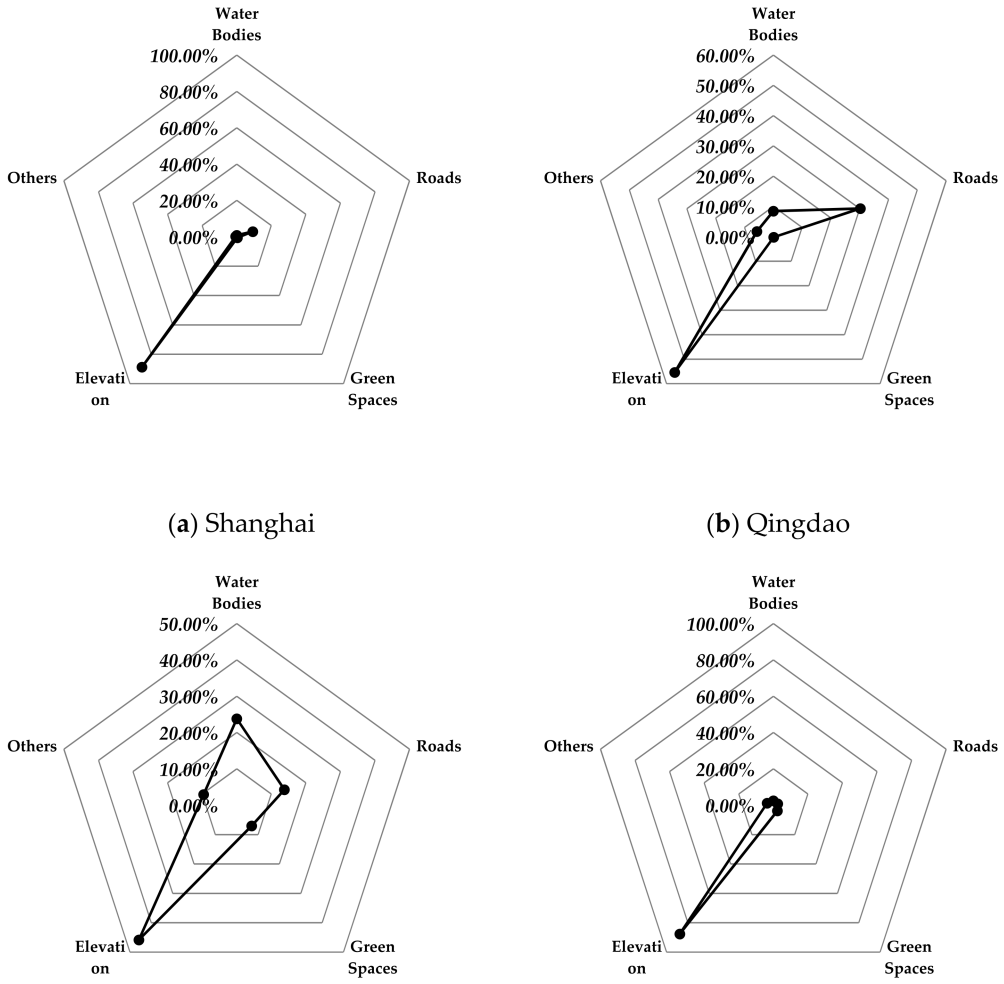
4.3. Analysis of Influencing Factors of Urban Waterlogging

To further probe into the primary factors affecting urban waterlogging in the research areas, the weight of each indicator is analyzed under the XGBoost model. For urban waterlogging points, the significance of each indicator is shown in Table 6. It can be seen from the table that there are significant differences among the factors affecting the urban waterlogging in the research areas. Among them, elevation is the primary factor affecting waterlogging in the four coastal cities, with a significance higher than 40%, which matches with the conclusions of most scholars [29,30]. This is because low-lying areas are more likely to become catchments than high-lying areas. Moreover, due to the increasing impermeable pavements and little vegetation coverage in the main urban area of the cities, which narrow the storm water infiltration area and cause increasing storm water runoff, stagnant water is more likely to form in low-lying areas.

Table 6. Significance of indicators of urban waterlogging points.

