electronics

Article

HA-Unet: A Modified Unet Based on Hybrid Attention for Urban Water Extraction in SAR Images

Huina Song [1,*], Han Wu [1], Jianhua Huang [1], Hua Zhong [1], Meilin He [1], Mingkun Su [1], Gaohang Yu [2], Mengyuan Wang [1] and Jianwu Zhang [1]

[1] School of Communication Engineering, Hangzhou Dianzi University, Hangzhou 310018, China
[2] School of Sciences, Hangzhou Dianzi University, Hangzhou 310018, China
* Correspondence: huinasong@hdu.edu.cn; Tel.: +86-1538-234-7570

Abstract: Urban water plays a significant role in the urban ecosystem, but urban water extraction is still a challenging task in automatic interpretation of synthetic aperture radar (SAR) images. The influence of radar shadows and strong scatters in urban areas may lead to misclassification in urban water extraction. Nevertheless, the local features captured by convolutional layers in Convolutional Neural Networks (CNNs) are generally redundant and cannot make effective use of global information to guide the prediction of water pixels. To effectively emphasize the identifiable water characteristics and fully exploit the global information of SAR images, a modified Unet based on hybrid attention mechanism is proposed to improve the performance of urban water extraction in this paper. Considering the feature extraction ability and the global modeling capability in SAR image segmentation, the Channel and Spatial Attention Module (CSAM) and the Multi-head Self-Attention Block (MSAB) are both introduced into the proposed Hybrid Attention Unet (HA-Unet). In this work, Resnet50 is adopted as the backbone of HA-Unet to extract multi-level features of SAR images. During the feature extraction process, CSAM based on local attention is adopted to enhance the meaningful water features and ignore unnecessary features adaptively in feature maps of two shallow layers. In the last two layers of the backbone, MSAB is introduced to capture the global information of SAR images to generate global attention. In addition, two global attention maps generated by MSAB are aggregated together to reconstruct the spatial feature relationship of SAR images from high-resolution feature maps. The experimental results on Sentinel-1A SAR images show that the proposed urban water extraction method has a strong ability to extract water bodies in the complex urban areas. The ablation experiment and visualization results vividly indicate that both CSAM and MSAB contribute significantly to extracting urban water accurately and effectively.

Keywords: hybrid attention; Unet; urban water extraction; synthetic aperture radar

Citation: Song, H.; Wu, H.; Huang, J.; Zhong, H.; He, M.; Su, M.; Yu, G.; Wang, M.; Zhang, J. HA-Unet: A Modified Unet Based on Hybrid Attention for Urban Water Extraction in SAR Images. *Electronics* **2022**, *11*, 3787. https://doi.org/10.3390/electronics11223787

Academic Editor: Chiman Kwan

Received: 28 October 2022
Accepted: 16 November 2022
Published: 17 November 2022

1. Introduction

Synthetic aperture radar (SAR) has widespread applications in remote sensing of the environment with the advantage of all-weather and all-time observation. With the dramatic increase in spaceborne SAR images, automatic interpretation of SAR images is a promising technology in the earth observation and surveillance [1–4]. As a vital factor for urban ecosystem, urban water is of great importance to urban ecological landscapes, urban development, and flood control [5–7]. Accurate and effective urban water extraction is a critical task in automatic interpretation of SAR images.

Various methods have been developed for the mapping of water in SAR images, which can be divided into traditional methods and deep learning methods. However, to our knowledge, most research pays attention to reservoirs and large lakes, and few studies focus on the exploration of urban water bodies including urban rivers, ponds and natural or human-made lakes [8,9]. A variety of traditional methods have achieved a great success for suburban water extraction, but the transferability of these methods for

urban water extraction is a critical issue [10–13]. Very recently, deep learning methods, especially Convolutional Neural Networks (CNNs), enabled remarkable performance in water extraction from SAR images by virtue of its powerful feature extraction ability without auxiliary data [14,15].

SAR image segmentation with CNNs is a task that requires the integration of feature maps from different spatial scales and a balance between local information and global information. Representative models including fully convolutional network, Unet, and the high-resolution network have been used to perform water segmentation [16]. Experimental results indicate that it is feasible to extract water bodies with CNNs. Due to the influence of radar shadow or strong scatter, prediction of urban water bodies with irregular shapes requires more contextual information in comparison with suburban water extraction [17]. For example, other dark surfaces (such as asphalt roads) and radar shadows caused by tall buildings in SAR images may be misclassified as water bodies while rough water surfaces may be blanketed. In the work by Denbina et al. [18], CNNs are adopted to detect flooding in urban areas, which suggests the potential of CNN-based urban water extraction in SAR images. Furthermore, a dense-coordinate-feature-concatenate network (DCFNet) is proposed to extract and fuse the water features [19]. The experimental results on Gaofen-3 and Sentinel-1 SAR images show that DCFNet can reduce the influence of ground interference and speckle noise to some extent. However, due to the inherent limitation of the local receptive field in CNNs, it is difficult to extract the global information of SAR images in these methods [20]. To solve these problems, a multiscale module is introduced into the CNN for urban water extraction [6]. In the work by Geng et al. [21], a recurrent layer based on long-short-term memory is introduced to extract spatial dependencies for SAR image segmentation. These modules can provide a description of the context of spatial locations to some extent, but the global information acquisition is still inadequate. Hence, it is meaningful to investigate more advanced models for water extraction to promote the application of automatic SAR image interpretation technology. Furthermore, accurate water mapping can present flooding disasters in real time, shoreline extraction, surface water monitoring and monitoring changes of river and lake in urban areas.

According to the above-mentioned analysis, this study aims to develop a deep learning method to improve the accuracy of urban water extraction in SAR images. This paper proposes Hybrid Attention Unet (HA-Unet) with the adaptive feature enhancement capability and the global modeling capability by introducing the hybrid attention mechanism into the classic Unet architecture. In the proposed HA-Unet, Resnet50 is adopted as the backbone to extract multi-layer features, and a total of 5 layers of feature maps are processed by the hybrid attention mechanism based on the Channel and Spatial Attention Module (CSAM) and the Multi-head Self-Attention Block (MSAB). Given that the channel attention and spatial attention mechanism can refine meaningful features to learn where and what to emphasize, CSAM is adopted to enhance the recognizable water features in the first two layers of the feature maps. Since the multi-head self-attention mechanism has the ability to model global information, MSAB is introduced into the last two layers of the feature maps to empower the proposed HA-Unet with a global receptive field. In addition, global attention maps of the last two layers are combined to further encode long-range dependencies of SAR images from high-resolution feature maps. The above series of operations enable the proposed HA-Unet to self-adaptive focus on target water and pay attention to the most semantic-relevant contextual features. Thus, the proposed HA-Unet can improve the performance of urban water extraction in terms of accuracy and efficiency.

