sustainability

MDPI

Article

Multi-Attention Network for Sewage Treatment Plant Detection

Yue Shuai [1,2], Jun Xie [3], Kaixuan Lu [1] and Zhengchao Chen [1,*]

[1] Airborne Remote Sensing Center, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China
[2] University of Chinese Academy of Sciences, Beijing 100049, China
[3] College of Geoscience and Surveying Engineering, China University of Mining and Technology, Beijing 100083, China
[*] Correspondence: chenzc@radi.ac.cn

Abstract: As an important facility for effectively controlling water pollution discharge and recycling waste water resources, accurate sewage treatment plant extraction is very important for protecting quality, function, and sustainable development of the water environment. However, due to the presence of rectangular and circular treatment facilities in sewage treatment plants, the shapes are diverse and the scales are different, resulting in the poor performance of conventional object detection algorithms. This paper proposes a multi-attention network (MANet) for sewage treatment plants using remote sensing images. MANet consists of three major components: a light backbone used to obtain multi-scale features, a channel and spatial attention module that realizes the feature representation of the channel dimension and spatial dimension, and a scale attention module to obtain scale-aware features. The results from the extensive experiments performed on the sewage treatment plant dataset suggest that our proposed MANet exhibits a superior performance compared with other competing methods. Meanwhile, we used a well-trained model to predict the sewage treatment plant from the GF-2 data for the Beijing area. By comparing the results with the data of manually obtained sewage treatment plants, our method can achieve an accuracy of 80.1% while maintaining the recall rate at a high level (90.4%).

Keywords: deep learning; sewage treatment plant detection; Beijing area; attention module; remote sensing images

check for updates

Citation: Shuai, Y.; Xie, J.; Lu, K.; Chen, Z. Multi-Attention Network for Sewage Treatment Plant Detection. *Sustainability* **2023**, *15*, 5880. https://doi.org/10.3390/su15075880

Academic Editors: Marc A. Rosen and Andrea G. Capodaglio

Received: 15 February 2023
Revised: 24 March 2023
Accepted: 27 March 2023
Published: 28 March 2023

1. Introduction

Wastewater treatment contributes to the achievement of 11 of the 17 Sustainable Development Goals that have currently been adopted globally [1]. As the main carrier of wastewater treatment, sewage treatment plants are important assistants for effectively curbing sewage discharge and recycling wastewater in industrial society, and they are increasingly important for water quality protection and the sustainable development of man and nature [2]. They are especially important in world-class cities such as Beijing, where a large amount of industrial sewage and domestic wastewater are produced every day; if they were directly discharged without being treated by the sewage treatment plant, they would cause a huge disaster to the natural environment and further affect people's lives, which is not conducive to the sustainable development of man and nature [3,4]. Realizing the automatic extraction of large-scale sewage treatment plants will provide basic data support for people to study the details of sewage treatment plants and provide further technical support for the realization of the Sustainable Development Goals.

The contributions made by sewage treatment plants are significant, but there are also negative problems. Due to the need to collect sewage for purification, the surrounding environment has a high level of pollutants, which has an adverse impact on the surrounding ecological environment and on people's lives [5,6]. Under the current conditions, we cannot immediately obtain the distribution information of sewage treatment plants in a certain area—especially on a large scale, such as the provincial or national level—or the sewage treatment plants that are built by some factories. In this study, we attempted to realize the automatic identification of large-scale sewage treatment plants through technical research, obtain their spatial location information and quantity information, and make this information easier for people or managers to obtain. Based on the relevant information of sewage treatment plants, they can provide a reference for the selection of the living location of the relevant population or allow people to further enhance the protection awareness of water resources with better understanding. In addition, the information can also provide a reference for city managers for optimizing the layout of sewage treatment plants and for the scientific construction of cities [7].

The goal of our technical research is to conduct a realization of large-scale sewage treatment plant extraction based on big data and computer vision methods. The convolutional neural network [8], designed by simulating the function of human neurons, has strong feature fitting and learning capabilities for input data through the stacking of network depths and the setting of nonlinear activation functions. Compared with the fully connected neural network, a convolutional neural network that extracts data features by setting the size of the convolution kernel area has a higher computing efficiency for image data; with the assistance of graphics computing hardware, it is the best choice for processing large image data [9]. The deep learning object detection algorithm based on the convolutional neural network has been rapidly developed after recent in-depth research; many classic algorithms have been sequentially proposed, such as Faster RCNN [10], SSD [11], RetinaNet [12], YOLO series [13–16], etc., and successfully applied in many fields of computer vision. A study of the feasibility detection algorithm that is based on the deep learning object detection algorithm and combined with the sewage treatment plant's characteristics will greatly improve the recognition efficiency and automation level of sewage treatment plants.

As a long-distance detection technology, remote sensing has the characteristics of wide monitoring ranges, short periods, and low costs [17]. It can be used as a technical means to objectively obtain sewage treatment plant distributions. In recent years, with the launch of a large number of satellites, it is very convenient to obtain large-scale, high-resolution, and short-period optical remote sensing image data. Using optical remote sensing data to carry out high-precision, high-frequency monitoring of sewage treatment plant times, extraction has become possible [18,19]. Different from natural images, optical remote sensing satellite images that are captured from the top-down view of the Earth contain rich and complicated ground object information. Directly transferring the object detection algorithm applied to natural images to optical remote sensing images will reduce the model's accuracy. As a building facility, sewage treatment plants contain modules such as circular and rectangular purification pools for filtering sewage, and the overall characteristics are consistent. However, there are also local differences in the scale and shape characteristics; furthermore, because there are many ground objects with similar characteristics, the characteristics cause certain challenges in detecting sewage treatment plants. There are a lot of studies on the detection difficulties caused by the multi-scale and large shape differences of objects similar to sewage treatment plants in optical satellite remote sensing images [20–23].

In view of the multi-scale characteristics of remote sensing ground objects, multi-scale information fusion modules are commonly designed for feature extraction. For example, the FPN [24] (feature pyramid network) can account for both deep and shallow features to preserve the multi-scale information of the object [25,26]. On the basis of multi-scale detection, Yan et al. [27] balanced the training weights of differently scaled objects for the loss function and strengthened the robustness of the algorithm to different scales. In addition, the attention mechanism introduced by the transformer model [28] has been proven to have a good effect in multi-scale object detection. Zhu et al. [29] used the transformer model to improve the prediction network of YOLOv5 and combine the self-attention mechanism to achieve multi-scale object detection. In view of the characteristics of the large differences in the shapes of remote sensing ground objects, the main method used adjusts the type and quantity of the anchor frames in the detection stage to adapt to different shapes of the same objects or multiple types of objects [25,30,31]. For example, reset the scale of the anchor box, the aspect ratio parameters, or increase the angle variable, etc., and use a deformable convolutional network [32–34] to adapt to the target shape. The disadvantage of these methods is that the increase in anchor frame parameters and addition of deformable convolution will add a large number of parameters to the network, increase the difficulty of the model training, and lead to unfavorable model convergence. The detection model based on key points can overcome the problem of large changes in the object's shape to a certain extent, but the detection accuracy is basically the same as that of the anchor frame method; there is still large room for improvement and optimization.