City	Water Bodies	Roads	Green Spaces	Elevation	Others
Haikou	0.238	0.138	0.070	0.458	0.096
Xiamen	0.025	0.025	0.038	0.876	0.036
Shanghai	0.008	0.093	0.006	0.887	0.006
Qingdao	0.085	0.302	0.002	0.554	0.057

The significance values of urban waterlogging factors are shown in Figure 5. The secondary factor affecting the waterlogging in both Shanghai and Qingdao is roads. As an example of urban impermeable surfaces, many hard-surface roads not only cut off the hydrological process between surface water and groundwater, but also change the original runoff generation, pooling, and flow conditions of the urban surface, which in turn affects the urban water circulation and increases the risk of waterlogging. For example, as of 2019, the land for roads and transport infrastructure in Shanghai accounts for more than 15% of the construction land, and the road network density in Huangpu and Hongkou is even higher (>8 km/km^2) [31]. High-density road construction significantly induces waterlogging on roads in main urban areas.



(**a**) Shanghai (**b**) Qingdao

(**c**) Haikou (**d**) Xiamen

Figure 5. Significance of waterlogging factors in each city under the XGBoost model.

For Haikou, water bodies are the secondary factor affecting urban waterlogging. As Haikou is a seaside city, the seawater level is higher than the river outlet level during heavy rainfall or at high tide, resulting in the backflow of sea water into rivers. In this case, the stagnant water in the city will be prevented from being discharged into the sea through rivers and ditches, which will then worsen the waterlogging situation. Haikou is a typical tropical coastal city established along a river. The city is impacted by typhoons and tropical storms all year round, with a very high annual rainfall. These climate conditions would have been more prone to waterlogging. However, the primary reason for waterlogging is that as the city develops, many ponds and wetlands around and in the city have been filled in for development, while the construction of supporting municipal facilities is lagging severely behind. The loss of natural control and detention capacity causes waterlogging, for example, in areas around Qiongshan Avenue.

For Xiamen, the four indicators other than the elevation are equally significant, which could all be regarded as secondary factors affecting the urban waterlogging. Most of the flood drainage systems on Xiamen Island are planned and constructed in 1980s. With the changes of times and the development of the city, the underground flood drainage systems have been damaged, the subsidence of land and pipe network has been aggravated, and the urban green spaces have shrunk. All of these have enormously weakened Xiamen's overall flood carrying capacity.

5. Conclusions and Policy Implications

5.1. Conclusions

The satellite image semantic segmentation model based on the U-net is capable of effectively recognizing ground objects in satellite images, while accurately extracting and distinguishing among different feature categories such as water bodies, roads, and green spaces in satellite images. The XGBoost model is capable of predicting the waterlogging points from the samples with high prediction accuracy and of analyzing the urban waterlogging risk factors by weighing each indicator. Moreover, the AUC of XGBoost model is 0.88 and larger the other common machined learning model, indicating the XGBoost has perfect prediction accuracy. Integrating the satellite image semantic segmentation model and XGBoost model provides a brand-new perspective to evaluate the urban waterlogging risk. Regarding the urban waterlogging predication, more attention should be paid to the elevation, which is the primary factor affecting the urban waterlogging.

It should be mentioned that more satellite image datasets should be collected and created, and more categories of ground objects and features that affect the urban waterlogging risk should be selected to improve the training accuracy of the semantic segmentation model and the recognition accuracy in satellite images. At the same time, other deep learning models should be considered to further integrate satellite images and the data from social networks to probe into urban waterlogging risk factors.

5.2. Policy Implications

Based on the assessment results of urban waterlogging risk factors in the research areas, the following countermeasures and recommendations are put forward in this paper for early warning, prevention, and control of urban waterlogging in the research areas:

Firstly, the four coastal cities, Haikou, Xiamen, Shanghai, and Qingdao, differ from each other in regard to the significance of the influence factors. Therefore, in urban waterlogging control, it is necessary to strengthen engineering measures and build a systematic urban drainage and waterlogging prevention system according to local conditions. In addition, it is necessary to improve the city's overall drainage capacity with reference to the construction experience of "sponge city" to reduce waterlogging.

Second, due to the high risk of waterlogging in the coastal cities with dense river networks, it is necessary to in real time monitor the changes in the water level of lakes and rivers and clean up and dredge the waterways in time to protect the lake water systems in cities and improve the water detention and flood control capacity of urban water systems.

Moreover, the proportion of urban green spaces should be increased to enhance the surface water circulation and reduce the flood control pressure on urban rivers.