The remainder of this paper is organized as follows. In Section 2, related work in this paper including Unet for water segmentation and attention mechanisms is briefly introduced. In Section 3, the research data are presented. In Section 4, HA-Unet for urban water extraction is proposed in detail. The experimental results based on Sentinel-1A (S1A) SAR images are presented in Section 5. The discussion for experimental results is provided in Section 6. Finally, the whole paper is concluded in Section 7.

2. Related Work

2.1. Unet for Water Segmentation

Unet, a popular model in the field of segmentation, is named for a U-shaped encoder-decoder architecture and transmits the shallow feature maps of the encoder to the deep through skip connections [22]. In recent years, Unet has been quite a popular model for SAR image segmentation. Many researchers have proved that the water segmentation result of Unet is better than traditional methods [23,24]. In 2019, Unet is adopted to perform an efficient segmentation of river [25]. After that, various improvement measures are being introduced into the original Unet to improve the performance of water segmentation in SAR images. The concept of inception is introduced into Unet to increase the receptive field of the convolution operation [26]. In addition, the spatial pyramid pooling module and the attention block are both adopted in Unet to construct a robust water extraction network [27]. The spatial pyramid pooling module fuses more contextual semantic information and the attention block makes the model focus on water extraction. Furthermore, Ren, Y. et al. introduce the position attention module and the channel attention module into the last layer of the encoder in Unet, which shows a 1% accuracy improvement than the original Unet in water segmentation [28]. However, there are still challenges in urban water extraction, such as the coherent speckle noise and complex terrain information in SAR images.

2.2. Attention Mechanisms

In the recent few years, attention mechanisms have played an increasingly important role in computer vision tasks. The attention mechanism can dynamically select representative features according to the importance of the input. It enables CNNs to pay attention to specific parts of the input image and select high-value information from massive information. Furthermore, related studies have shown that the attention mechanism is a means for deep learning models to understand SAR images with complex scenes in terms of accuracy and efficiency [14].

The unified model of human attention that focuses on the representative parts and analyses these parts can be defined as

$$Attention = f(g(x), x), \tag{1}$$

where x represents the image, $g(x)$ represents the process of generating attention according to the representative parts, and $f(x)$ means the process of analyzing the image based on the attention generated by $g(x)$ to obtain high-value information.

The implementation of attention mechanisms in deep learning can be also divided into two steps. At first, the attention distribution is computed based on the input information. Then, the high-value information is output according to the attention mechanism. The general form of attention mechanisms in deep learning is presented in Figure 1. First, the energy score that reflects the importance of input information is obtained through the score function $g(\cdot)$. The score function $g(\cdot)$ is a key role in the process of generating the attention distribution. Different operations such as addition, similarity, multiplication, convolution, etc. can yield different energy scores, which determines the name of the attention mechanism. Then, the energy score is normalized between 0 and 1 by a distribution function to obtain the attention weight. The distribution function usually corresponds to the $softmax$ function, which is defined as

$$Softmax(z_i) = \frac{e^{z_i}}{\sum_{n=1}^{n=N} e^{z_n}}, \tag{2}$$

where z_i denotes the energy score, e represents Natural Constant. Finally, the product of the attention weight and input information is calculated to generate weighted values.

The main idea behind the above-explained attention mechanisms is to give different weights to different features. In other words, the attention mechanism enables the model to exploit the most relevant parts of the input information flexibly. Thus, CNNs with

attention mechanisms for water extraction can give higher weights to relevant water body information to attract the attention of the CNN to water features.

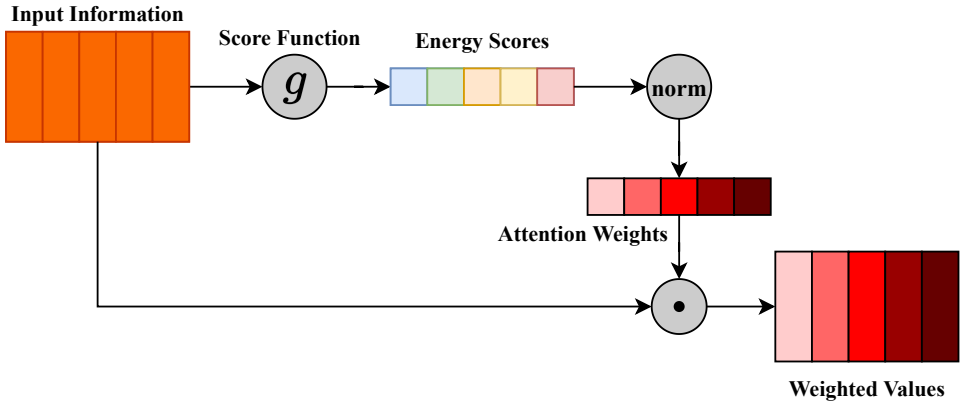

Figure 1. The general form of attention mechanisms.

3. Materials

3.1. Study Area

The study area is a typical watershed-rich region in east Hangzhou, Zhejiang Province, China. Hangzhou, the provincial capital of Zhejiang, is located in the north of Zhejiang Province, east of Hangzhou Bay. The climate in Hangzhou is a subtropical monsoon, with an average annual temperature of 17.8 °C and an annual rainfall of 1454 mm. The rainfall usually concentrates in March and April each year. On 26 March 2022, the temperature in Hangzhou was 11–17 °C, with the moderate breeze. Hangzhou is densely populated with lakes. In particular, West Lake is the symbol of Hangzhou, as well as one of the most beautiful sights in China and the Qiantang River flows through most of the city from southwest to northeast. Therefore, the distribution of water bodies is of great significance to the development of Hangzhou, and the region related to West Lake and parts of the Qiantang River is selected as the study area. The location of the study area is shown in Figure 2. Furthermore, the Sentinel-2A optical remote sensing image of the study area, collected on 24 March 2022, at 2:35 UTC, is presented in Figure 3. In the optical remote sensing image corresponding to the study area, both West Lake and Qiantang River to be extracted are included.