However, sewage treatment plant characteristics in remote sensing images are different from other ground objects, and the above work still cannot directly meet the detection needs of sewage treatment plants. In this paper, starting from the detection of sewage treatment plants in remote sensing images in Beijing, a MANet sewage treatment plant detection network is proposed to solve the problems of the large differences in the shape and scale of sewage treatment plants, as well as their inconsistent local features. MANet integrates the channel and spatial attention in the feature extraction module and innovates a scale attention algorithm for network feature optimization, which better solves for the detection difficulties of sewage treatment plants and greatly improves the interpretation of sewage treatment plant target precision. The main contributions of this paper are as follows:

(1) We introduced a lightweight channel and spatial attention module (CSAM) to improve the feature expression ability of the extracted target in the spatial and channel dimensions;

(2) We innovated a novel scale attention modeule (SAM) algorithm to improve the feature learning ability of the network at different levels for targets with large-scale changes;

(3) We added the above two attention modules based on RetinaNet, proposed a MANet sewage treatment plant detection network, and achieved better results in the dataset test. In the actual scene, based on GF-2 remote sensing images, the sewage treatment plant detection in the Beijing area was realized. The results show that our method can achieve an accuracy of 80.1% while maintaining the recall rate at a high level (90.4%).

2. Methodology

In this section, we introduce the architecture of our proposed approach, MANet (Section 2.1), the backbone for the feature extraction (Section 2.2), the channel and spatial attention module (Section 2.3), and the scale attention module (Section 2.4).

2.1. Model Overview

The attention mechanism in deep learning approaches imitates the human visual system. When a human being observes an object, they first quickly scan the entire area, select the target from the area, and invest more visual resources to obtain more detailed information; however, the neural network needs to scan each pixel when scanning an image. The attention module is used to ensure that the CNN learns and pays more attention to key features instead of learning useless background information. In the object detection task, the useful information refers to the target's location and category information on the image, which essentially uses the $C \times H \times W$ feature map as an input and provides $1 \times H \times W$ as the output attention map. This attention map is then element-wise multiplied with the input feature map to obtain a more refined and salient output. In general, the attention mechanism is mainly applied to the spatial dimension or the channel dimension and is integrated in the residual structure of the network.

Sewage treatment plants have different shapes, including rectangular and circular treatment devices. Targets usually appear in completely different shapes, rotations, and positions, and the spatial variation of targets needs to be considered. At the same time, sewage treatment plants have different scales, ranging from large to small. The scale variation of the targets needs to be considered. Due to the complexity of sewage treatment plants' characteristics, the traditional deep learning object detection model has missed detection and falsely detected plants in sewage treatment plant recognition [35]. By analyzing the characteristics of sewage treatment plants using high-resolution remote sensing images and aiming to resolve the detection difficulties of sewage treatment plants, we designed MANet. Its overall architecture is shown in Figure 1. It mainly includes three parts: (1) the feature extraction part, which contains a backbone for obtaining multi-scale feature maps; (2) CSAM, which includes the spatial attention module and channel attention module, which learns the best features of the target from the two dimensions of the space and channel; and (3) SAM, which is only processed in the feature layer dimension, which learns the relative importance of multiple semantic layers and enhances features at the appropriate level according to the scale of the object.

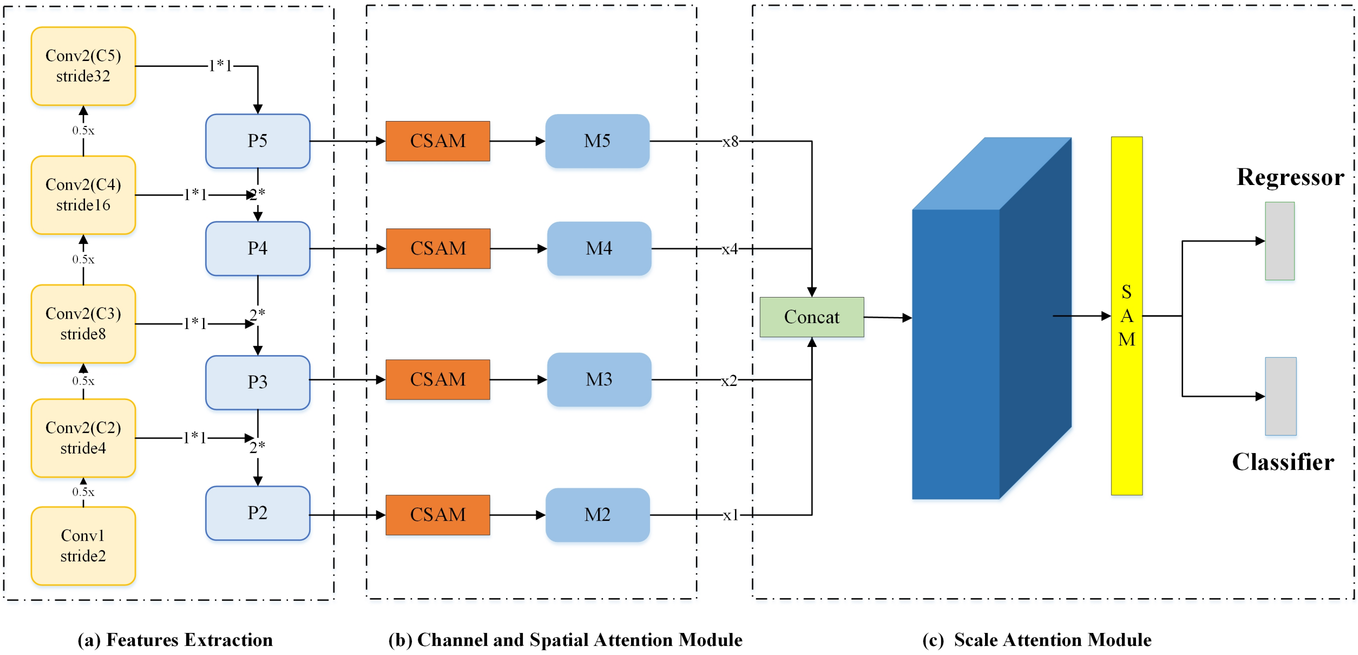

Figure 1. The framework of our proposed approach.

2.2. Feature Extraction

This part uses the structure of the ResNet+Feature Pyramid Network (FPN) to extract low-resolution features from images (input RGB images, size $H \times W \times 3$) and obtain multi-scale feature maps ($\frac{H}{S} \times \frac{W}{S} \times C$) through different stage steps (S4, 8, 16, and 32). In order to reduce the weight, this paper chooses a relatively simple ResNet-50+FPN structure. The basic structural unit of ResNet-50 is the residual structure. As shown in Figure 1, the entire network is divided into five blocks, namely conv1, conv2-x, conv3-x, conv4-x, and conv5-x. The convolution kernel size of conv1 is set to 7×7, the step size is set to 2, and the expansion is set to 3; then, the maximum pooling is performed. The pooled convolution kernel size is set to 3×3, the step size is set to 2, and the expansion setting is 0. There are three convolution blocks in the conv2-x part, where the convolution kernel sizes are set to 1×1, 3×3, and 1×1, respectively. The three parts of conv3-x, conv4-x, and conv5-x are similar to the structure of conv2-x; the difference is that the number of convolution blocks is different. The conv3-x part has four convolution blocks, the conv4-x part has six convolution blocks, and the conv5-x part has three convolution blocks. Finally, the multi-scale features are obtained after the blocks calculate the image data.

The structure of a light FPN typically consists of the following components: (1) top-down pathway: this pathway starts from the high-level semantic feature maps obtained from the backbone network and passes them through up-sampling operations to obtain feature maps at lower scales; (2) bottom-up pathway: this pathway starts from the low-level feature maps obtained from the backbone network and passes them through up-sampling operations to obtain feature maps at higher scales; (3) fusion layer: the feature maps from the top-down and bottom-up pathways are combined using element-wise summation or concatenation to obtain the final feature maps at each scale. By combining the features from multiple scales, the FPN can capture both the fine-grained details and the high-level context of the input image, leading to improved performance in object detection tasks.