Thirdly, the government should comprehensively improve its emergency management for drainage and waterlogging prevention and improve and optimize the contingency plans concerning urban drainage and waterlogging prevention. The inspection, maintenance, and potential hazard identification system for urban drainage and waterlogging prevention facilities as well as safe operation procedures should be implemented strictly as required. It should be ensured that potential hazards are fully investigated, identified, and eliminated before the flood season. In addition, more effort should be put into routine maintenance on drainage facilities. Moreover, the government may launch catastrophe insurance, social assistance, and other safeguard mechanisms to improve the resilience of residents and enterprises against disasters.

Author Contributions: Conceptualization, J.T. and H.L.; methodology, F.G., H.L. and H.Z.; software, F.G., J.H. and G.L.; data curation, F.G. and Q.D.; writing- original draft, F.G.; writing—reviewing & editing, F.G. and Q.D.; supervision, J.T. and H.L.; funding acquisition, J.T., H.L. and F.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (Grant number 91846203, 42171081 and 72174504); Zijin Cultural Talent Project in Jiangsu Province (Grant number PDJPC [2020] No.96); Graduate Student Scientific Research Innovation Projects in Jiangsu Province (Grant number KYCX22_2980).

Data Availability Statement: The source code and manually collected Weibo posts that report flood deposits can be downloaded from Github: https://github.com/hliu2016/waterlogging (accessed on 30 March 2021).

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

References

1. Zhang, J.; Wang, Y.; He, R.; Hu, Q.; Song, X. Discussion on the urban flood and waterlogging and causes analysis in China. *Adv. Water Sci.* **2016**, *27*, 485–491, (In Chinese with English Abstract).
2. Yang, P.; Jin, J.; Zhao, D.; Li, J. An Urban Vulnerability Study Based on Historical Flood Data: A Case Study of Beijing. *Sci. Geogr. Sin.* **2016**, *36*, 733–741, (In Chinese with English Abstract).
3. Lu, H.; Zhou, Y.; Sun, J.; Zhu, Q.; Niu, S.; Li, X. Simulation of waterlogging control effect in Tiexinqiao experimental base based on SWMM. *Water Resour. Prot.* **2020**, *36*, 58–65, (In Chinese with English Abstract).
4. Quan, R. Exposure Analysis of Rainstorm Waterlogging on Subway in Central Urban Area of Shanghai Based on Multiple Scenario Perspective. *Sci. Geogr. Sin.* **2015**, *35*, 471–475, (In Chinese with English Abstract).
5. Chen, J.; Li, Q.; Deng, M.; Pei, J. Urban flood vulnerability assessment based on random forests and variable fuzzy sets. *Resour. Environ. Yangtze Basin* **2020**, *29*, 2551–2562, (In Chinese with English Abstract).
6. Huang, G.; Luo, H.; Lu, X.; Yang, C.; Wang, Z.; Huang, T.; Ma, J. Study on risk analysis and zoning method of urban flood disaster. *Water Resour. Prot.* **2020**, *36*, 1–6, (In Chinese with English Abstract).
7. Sun, Q.; Fang, J.; Dang, X.; Xu, K.; Fang, Y.; Li, X.; Liu, M. Multi-scenario urban flood risk assessment by integrating future land use change models and hydrodynamic models. *Nat. Hazards Earth Syst. Sci.* **2022**, *22*, 3815–3829. [CrossRef]
8. Gentilucci, M.; Barbieri, M.; Burt, P.; D'Aprile, F. Preliminary Data Validation and Reconstruction of Temperature and Precipitation in Central Italy. *Geosciences* **2018**, *8*, 202. [CrossRef]
9. Lin, L.; Wu, Z.; Liang, Q. Urban flood susceptibility analysis using a GIS-based multi-criteria analysis framework. *Nat. Hazards* **2019**, *97*, 455–475. [CrossRef]
10. Gentilucci, M.; Barbieri, M.; Lee, H.S.; Zardi, D. Analysis of Rainfall Trends and Extreme Precipitation in the Middle Adriatic Side, Marche Region (Central Italy). *Water* **2019**, *11*, 1948. [CrossRef]
11. Gentilucci, M.; Barbieri, M.; Pambianchi, G. Reliability of the IMERG product through reference rain gauges in Central Italy. *Atmos. Res.* **2022**, *278*, 106340. [CrossRef]
12. Wang, Z.; Lai, C.; Chen, X.; Yang, B.; Zhao, S.; Bai, X. Flood hazard risk assessment model based on random forest. *J. Hydrol.* **2015**, *527*, 1130–1141. [CrossRef]
13. Zhu, Q.; Zhong, Y.; Zhao, B.; Xia, G.; Zhang, L. Bag-of-Visual-Words Scene Classifier With Local and Global Features for High Spatial Resolution Remote Sensing Imagery. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 747–751. [CrossRef]
14. Pal, M. Ensemble of support vector machines for land cover classification. *Int. J. Remote Sens.* **2008**, *29*, 3043–3049. [CrossRef]
15. Chen, Y.; Fan, R.; Yang, X.; Wang, J.; Latif, A. Extraction of Urban Water Bodies from High-Resolution Remote-Sensing Imagery Using Deep Learning. *Water* **2018**, *10*, 585. [CrossRef]