3.2. Dataset

In this work, VH polarization SAR images in Ground Range Detected (GRD) format acquired by S1A are used to generate the dataset. The Sentinel-1 satellite is C-band multi-polarization SAR, which offers a continuity of wide area coverage, achieving higher resolution and global coverage over landmasses. It has been used in water conservancy, disaster reduction, marine and other fields [29]. Sentinel-1 data are made available systematically and can be downloaded from the Alaska Satellite Facility (https://search.asf.alaska.edu/, accessed on 15 September 2022). The basic information of the data source in our experiment is presented in Table 1.

In order to enhance the readability of SAR images, S1A SAR images are pre-processed in this work on the SentiNel Application Platform provided by the Scientific Data Hub of European Space Agency (ESA). First, SAR images are subjected to orbit correction. Then, radiometric calibration is applied to reduce the radiometric irregularities. Next, range Doppler terrain correction is conducted to reduce the influence of local terrain on backscatter. Subsequently, the SAR block-matching 3D (SAR BM3D) algorithm based on non-local approach is introduced to reduce the negative impact of speckle noise [30]. Last but not least, the backscatter coefficient is converted to dB using a logarithmic transformation for the readability improvement.

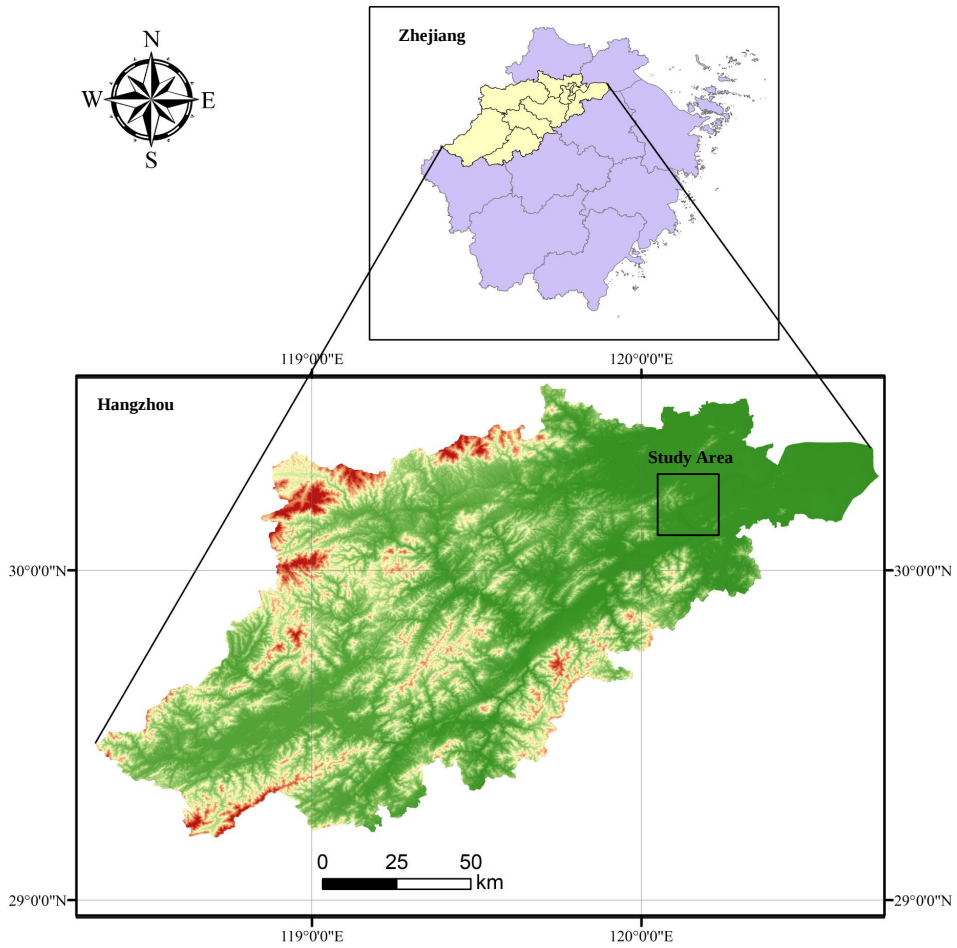

Figure 2. The location of the study area.

Figure 3. The optical remote sensing image of the study area.

Table 1. The basic information of the data source.

Sentinel-1A	Parameter
Product format	GRD
Product level	Level-1
Beam mode	Interferometric Wide swath
Polarization	VH
Resolution	20 × 22 m
Band	C
Number of looks	5 × 1
Size	2048 × 2048 pixels

To obtain comprehensive and accurate ground truths, several SAR images are labeled manually according to OpenStreetMap (https://www.openstreetmap.org/, accessed on 20 April 2022) on the Labelme software. Then, SAR images and the ground truth are randomly cut into a common size of 512 × 512 pixels, and a total of 522 subimages and the corresponding labels are obtained in our experiment. The subimages and the corresponding labels are divided into two parts: 470 training subimages (about 90%) and 52 validating subimages (about 10%).

The tested S1A SAR image, related to the study area in east Hangzhou, was collected on 26 March 2022, at 10:03 UTC. The tested data item is a 2048 × 2048 pixel SAR image and is also cropped into a common size of 512 × 512 pixels for water segmentation. Finally, all subimages are returned to their original locations to obtain the final urban water extraction result.

4. Methodology

4.1. Overview

Considering the complexity of water bodies in urban areas, a modified urban water extraction method is proposed based on HA-Unet to exploit the global information as well as local water features. The chain of the proposed urban water extraction method is presented in Figure 4. After data acquisition and SAR image pre-processing, the S1A SAR images and the ground truth are fed into the proposed HA-Unet for training. Finally, the tested SAR image related to the study area is segmented by the trained HA-Unet for the map of urban water.

4.2. Overall Structure of the Proposed HA-Unet

Unet and its variants have been widely used in SAR image segmentation in recent years. However, the redundancy of features captured by convolution layers in Unet may lead to the misclassification of water pixels. Furthermore, the local receptive field of small convolution kernels in Unet cannot leverage global interaction well [31]. To overcome the above problems, the hybrid attention is introduced into the original Unet based on local attention mechanism and global attention mechanism, and the proposed network is known as HA-Unet.