2.3. Channel and Spatial Attention Module

In addition to the variable scale of sewage treatment plants mentioned in this paper, the characteristics of different shapes and colors make it difficult for the network to distinguish between them. Therefore, we introduce a CSAM to further optimize the features and cause these feature pairs to be more distinguishable. It is a very lightweight module that does not incur excessive memory and computational overhead. As shown in Figure 2, CSAM consists of two sub-modules, a channel attention module, and a spatial attention module to help strengthen the useful information in extracted features.

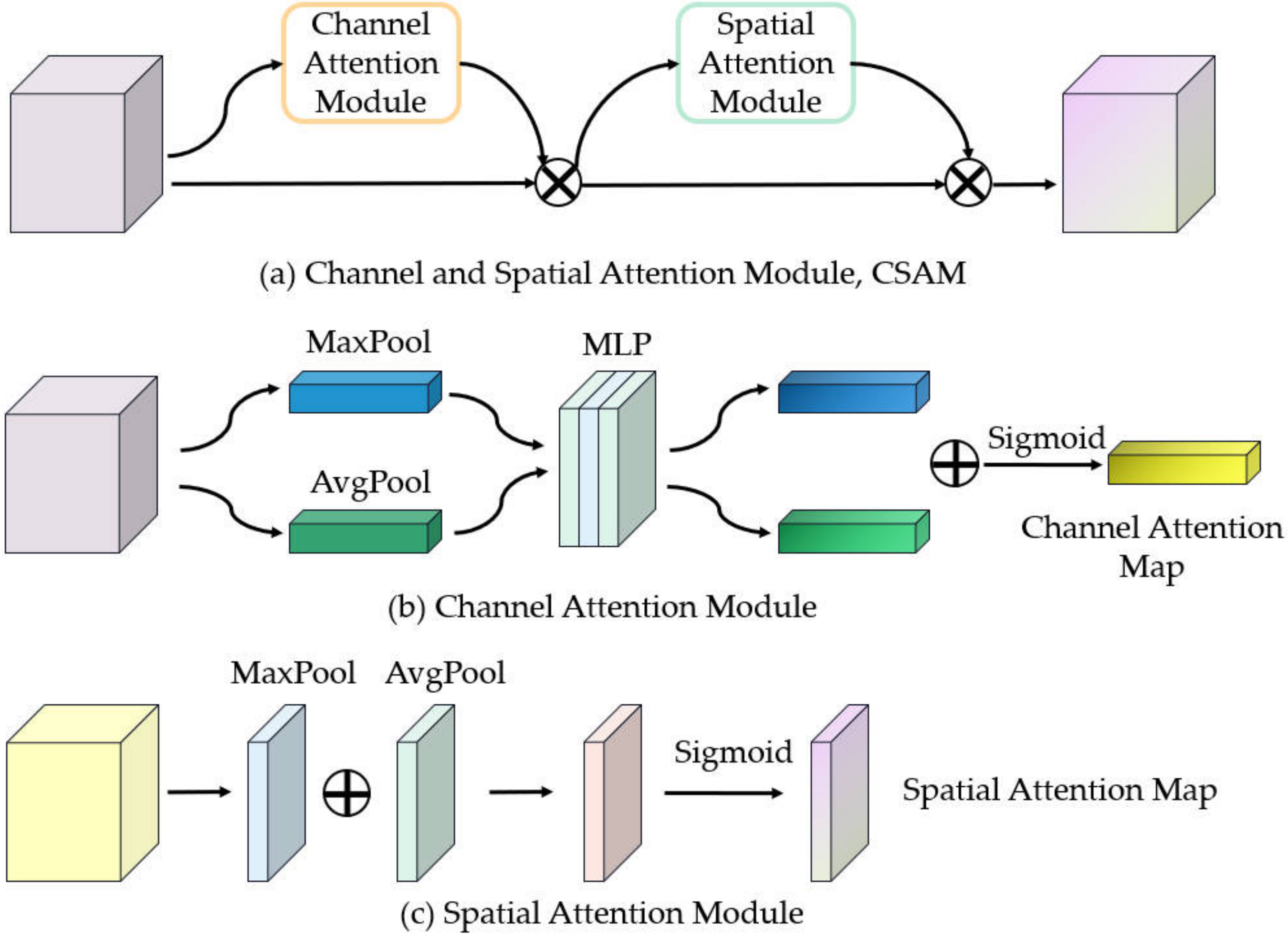

Figure 2. The structure of the channel and spatial attention module, CSAM. (**a**) Modules included in the CSAM. (**b**) Structural details of the channel attention module. (**c**) Structural details of the spatial attention module.

The channel attention module is a channel-based attention module in the convolutional neural networks that aims to capture the long-term contextual information of channel directions through channel attention maps. To efficiently compute channel attention, we aggregate the spatial information of the feature maps using two pooling operations (average pooling and max pooling) to generate two 2D feature maps $F_1 \in R^c$ and $F_2 \in R^c$. These two feature maps represent the average pooled features and max pooled features in the channel, respectively. The channel attention module can be calculated using the following formula:

$$M_c(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \quad (1)$$

where σ denotes the sigmoid function.

The spatial attention module is a module that is applied to spatial attention in the convolutional neural network. It uses the spatial relationship of features to generate a spatial attention map and concentrates on mining target location information. To compute the spatial attention, we aggregate the channel information of feature maps through two pooling operations (average pooling and max pooling) to generate two 2D feature maps $F_1 \in R^{H\times W}$ and $F_2 \in R^{H\times W}$. These two feature maps represent the average pooled feature and maximum pooled feature in the channel, respectively, and they are concatenated and convolved by a standard convolutional layer to generate a two-dimensional spatial attention map $Attn_s \in R^{H\times W}$. This attention map shows how much the model pays attention to the position. The spatial attention module can be calculated using the following formula:

$$M_s(F) = \sigma(f([AvgPool(F)]; MaxPool(F))) \quad (2)$$

where σ denotes the sigmoid function and f represents a convolution operation.

2.4. Scale Attention Module

The object scale difference is related to the features of different levels. Improving the representation learning ability of different feature levels is conducive to improving the detection accuracy of target detection. However, the features at different levels are usually extracted from different depths of the network, which results in an obvious semantic gap, and it is not optimal to directly fuse feature layers at different levels. To solve this problem, we introduce a SAM to dynamically fuse the features of different scales based on semantics. The structure diagram of SAM is shown in Figure 3.

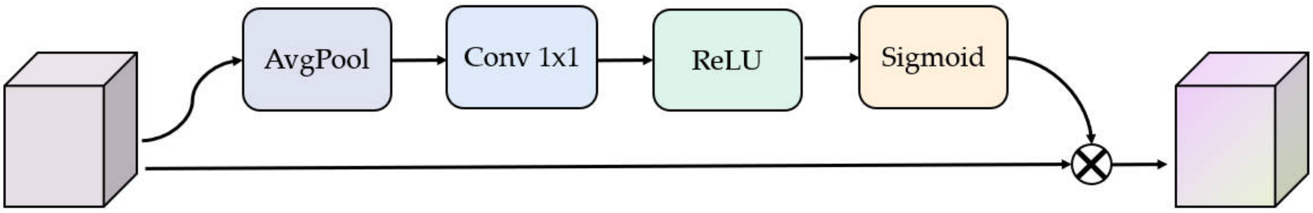

Figure 3. Structural details of the scale attention module, SAM.