16. Shelhamer, E.; Long, J.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 640–651. [CrossRef]
17. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
18. Lv, Y. *Research and Application of Remote Sensing Image Feature Extraction Technology Based on Deep Learning*; Beijing University of Posts and Telecommunications: Beijing, China, 2020; (In Chinese with English Abstract).
19. Xu, S.; Wang, J.; Shi, C.; Yan, J. Research on natural disaster risk in coastal cities. *Acta Geogr. Sin.* **2006**, *2*, 127–138. (In Chinese)
20. Liu, J.; Li, Z.; Mei, C.; Wang, K.; Zhou, G. Urban flood analysis for different design storm hyetographs in Xiamen Island based on TELEMAC-2D. *Chin. Sci. Bull.* **2019**, *64*, 2055–2066, (In Chinese with English Abstract). [CrossRef]
21. Yu, H.; Zhao, Y.; Fu, Y.; Li, L. Spatiotemporal Variance Assessment of Urban Rainstorm Waterlogging Affected by Impervious Surface Expansion: A Case Study of Guangzhou, China. *Sustainability* **2018**, *10*, 3761. [CrossRef]
22. Zhang, H.; Wu, C.; Chen, W.; Huang, G. Assessing the Impact of Climate Change on the Waterlogging Risk in Coastal Cities: A Case Study of Guangzhou, South China. *J. Hydrometeorol.* **2017**, *18*, 1549–1562. [CrossRef]
23. Bottou, L.; Curtis, F.; Nocedal, J. Optimization Methods for Large-Scale Machine Learning. *SIAM Rev.* **2018**, *60*, 223–311. [CrossRef]
24. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2015**, arXiv:1409.1556.
25. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 13–17 August 2016.
26. Zhao, G.; Pang, B.; Xu, Z.; Yue, J.; Tu, T. Mapping flood susceptibility in mountainous areas on a national scale in China. *Sci. Total Environ.* **2018**, *615*, 1133–1142. [CrossRef]
27. Schumann, G.; Vernieuwe, H.; De Baets, B.; Verhoest, N. ROC-based calibration of flood inundation models. *Hydrol. Process.* **2014**, *28*, 5495–5502. [CrossRef]
28. Geisser, S. The predictive sample reuse method with applications. *J. Am. Stat. Assoc.* **1975**, *70*, 320–328. [CrossRef]
29. Lei, X.; Chen, Y.; Pan, X.; Zhang, L.; Li, Y.; Hu, T. Risk zonation of rainstorm flooding disaster in Hangzhou Main City Zone. *J. Hangzhou Norm. Univ. (Nat. Sci. Ed.)* **2019**, *18*, 105–112, (In Chinese with English Abstract).
30. Mahyat, S.; Simon, J.; Farzin, S. Identifying the essential flood conditioning factors for flood prone area mapping using machine learning techniques. *CATENA* **2019**, *175*, 174–192.
31. Urban Traffic Engineering Technology Center of the Ministry of Housing and Urban Rural Development. *Annual Report on Road Network Density in Major Chinese Cities*; Urban Traffic Engineering Technology Center of Ministry of Housing and Urban-Rural Development; China Academy of Urban Planning and Design; NavInfo Co., Ltd.: Beijing, China, 2020. (In Chinese)