As presented in Figure 5, HA-Unet for urban water extraction retains the essential structure of Unet with Resnet50 [32] as backbone. In the encoder, the five stages of feature maps extracted by Resnet50 are defined as $feat_i$ ($i = 1, 2, 3, 4, 5$). Furthermore, the feature maps of the decoder are defined as up_i ($i = 1, 2, 3, 4$). Based on the local attention, CSAM is adopted at early stages to enhance the meaningful water features and filter out non-semantic features in $feat_1$ and $feat_2$ before skipping connections. Based on the global attention, MSAB is introduced to model long-range dependencies for CNN backbone at late stages, and $feat_4$ and $feat_5$ activated by MSAB are aggregated to construct high-resolution feature maps. Finally, the pixel-level water segmentation results are generated in the decoder.

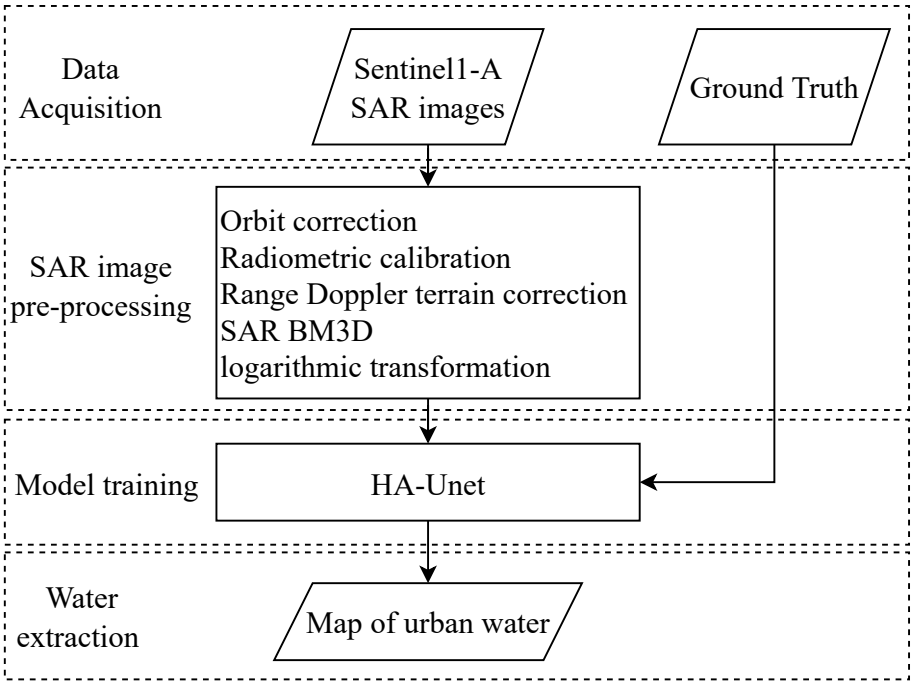

Figure 4. The chain of the proposed urban water extraction method.

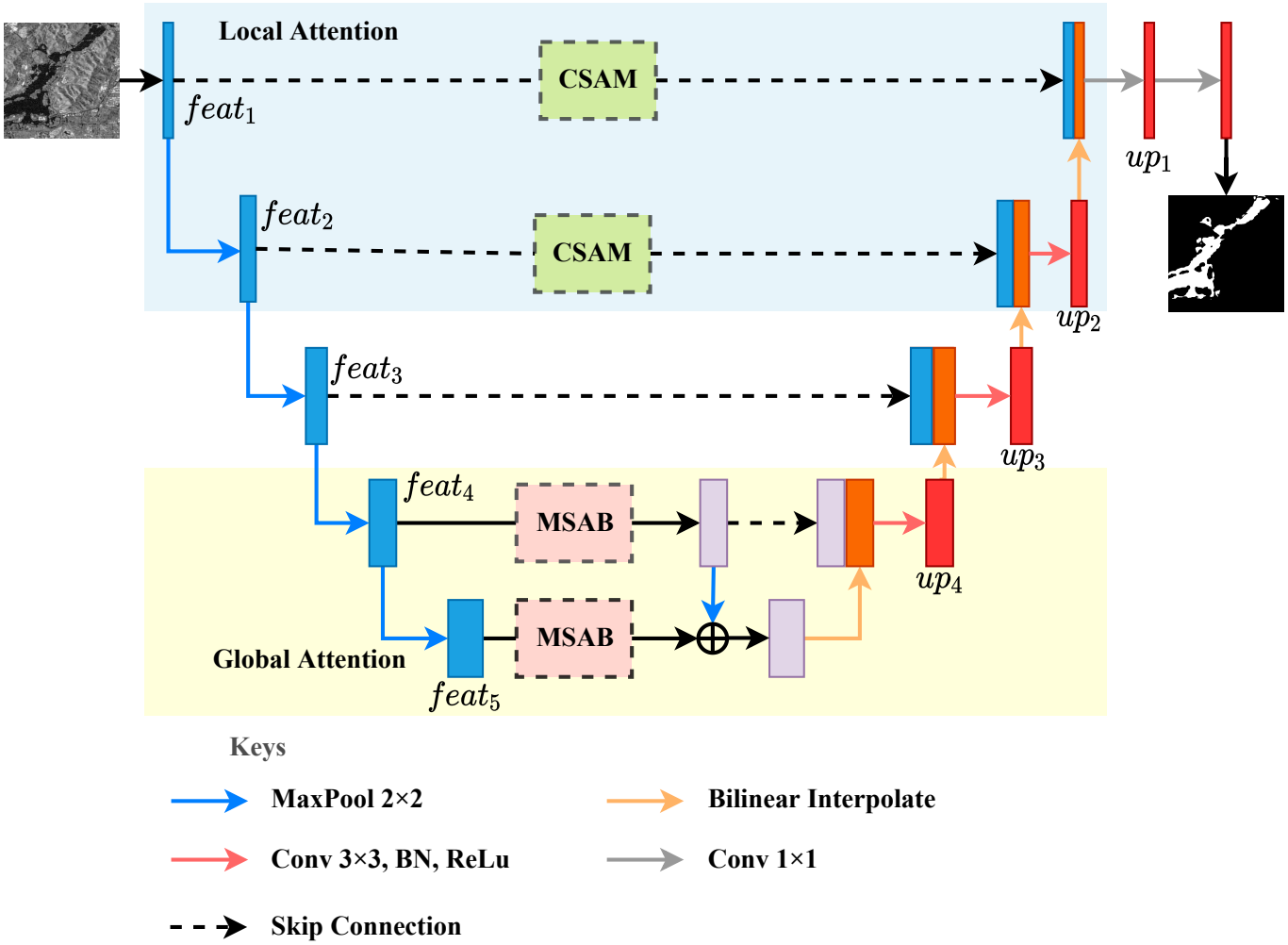

Figure 5. Overall structure of the proposed HA-Unet. ⊕ denotes element-wise addition.