We first sample features at different scales to $\frac{H}{4} \times \frac{W}{4}$ and connect them together to form $F_{level} \in R^{L\times C\times \frac{H}{4}\times \frac{W}{4}}$, where L is 4 and C is 256. Next, we use the scale attention module to obtain the scale attention feature map. Then, the scale attention module is composed of average pooling, 1 × 1 convolution, and relu. Finally, we use a sigmoid normalization to obtain the final scale attention feature map $Attn_{level} \in R^{L\times 1\times 1\times 1}$. The scale attention calculation formula is as follows:

$$M(F) = \sigma(f(AvgPool(\sum F))) \quad (3)$$

where σ denotes the sigmoid function and f represents a 1×1 convolution operation.

3. Experimental Results and Discussion

We performed related experiments to evaluate the proposed MANet architecture's effectiveness. In this section, we introduce the selected study area; the used experimental data, experimental setting, and evaluation metrics; the comparative experiment used to verify the performance of MANet; the ablation experiment to compare the effects of multiple attention modules; and, finally, the detection results of MANet in the actual sewage treatment plant scene.

3.1. Study Area and Experimental Data

We chose Beijing as the study area. Beijing is located in the northern part of the North China Plain, which is adjacent to Tianjin. It is located at 115.7°–117.4° east longitude and 39.4°–41.6° north latitude with a total area of 16,410.54 square kilometers. The climate is a typical northern temperate semi-humid continental monsoon climate. As a world-class city, Beijing has a dense population and a large number of factories engaged in production; thus, it requires sewage treatment plants with better operation layouts. Realizing the extraction of sewage treatment plants in this area, in addition to obtaining the number and location information of sewage treatment plants, will provide a reference for the subsequent development and construction planning of emerging cities. Figure 4 is a regional image of Beijing.

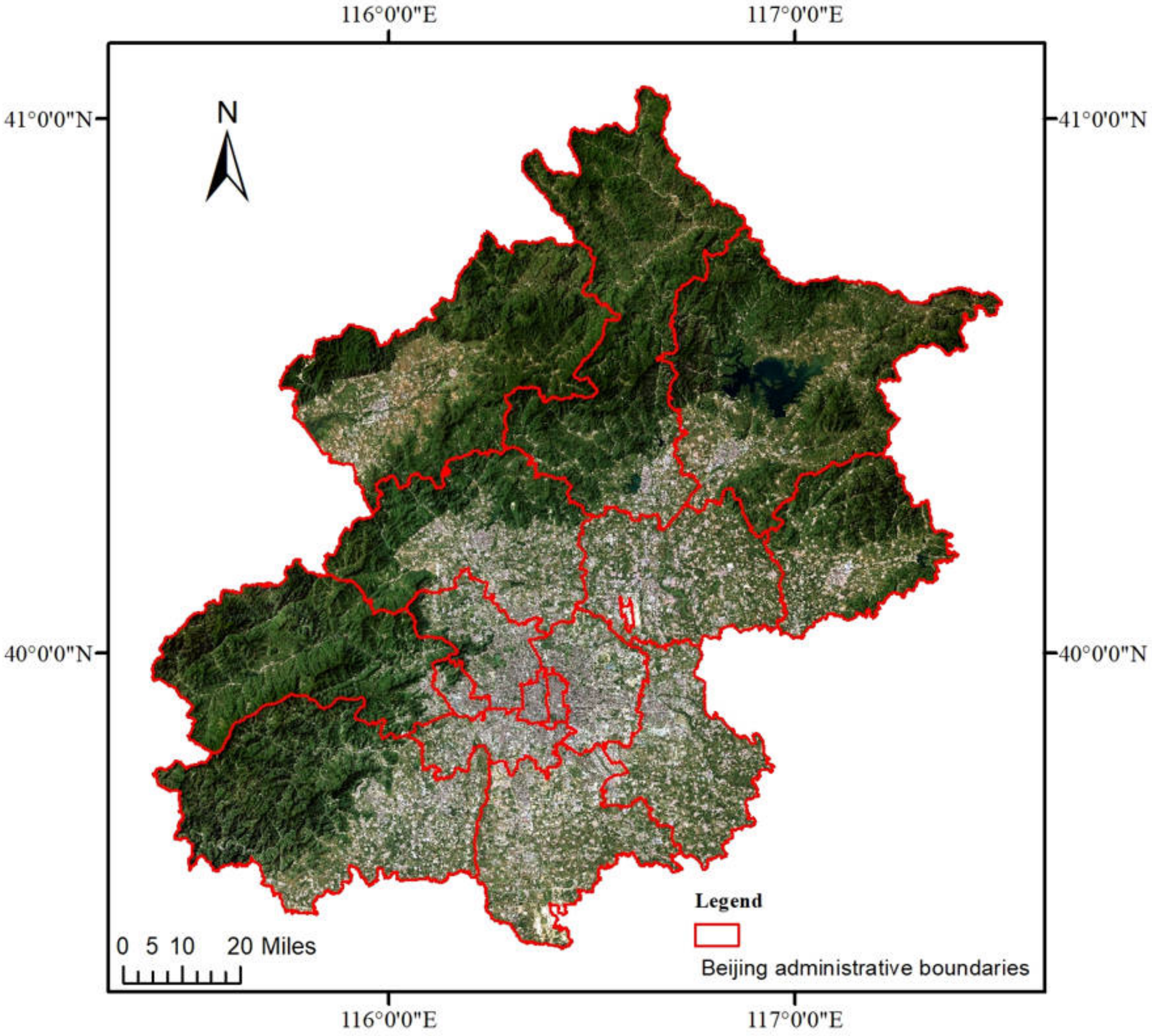

Figure 4. Beijing area.

In order to ensure the clarity of the sewage treatment plant in the remote sensing image, we use the 2 m resolution GF-2 satellite image data domestically produced in China to create a sample dataset of the sewage treatment plant. As shown in Figure 5, we use a sewage treatment plant containing circular and rectangular sedimentation tank structures as detection targets for the sample labeling. Considering the size of the sewage treatment plant target in the 2 m remote sensing image, we use a resolution of 1536 × 1536 for slice production. After manual labeling, 3000 samples of sewage treatment plants were obtained, and the dataset was divided into training and validation sets according to the ratio of 10:1.

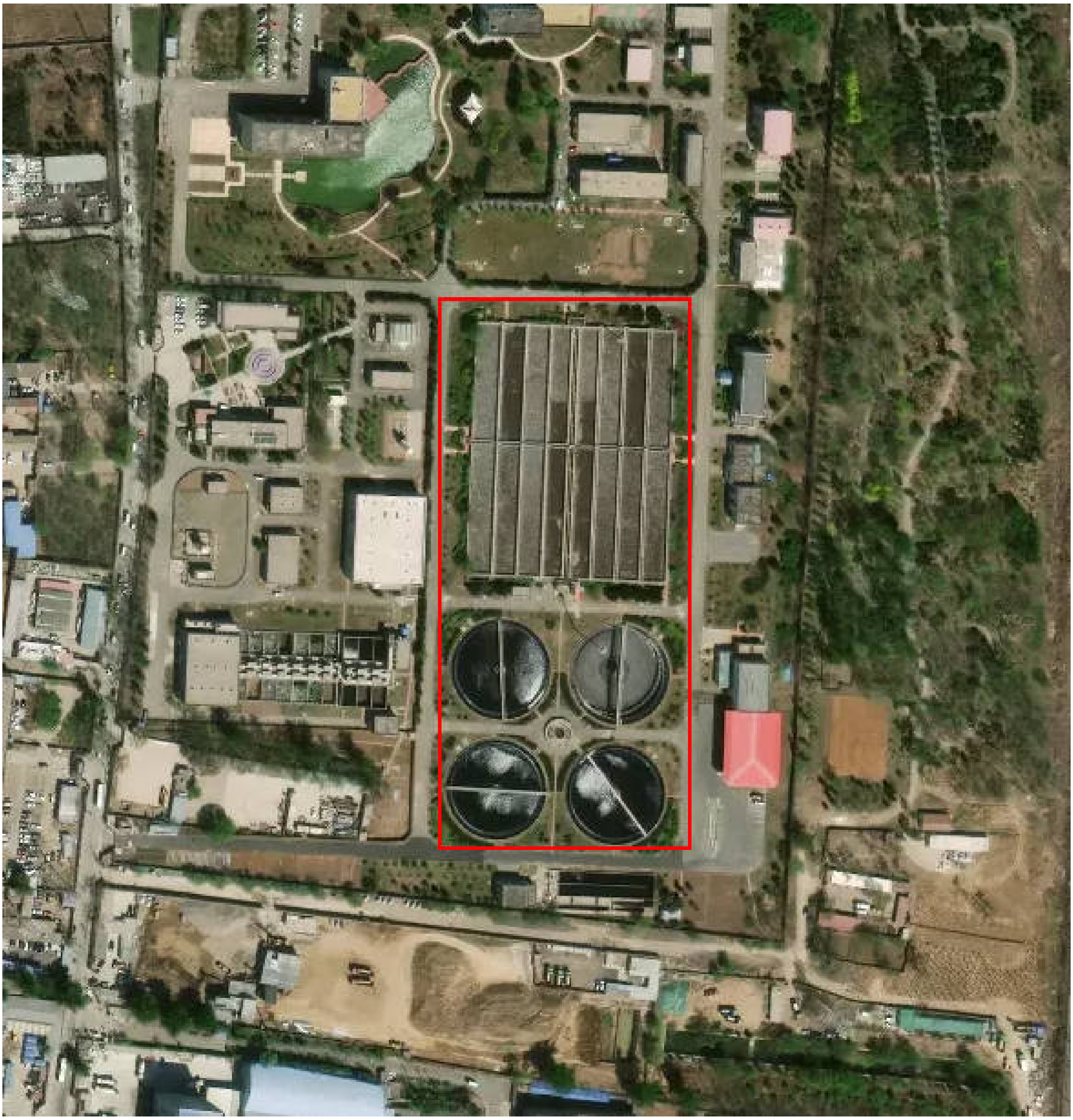

Figure 5. Sewage treatment plant in remote sensing images.

3.2. Experiment Setting and Evaluation Metrics

The method proposed in this paper and the related experiments were all run on the Ubuntu 16.04.7 LTS operating system using an NVIDIA GeForce RTX 3090 GPU with a 24GB memory size. All of the algorithm model experiments were carried out on the PyTorch deep learning framework, and the relevant parameters of the model were kept consistent during the training and testing processes. The input size and batch size were set to 1024 × 1024 × 3 and 8, respectively. The total number of iterations for all experiments was 12 epochs, and all backbones were pre-trained on the ImageNet-1K dataset. The experiment used the stochastic gradient descent (SGD) optimizer, and the original learning rate parameter of network training was set to 0.01. The momentum parameter used to accelerate and stabilize the optimal solution of the function was set to 0.9. The weight decay parameter, which is conducive to the network convergence and fitting data, was set to 0.0001.

Regarding the evaluation metrics of the experiments, we adopt a confusion matrix, which is often used to evaluate object detection results. The confusion matrix's composition is shown in Table 1. Among them, TP indicates that the target is a sewage treatment plant and is correctly predicted; TN indicates that the target is not a sewage treatment plant and is correctly predicted; FP indicates that the target is not a sewage treatment plant but is predicted to be one; and FN indicates that the target is a sewage treatment plant but is predicted to not be one.

Table 1. Confusion matrix.

		Ground Truth	
		True	**False**
Predicted Label	True	TP (True Positive)	FP (False Positive)
	False	FN (False Negative)	TN (True Negative)

Based on the confusion matrix, we further use precision AP (average precision) and recall AR (average recall) to evaluate the detection results. The relevant calculation formulae are as follows:

$$\mathrm{AP} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{4}$$

$$\mathrm{AR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{5}$$

where AP indicates the proportion of correctly predicted sewage treatment plants in the detection results and AR indicates the proportion of correctly predicted sewage treatment plants in the validation set. When its threshold is set to 0.5, AP50 indicates the proportion of correctly predicted sewage treatment plants in the detection results.

3.3. Experimental Results

MANet uses RetinaNet as the baseline network and integrates CSAM and SAM modules based on the ResNet-50+FPN structure. We selected some advanced and representative object detection algorithms to conduct comparative experiments using MANet and used AP and AP50 to evaluate the experimental results. Then, the effectiveness of CSAM and SAM was verified through ablation experiments, and the effects of different modules on MANet performance were analyzed.

3.3.1. Comparison of Model Performance

To evaluate the performance of MANet, we selected four of the most advanced and mature object detection methods (RetinaNet, Cascade RCNN [36], ATSS [37], and Faster RCNN) to conduct the experiments and compare the results in the same environment and settings. The experimental settings are in Section 3.2. The *AP* and *AP*50 curves obtained from the experimental results of the five networks are shown in Figure 6.

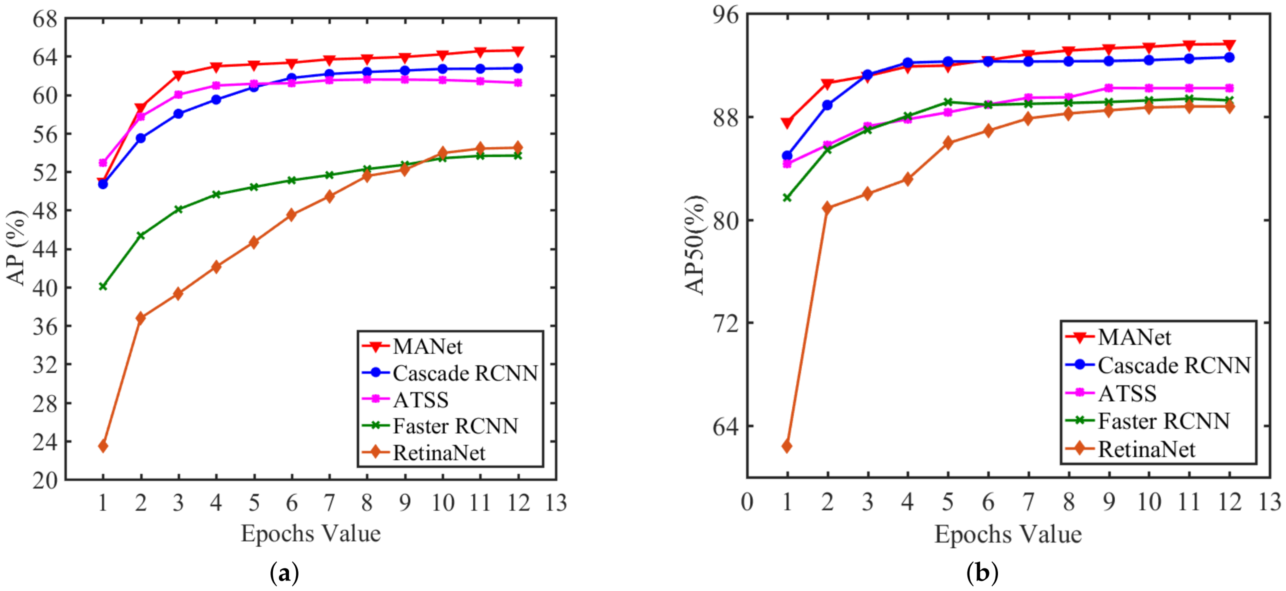

Figure 6. Comparative experimental results of MANet and related networks; (**a**) AP results; (**b**) AP50 results.