4.2.1. Backbone

Resnet is widely used in semantic segmentation and target detection, and shows outstanding performance in remote sensing images [33]. The key idea of Resnet is the deep residual learning framework, and the deep residual learning framework in Resnet is presented in Figure 6. Instead of directly fitting an underlying mapping $F(x)$, the stacked layers in Resnet fit a residual mapping $F(x) + x$ which is performed by a shortcut connection and element-wise addition. This framework with a shortcut connection helps Resnet extend the depth of the network to learn richer features without gradient degradation [32]. Taking into account the number of parameters and training difficulty, Resnet50 is adopted for water feature extraction in this work and the structure of Resnet50 backbone is presented in Table 2.

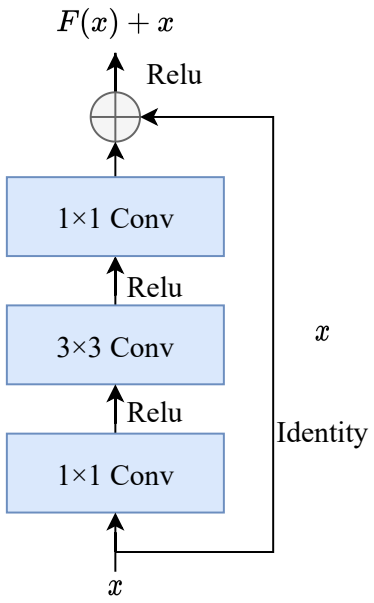

Figure 6. The deep residual learning framework.

Table 2. The structure of Resnet50 used in this paper.

Layer Name	Operator	Output Name	Output Size	Output Dimension
conv1	7 × 7 Conv, stride = 2, padding = 3	$feat_1$	256 × 256	64
conv2x	3 × 3 Pool, stride = 2 $\begin{bmatrix} 1\times1\ \text{Conv} \\ 3\times3\ \text{Conv} \\ 1\times1\ \text{Conv} \end{bmatrix} \times 3$	$feat_2$	128 × 128	64
conv3x	$\begin{bmatrix} 1\times1\ \text{Conv} \\ 3\times3\ \text{Conv} \\ 1\times1\ \text{Conv} \end{bmatrix} \times 3$	$feat_3$	64 × 64	512
conv4x	$\begin{bmatrix} 1\times1\ \text{Conv} \\ 3\times3\ \text{Conv} \\ 1\times1\ \text{Conv} \end{bmatrix} \times 3$	$feat_4$	32 × 32	1024
conv5x	$\begin{bmatrix} 1\times1\ \text{Conv} \\ 3\times3\ \text{Conv} \\ 1\times1\ \text{Conv} \end{bmatrix} \times 3$	$feat_5$	16 × 16	2048

4.2.2. CSAM

Identifiable and representative feature representations are essential in high accuracy segmentation. However, these features extracted by CNNs are redundant, especially at low stage, and they may influence the implicit representation of CNNs [34]. Thus, a feature selecting approach is necessary for urban water extraction. In the proposed HA-Unet, CSAM based on channel attention and spatial attention is adopted at the early stage of the encoder for adaptive feature enhancement in complex scenes of SAR images. The schematic of CSAM is presented in Figure 7.

Given a feature map x as input, the overall progress of CSAM can be described as follows:

$$y = M_s(M_c(x) \otimes x) \otimes (M_c(x) \otimes x), \tag{3}$$

$$M_c = sigmoid\left[mlp\left(x_{\text{avgpool}}\right) + mlp\left(x_{\text{maxpool}}\right)\right], \tag{4}$$

$$M_s = sigmoid\left[conv\left(concat\left(x_{\text{avgpool}}, x_{\text{maxpool}}\right)\right)\right], \tag{5}$$

where y represents the emphasized feature map, M_c and M_s denote the activation maps after a *sigmoid* function generated by channel attention and spatial attention, respectively, $\otimes$ represents element-wise multiplication, mlp represents multi-layer perceptron, and $conv$ represents a convolution layer. After CSAM, salient parts of significant properties of water in the feature maps $feat_1$ and $feat_2$ are focused on water bodies by adaptive enhancement in both channel dimension and spatial dimension, while the unnecessary ones are suppressed.

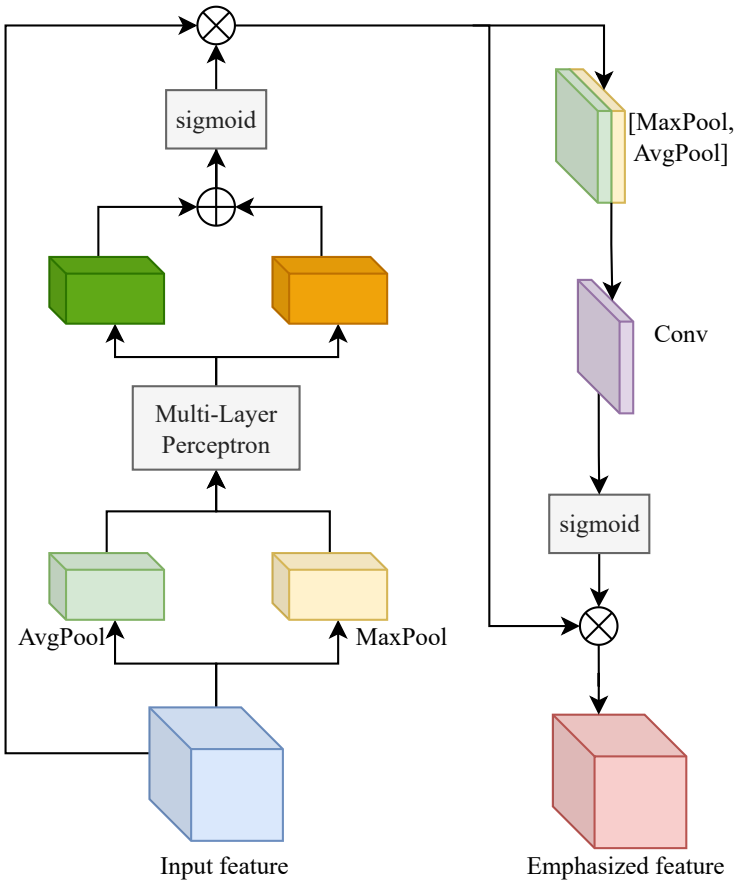

Figure 7. The schematic of CSAM. $\oplus$ and $\otimes$ denote element-wise addition and element-wise multiplication, respectively.