As shown in Figure 6, the proposed MANet method achieved higher precision than the other four methods. MANet can obtain the highest AP value of 64.6%, which is significantly higher than the baseline network RetinaNet's value of 54.48%. The AP values of Cascade RCNN and ATSS are closer to MANet but are still below the AP curve of MANet. When the threshold is 0.5, the AP50 value of MANet is still the highest at up to 93.62%, and the AP50 value is basically ahead of the other four networks during the training process. It can be seen that MANet has obvious advantages in the task of detecting sewage treatment plant targets, can more effectively learn the remote sensing image features of sewage treatment plants, and can achieve higher recognition capabilities.

3.3.2. Ablation Studies

In order to verify the effectiveness of CSAM and SAM for MANet to identify sewage treatment plant targets, we conducted ablation experiments on CSAM and SAM. The experiment used RetinaNet as the baseline network, which is based on the ResNet-50+FPN structure, and used the control variable method to experiment with CSAM or SAM. CSAM was disassembled into a channel attention module (Channel-AM) and a spatial attention module (Spatial-AM) for the experiments. The experimental results were evaluated using the AP and AP50 values, and the training time was also involved in the comparison.

Table 2 shows the results of the ablation experiments. Analyzing the experimental results reveals that both of the proposed CSAM and SAM models can improve the network's performance and that the AP values are increased by 4.02% and 6.01%, respectively, compared with the baseline network. When the two models work together, the AP value increases by 10.12%, the AP value reaches 64.6%, and the AP50 can reach 93.62%. These results show that both the CSAM and SAM modules can efficiently extract the sewage treatment plant's features and have strong robustness regarding the shape and scale changes of the sewage treatment plant. In particular, SAM's processing of differently scaled features at different levels contributes more to the performance of the model. Compared with the baseline, the channel attention module and spatial attention module of CSAM have improved by 1.8% and 2.22%, respectively, indicating that CSAM has improved the ability to learn the features of sewage treatment plants in the channel and space dimensions.

Table 2. Ablation experiment results of CSAM and SAM.

Baseline-RetinaNet	Channel-AM	Spatial-AM	Scale-AM	AP (%)	AP50 (%)	Time/h
✓				54.48	88.78	7.97
✓	✓			56.28	89.65	8.14
✓		✓		56.70	89.85	8.18
✓			✓	60.49	91.68	8.54
✓	✓	✓		58.50	90.72	8.35
✓	✓	✓	✓	**64.60**	**93.62**	**8.92**

3.4. Extraction Results of Beijing Sewage Treatment Plant

We used the model obtained by training MANet to detect the sewage treatment plant on the 2 m GF-2 remote sensing image of Beijing, compared the detection results with the actual number of the manual statistics, and used the confusion matrix to evaluate the detection results. The model is set at thresholds of 0.5, 0.6, 0.7, 0.8, and 0.9 when detecting sewage treatment plants, and the results are shown in Table 3.

Table 3. Assessment of detection results of sewage treatment plant in the Beijing area.

Threshold	Actual Amount	Predicted Amount	TP	FP	FN	AP (%)	AR (%)
0.5	151	210	**149**	61	**2**	70.9	**97.7**
0.6		203	146	57	5	71.6	95.4
0.7		180	140	40	11	77.3	92.4
0.8		170	137	33	14	80.1	90.4
0.9		160	130	**30**	21	**81.5**	85.4

The detection results in Table 3 highlight that the AP of the detection results increases with the increase in the threshold, while the AR decreases. When the threshold is 0.5, the number of detections of sewage treatment plants is the largest, reaching 149; however, the false positive detections are serious. Additionally, the AP is the lowest at 70.9%, and the AR can reach 97.7%. When the threshold is 0.9, the predicted number of the sewage treatment plant is the lowest, but the AP is the highest at a value of 81.5%; the AR is 85.4%. Overall, the network has a good detection effect.

According to the sewage treatment plant detection results, we obtained the location information of the sewage treatment plants and created a distribution map of them for the Beijing area in a TIF image. As shown in Figure 7, the red dots represent the detected targets of the sewage treatment plants. From the picture, we can learn more about the sewage treatment plant distribution in Beijing, which can provide a reference for the site selection of sewage treatment plants and help future urban construction.

To further demonstrate the detection effect, we selected some sewage treatment plant targets from the detection results to assist in the description, as shown in Figure 8. It can be seen that MANet can more effectively overcome the problems of the varying shapes and scales of sewage treatment plants and inconsistent local features, has good generalization performance, and can accurately realize the detection of sewage treatment plants.

3.5. Discussion

In this study, we designed a multi-attention network MANet containing multiple modules for the characteristics of large scale changes, large shape differences, and complex semantic information of sewage treatment plants as derived from remote sensing images. The CSAM of the network improves its ability to extract target features from the spatial and channel dimensions, and the SAM processes feature maps of different scales from the scale dimension to reduce the impact of target scale changes. Based on this, MANet was constructed and a deep learning model was trained. The performance of the model was verified in experiments. Finally, the extraction of sewage treatment plants in Beijing was realized based on 2 m GF-2 satellite remote sensing images. The accurate and fast extraction of sewage treatment plants in a large area fully demonstrates the advantages of deep learning methods in the interpretation of remote sensing image objects.