4.2.3. MSAB

Segmentation is a task that requires accurate pixel-level predictions. Not only fine-grained features, but also long-range dependencies are crucial to resolving the ambiguities of local pixel prediction [35]. In large-scene SAR images, intrinsic correlations among pixels are beneficial to improve classification accuracy, especially for small regional segmentation [21]. Nevertheless, CNNs have difficulty in capturing the latent contextual correlations of the whole image, since they only process a local neighborhood because of their local receptive field. Based on self-attention, MSAB shown in Figure 8 is introduced into the late stages of the encoder to model the long-range dependencies of SAR images. In MSAB, the multi-head self-attention (MHSA) layer captures the multiple complex relationships by a concatenation of outputs of n self-attention heads and 1×1 convolutions are used to transform the dimensions of output feature maps.

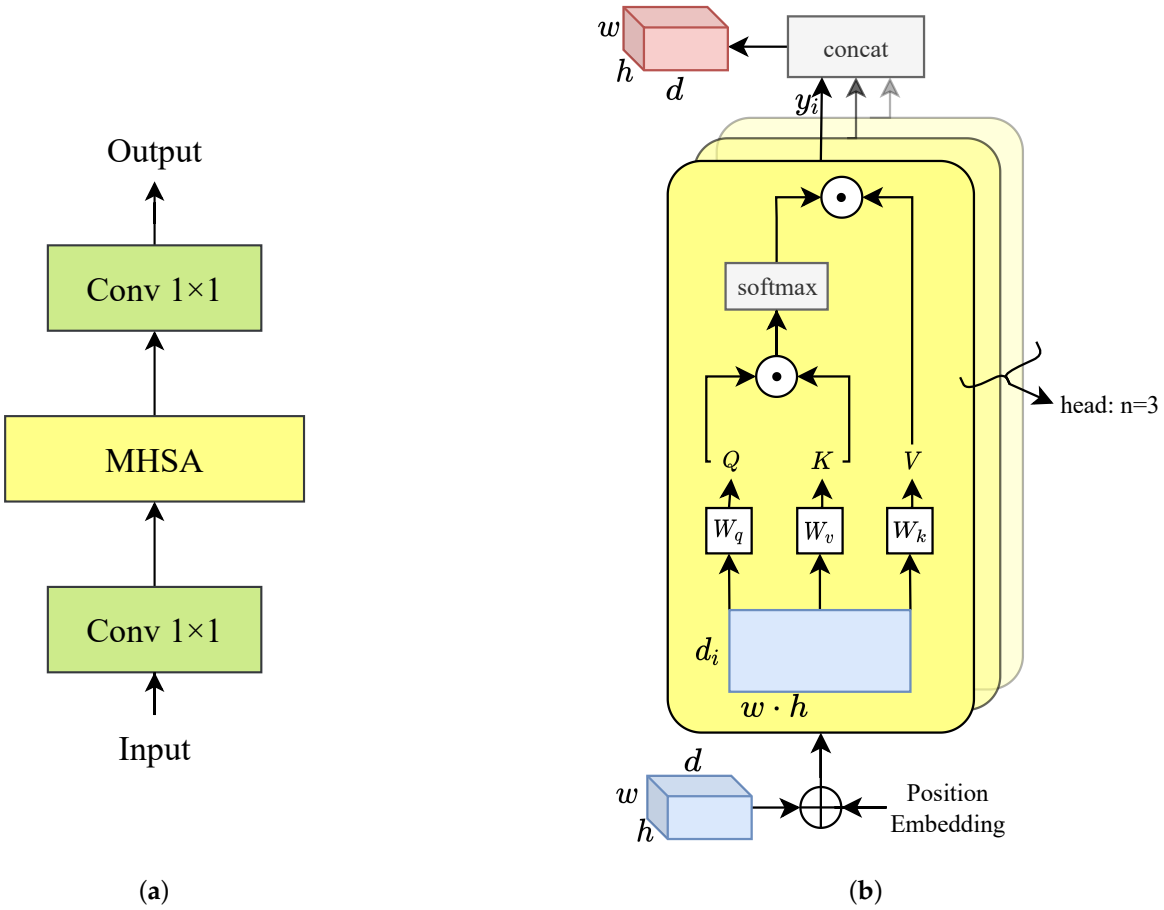

Figure 8. (**a**) The schematic of MSAB. (**b**) Detailed calculation process of MHSA layer, where h, w and d denote height, width and dimension of the input feature and output feature of MHSA layer, respectively, $\odot$ denotes matrix multiplication.

Let x denote the input feature, the output feature map y of MSAB is given as

$$y = concat(y_1, y_2, \ldots, y_n), \tag{6}$$

and the detailed calculation process in each self-attention head is given by

$$y_i = softmax\left(\frac{Q \cdot K^\top}{\sqrt{d_i}}\right) \cdot V, \tag{7}$$

$$Q = W_q x, K = W_k x, V = W_v x, \tag{8}$$

where y_i represents the output feature map of the ith self-attention head, $d_i = d/n$ denotes the dimension of the ith self-attention head. Three different matrixes Q, K, and V, generated by trainable 1×1 convolutions W_q, W_k and W_v times input x, denote queries, keys, and values. In formula (7), attention weights of each self-attention head are generated by $Q \cdot K^\top$ first. Then attention maps are normalized by $\sqrt{d_i}$ and softmax function to obtain the attention scores that contain global contextual information. Finally, the output y_i is yielded by assigning the values of V according to attention scores.

As shown in Figure 8, the input feature x of MHSA layer is appended with positional embedding. With the parallel execution of n heads, the MHSA layer is able to learn the richer non-local context. Considering the high computational complexity in MHSA, only three MHSA layers with four heads are used to construct MSAB in our experiment. In the late stage of the encoder, global attention maps are extracted from $feat_4$ and $feat_5$ with MSAB, respectively, and global attention maps of different stages are aggregated together to capture long-range dependencies from high-resolution feature maps of SAR images.