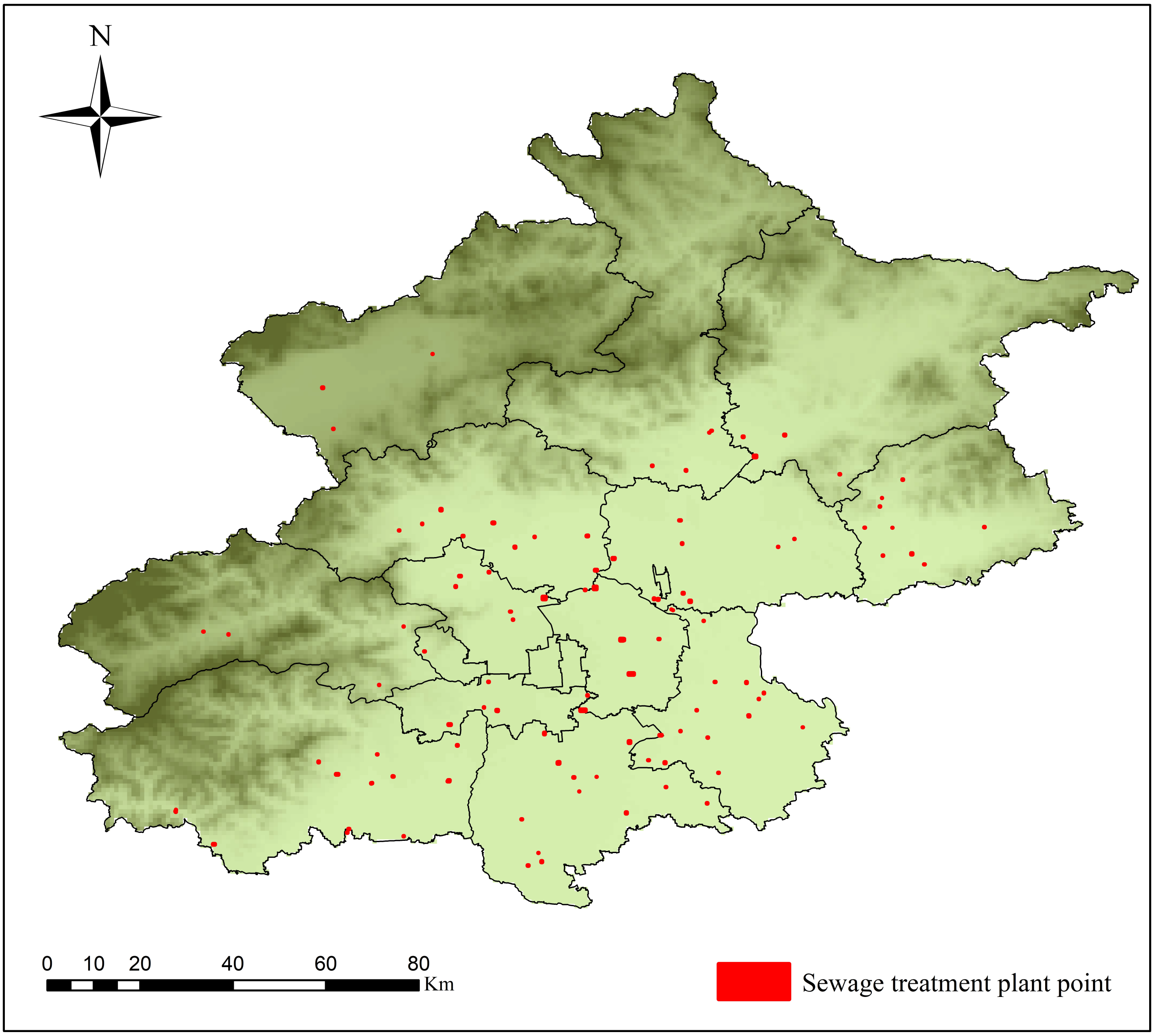

Figure 7. Distribution map of the detection results of Beijing sewage treatment plants.

During the research, we found that the number of samples restricted the accuracy of the model to a certain extent; however, the most fundamental problem was that there were not many sewage treatment plants in reality. Therefore, under the premise of ensuring the generalization performance of the model, we have performed data enhancement on the samples to increase the number of samples. Based on the characteristics of the circular and rectangular sedimentation tank structures contained in the sewage treatment plant, we first focused on the method of sample labeling. On the basis of the original labeling, mark the modules inside the sewage treatment plants and establish a mathematical model for the spatial distance between the modules to judge the target of the sewage treatment plants. In addition, based on the extraction results of the sewage treatment plants, how to fully mine its information according to the key structure of the object, such as further estimating the sewage treatment capacity of the sewage treatment plant based on the target recognition network, may require the designing of a more powerful network.

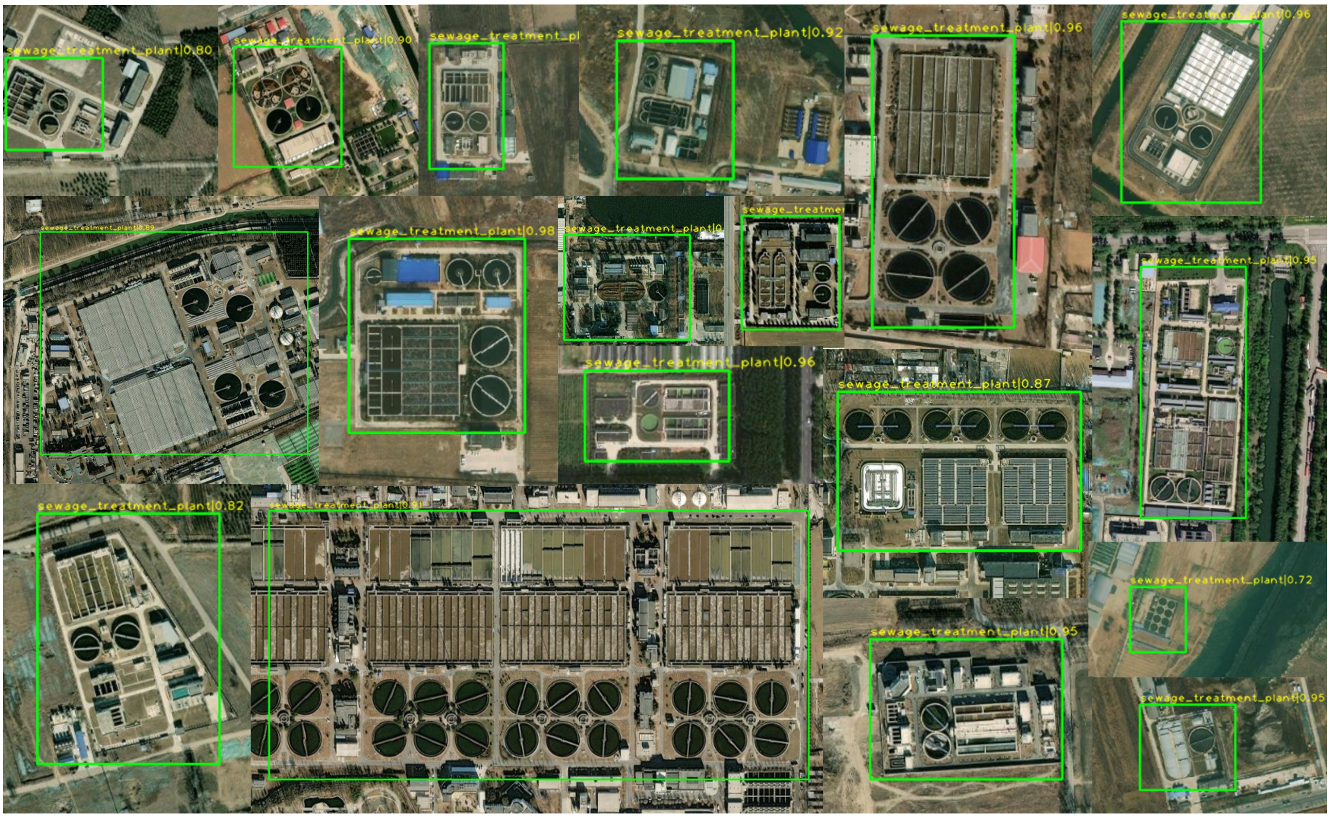
sewage_treatment_plant|0.80
sewage_treatment_plant|0.90
sewage_treatment_pl
sewage_treatment_plant|0.92
sewage_treatment_plant|0.96
sewage_treatment_plant|0.96
sewage_treatment_plant|0.98
sewage_treatment_plant|0
sewage_treatmer
sewage_treatment_plant|0.95
sewage_treatment_plant|0.96
sewage_treatment_plant|0.87
sewage_treatment_plant|0.82
sewage_treatment_plant|0.72
sewage_treatment_plant|0.95
sewage_treatment_plant|0.95

Figure 8. Detection results of sewage treatment plants.

4. Conclusions

Sewage treatment plants in remote sensing images have the characteristics of varying shapes and scales and inconsistent local features. It is difficult to detect sewage treatment plants using traditional deep learning object detection algorithms. In this study, we proposed a novel and effective sewage treatment plant detection network, MANet, which has obvious advantages compared with other advanced object detection algorithms. Based on the model trained by MANet, we realized the detection of the sewage treatment plant on the 2 m resolution GF-2 satellite remote sensing image in the Beijing area and obtained a distribution location map. The conclusions are drawn as follows:

(1) We introduced a lightweight CSAM using channel attention and spatial attention, which can efficiently improve the feature learning ability of MANet in spatial and channel dimensions;

(2) A novel SAM was proposed, which can improve the feature learning ability of MANet at different levels when extracting sewage treatment plant targets with large-scale changes;

(3) Based on the addition of CSAM and SAM to the RetinaNet model, a sewage treatment plant detection network called MANet was proposed, and better results were achieved in the dataset experiments. In the actual scene, based on GF-2 remote sensing images, sewage treatment plant detection for the Beijing area was realized. The results show that our method can achieve an accuracy of 80.1% while maintaining the recall rate at a high level (90.4%).

Author Contributions: Conceptualization, Y.S.; Methodology, Y.S. and K.L.; Data Curation, J.X.; Investigation, Y.S.; Resources, J.X.; Visualization, Y.S.; Writing—Original Draft, Y.S. and K.L.; Review-

ing and Editing, Y.S., Z.C. and J.X. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the National Key Research and Development Program of China (Grant No. 2021YFB3901202).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: Thanks to Haoyang Wang for his support in making the distribution map.

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

References

1. Obaideen, K.; Shehata, N.; Sayed, E.T.; Abdelkareem, M.A.; Mahmoud, M.S.; Olabi, A. The role of wastewater treatment in achieving sustainable development goals (SDGs) and sustainability guideline. *Energy Nexus* **2022**, *7*, 100112. [CrossRef]
2. Shanmugam, K.; Gadhamshetty, V.; Tysklind, M.; Bhattacharyya, D.; Upadhyayula, V.K. A sustainable performance assessment framework for circular management of municipal wastewater treatment plants. *J. Clean. Prod.* **2022**, *339*, 130657. [CrossRef]
3. Gautam, S.K.; Sharma, D.; Tripathi, J.K.; Ahirwar, S.; Singh, S.K. A study of the effectiveness of sewage treatment plants in Delhi region. *Appl. Water Sci.* **2013**, *3*, 57–65. [CrossRef]
4. Jin, L.; Zhang, G.; Tian, H. Current state of sewage treatment in China. *Water Res.* **2014**, *66*, 85–98. [CrossRef] [PubMed]
5. Bao, R.; Wang, Z.; Qi, H.; Mehmood, T.; Cai, M.; Zhang, Y.; Yang, R.; Peng, L.; Liu, F. Occurrence and distribution of microplastics in wastewater treatment plant in a tropical region of China. *J. Clean. Prod.* **2022**, *349*, 131454. [CrossRef]
6. Xie, J.; Jin, L.; Wu, D.; Pruden, A.; Li, X. Inhalable antibiotic resistome from wastewater treatment plants to urban areas: Bacterial hosts, dissemination risks, and source contributions. *Environ. Sci. Technol.* **2022**, *56*, 7040–7051. [CrossRef]
7. Liu, B.; Tang, J.; Qu, Y.; Yang, Y.; Lyu, H.; Dai, Y.; Li, Z. A GIS-based method for identification of blindness in former site selection of sewage treatment plants and exploration of optimal siting areas: A case study in Liao River basin. *Water* **2022**, *14*, 1092. [CrossRef]
8. Hubel, D.H.; Wiesel, T.N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.* **1962**, *160*, 106. [CrossRef]
9. Raina, R.; Madhavan, A.; Ng, A.Y. Large-scale deep unsupervised learning using graphics processors. In Proceedings of the 26th Annual International Conference on Machine Learning, Montreal, QC, Canada, 14–18 June 2009; pp. 873–880.
10. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1137–1149. [CrossRef]
11. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Part I 14; Springer: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
12. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. *arXiv* **2017**, arXiv:1708.02002.
13. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. *arXiv* **2015**, arXiv:1506.02640.
14. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. *arXiv* **2016**, arXiv:1612.08242.
15. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
16. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
17. Rogan, J.; Chen, D. Remote sensing technology for mapping and monitoring land-cover and land-use change. *Prog. Plan.* **2004**, *61*, 301–325. [CrossRef]
18. Zhao, S.; Wang, Q.; Li, Y.; Liu, S.; Wang, Z.; Zhu, L.; Wang, Z. An overview of satellite remote sensing technology used in China's environmental protection. *Earth Sci. Inform.* **2017**, *10*, 137–148. [CrossRef]
19. Toth, C.; Jóźków, G. Remote sensing platforms and sensors: A survey. *ISPRS J. Photogramm. Remote. Sens.* **2016**, *115*, 22–36. [CrossRef]
20. Ren, Y.; Zhu, C.; Xiao, S. Small object detection in optical remote sensing images via modified faster R-CNN. *Appl. Sci.* **2018**, *8*, 813. [CrossRef]
21. Chen, J.; Wan, L.; Zhu, J.; Xu, G.; Deng, M. Multi-scale spatial and channel-wise attention for improving object detection in remote sensing imagery. *IEEE Geosci. Remote. Sens. Lett.* **2019**, *17*, 681–685. [CrossRef]
22. Liu, W.; Ma, L.; Chen, H. Arbitrary-oriented ship detection framework in optical remote-sensing images. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 937–941. [CrossRef]
23. Deng, Z.; Sun, H.; Zhou, S.; Zhao, J.; Lei, L.; Zou, H. Multi-scale object detection in remote sensing imagery with convolutional neural networks. *ISPRS J. Photogramm. Remote. Sens.* **2018**, *145*, 3–22. [CrossRef]
24. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. *arXiv* **2016**, arXiv:1612.03144.

25. Yang, X.; Sun, H.; Fu, K.; Yang, J.; Sun, X.; Yan, M.; Guo, Z.J. Automatic ship detection in remote sensing images from google earth of complex scenes based on multiscale rotation dense feature pyramid networks. *Remote Sens.* **2018**, *10*, 132. [CrossRef]
26. Li, Y.; Huang, Q.; Pei, X.; Jiao, L.; Shang, R.J. RADet: Refine feature pyramid network and multi-layer attention network for arbitrary-oriented object detection of remote sensing images. *Remote Sens.* **2020**, *12*, 389. [CrossRef]
27. Yan, J.; Wang, H.; Yan, M.; Diao, W.; Sun, X.; Li, H.J. IoU-adaptive deformable R-CNN: Make full use of IoU for multi-class object detection in remote sensing imagery. *Remote Sens.* **2019**, *11*, 286. [CrossRef]
28. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762.
29. Zhu, X.; Lyu, S.; Wang, X.; Zhao, Q. TPH-YOLOv5: Improved YOLOv5 Based on Transformer Prediction Head for Object Detection on Drone-captured Scenarios. *arXiv* **2021**, arXiv:2108.11539.
30. Long, H.; Chung, Y.; Liu, Z.; Bu, S. Object detection in aerial images using feature fusion deep networks. *IEEE Access* **2019**, *7*, 30980–30990. [CrossRef]
31. Yang, X.; Yang, J.; Yan, J.; Zhang, Y.; Zhang, T.; Guo, Z.; Xian, S.; Fu, K. SCRDet: Towards More Robust Detection for Small, Cluttered and Rotated Objects. *arXiv* **2018**, arXiv:1811.07126.
32. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable Convolutional Networks. *arXiv* **2017**, arXiv:1703.06211.
33. Zhu, J.; Fang, L.; Ghamisi, P. Deformable convolutional neural networks for hyperspectral image classification. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 1254–1258. [CrossRef]
34. Guo, H.; Bai, H.; Yuan, Y.; Qin, W. Fully deformable convolutional network for ship detection in remote sensing imagery. *Remote Sens.* **2022**, *14*, 1850. [CrossRef]
35. Li, H.; Zech, J.; Hong, D.; Ghamisi, P.; Schultz, M.; Zipf, A. Leveraging openstreetmap and multimodal remote sensing data with joint deep learning for wastewater treatment plants detection. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *110*, 102804. [CrossRef] [PubMed]
36. Cai, Z.; Vasconcelos, N. Cascade R-CNN: Delving into High Quality Object Detection. *arXiv* **2017**, arXiv:1712.00726.
37. Zhang, S.; Chi, C.; Yao, Y.; Lei, Z.; Li, S.Z. Bridging the Gap Between Anchor-based and Anchor-free Detection via Adaptive Training Sample Selection. *arXiv* **2019**, arXiv:1912.02424.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.