5. Experimental Results

5.1. Training

The loss function is an essential parameter for CNN training. Considering the imbalanced categorical distribution in the training set for urban water extraction, the loss function based on cross-entropy loss and dice loss is introduced in this work, which is defined as

$$L = L_{ce} + L_d, \tag{9}$$

and

$$L_{ce} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log p_i, \tag{10}$$

$$L_d = 1 - \frac{1}{N}\sum_{i=1}^{N} \frac{2y_i\hat{y}_i}{y_i + \hat{y}_i}, \tag{11}$$

where L_{ce} and L_d represent cross-entropy loss and dice loss, respectively, y_i denotes the ground truth of each pixel, p_i denotes the prediction probability of each pixel, and N denotes the total number of pixels. Cross-entropy loss L_{ce} may have a poor performance since the pixel-wise error is calculated equally for each pixel in Formula (10). Nevertheless, assuming that 0 represents the background and 1 represents water, dice loss L_d can pay attention to water if $y_i\hat{y}_i = 0$ in Formula (11). Therefore, the joint loss function L enables the proposed HA-Unet to perform better for unbalanced samples.

All experiments are based on Pytorch Library, accelerated by 6GB NVIDA 2060 MAX-Q GPU. In order to avoid the impact of hyper-parameters on the experimental results, all hyper-parameters of different models are set as follows: optimizer Adam, initial learning rate 1×10^{-4}, epoch 300 and batch size 2. Finally, all testing samples are fed into the well-trained model to obtain the mapping of urban water.

5.2. Results

The tested SAR image and the ground truth of water bodies are presented in Figure 9a,b. The urban water extraction results generated by DeeplabV3+ [36], original Unet [22], and the proposed HA-Unet are presented in Figure 9c–e. As shown in Figure 9, DeeplabV3+ is good at the extraction of large water bodies but over-detects more radar shadow as water. Compared with DeeplabV3+, Unet performs better in terms of over-detection but misses more water pixels affected by strong scatters. Benefiting from the structural design of hybrid attention, errors in the mapping of urban water generated by the proposed HA-Unet are fewer than the original Unet and DeeplabV3+ in general. To observe the results generated by different models in detail, regions A-C enclosed by green rectangles have been enlarged in Figure 10. Even in complex areas, the omission errors and commission errors in the extraction result generated by HA-Unet are fewer owing to the hybrid attention

mechanism. For example, in the water extraction result generated by the original Unet, more commission errors are caused by radar shadows and some of the hills are misclassified as water bodies in Region A. Additionally, due to the interference of strong scatter, part of water bodies near the strong scatter are misclassified as background in Region B. However, in the water extraction result generated by HA-Unet, by means of the identifiable water features emphasized by CSAM and the guidance of global information captured by MSAB, both commission errors and omission errors are significantly reduced. The experimental results intuitively indicate that HA-Unet can still locate and extract water bodies accurately in urban areas.

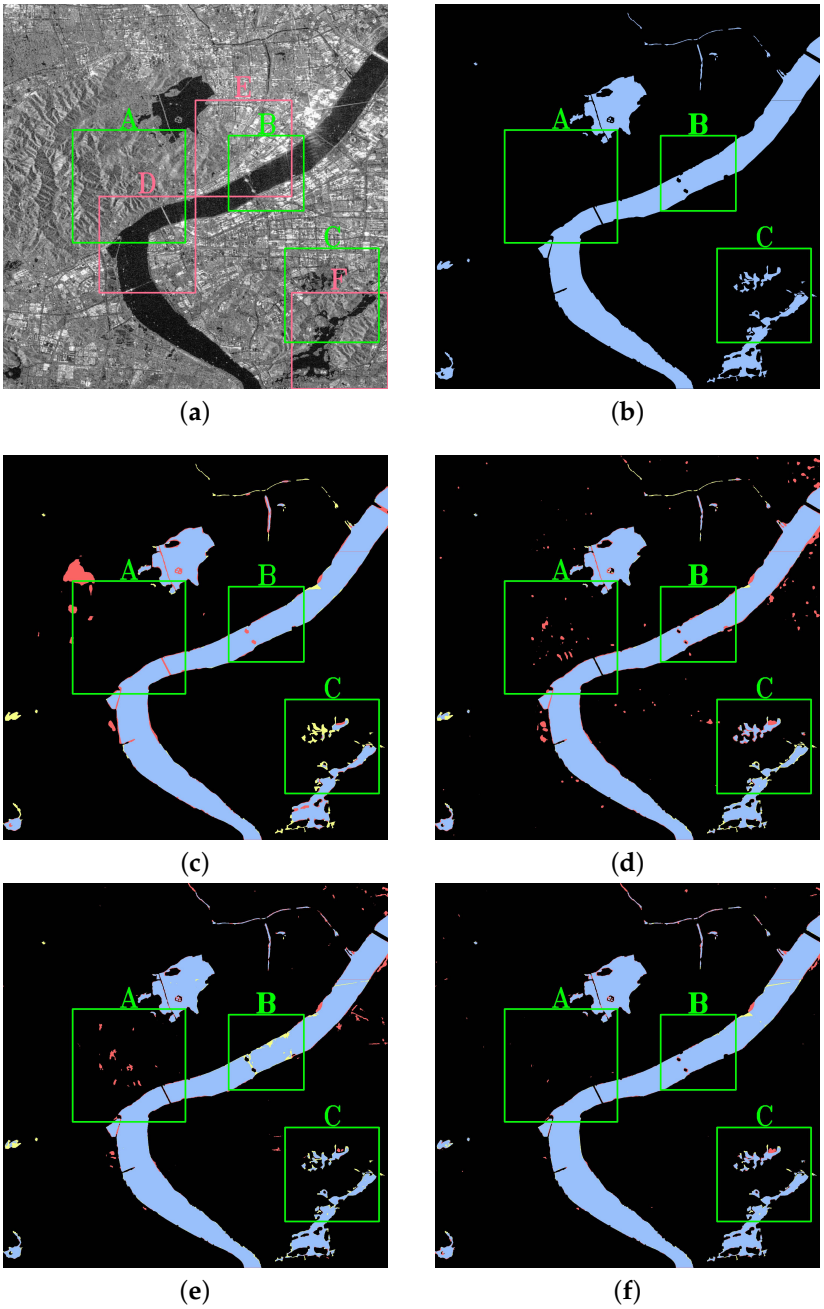

Figure 9. (**a**) The tested SAR image. (**b**)The ground truth of water bodies. Urban water extraction results generated by (**c**) PSPNet, (**d**) DeeplabV3+, (**e**) original Unet and (**f**) the proposed HA-Unet. Yellow denotes omission errors, red denotes commission errors and blue denotes correctly classified water.

To quantitatively evaluate the accuracy of urban water extraction, intersection over union (IoU) and pixel accuracy (PA) based on confusion matrix shown in Table 3 are calculated as follows:

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}, \tag{12}$$

$$\mathrm{PA} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}, \tag{13}$$

where true positive (TP) indicates the number of water pixels that are correctly predicted, false negative (FN) indicates the number of water pixels that are incorrectly predicted as background, false positive (FP) indicates the number of background pixels that are incorrectly predicted as water, and true negative (TN) indicates the number of background pixels that are correctly predicted. As shown in Table 4, the proposed HA-Unet achieves 93.06% IoU and 95.35% PA, outperforming both metrics on the tested SAR image. In comparison with the original Unet, the proposed HA-Unet represents a 6.02% IoU improvement and keeps 7.64% better PA, i.e., the constructed HA-Unet has a better performance in the accuracy of urban water extraction.

Table 3. The confusion matrix.

		Prediction	
		Flood	**Background**
Ground Truth	flood	TP	FN
	background	FP	TN

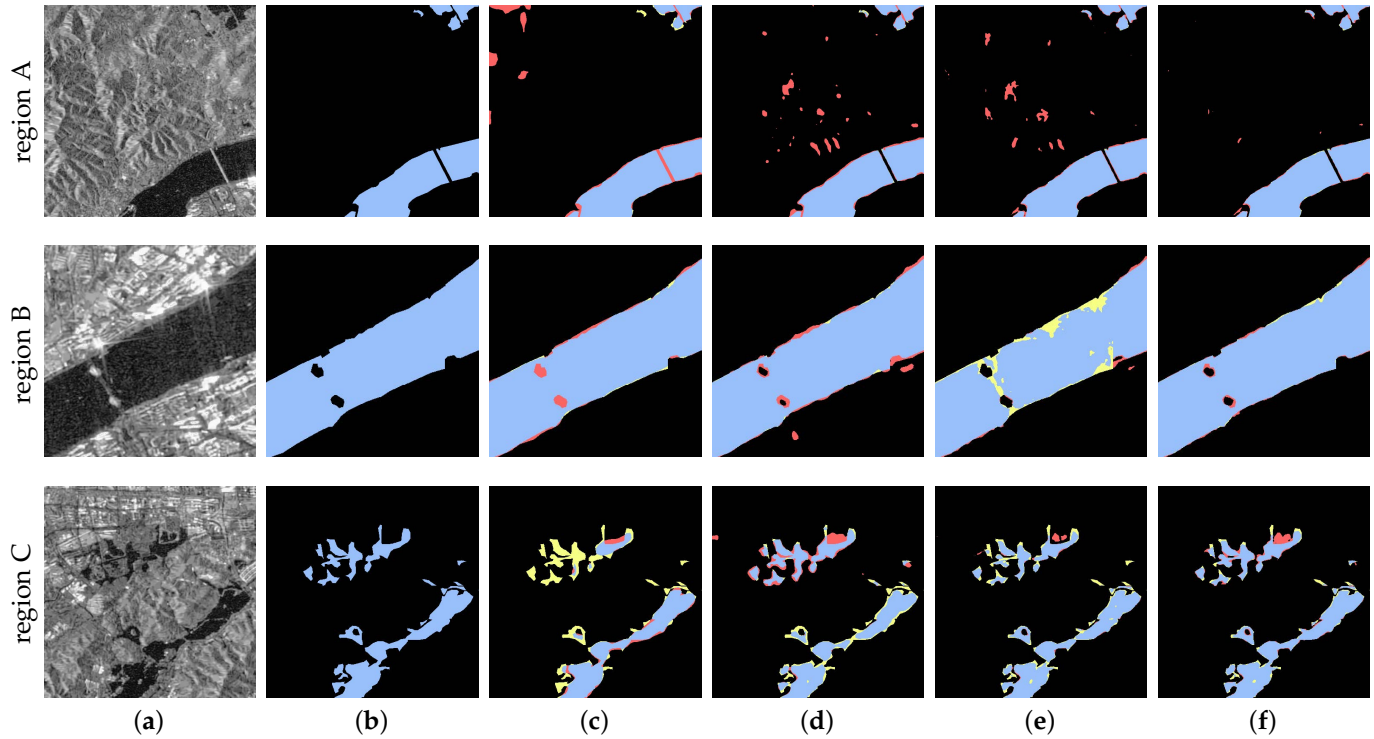

Figure 10. (**a**) Enlarged versions of regions A–C in Figure 9a. (**b**) Ground truth of region A and region B. Urban water extraction results of region A and region B generated by (**c**) PSPNet, (**d**) DeeplabV3+, (**e**) original Unet and (**f**) the proposed HA-Unet.

Table 4. Evaluation of urban water extraction.

	DeeplabV3+	Unet	HA-Unet
IoU(%)	88.56	87.04	**93.06**
PA(%)	90.05	87.71	**95.35**

As mentioned in Section 4, two critical attention modules, CSAM and MSAB, are introduced into shallow layers and deep layers, respectively, in this work. To further gain the deep insights of the improvements obtained by the proposed HA-Unet, an ablation experiment is also performed in our work. The evaluation of ablation experiment is presented in Table 5. In CSAM+Unet, only CSAM is adopted to refine feature maps $feat_1$ and $feat_2$. Furthermore, in MSAB+Unet, only MSAB is introduced into feature maps $feat_4$ and $feat_5$ for global information extraction. It can be seen that either attention module can improve the accuracy of the model compared with the original Unet. In other words, the contributions and effectiveness of the two attention modules in the proposed HA-Unet are undeniable. Finally, combining both the two attention modules yields the best extraction performance

Table 5. The valuation of ablation experiment.

	Unet	CSAM+Unet	MSAB+Unet	HA-Unet
IoU(%)	87.04	90.77	87.87	93.06
PA(%)	87.71	91.89	90.55	95.35

5.3. Visualization

To further explore the role of the hybrid attention mechanism in HA-Unet, the attention maps of regions D-F enclosed by red rectangles in Figure 9a are generated with gradient-weighted class activation mapping (Grad-CAM) [37]. It can be seen that original Unet only pays attention to the local water bodies due to the limited receptive field. As shown in Figure 11c, the global information in the attention map of $feat_5$ of the backbone can be captured based on the long-range dependencies constructed by MSAB. Compared with original Unet, the proposed HA-Unet places more emphasis on the water bodies in the feature map up_1 after CSAM refining the representative water features. These Grad-CAM visualization results indicate that HA-Unet can not only capture the global information in SAR images but also focus on identifiable water features.

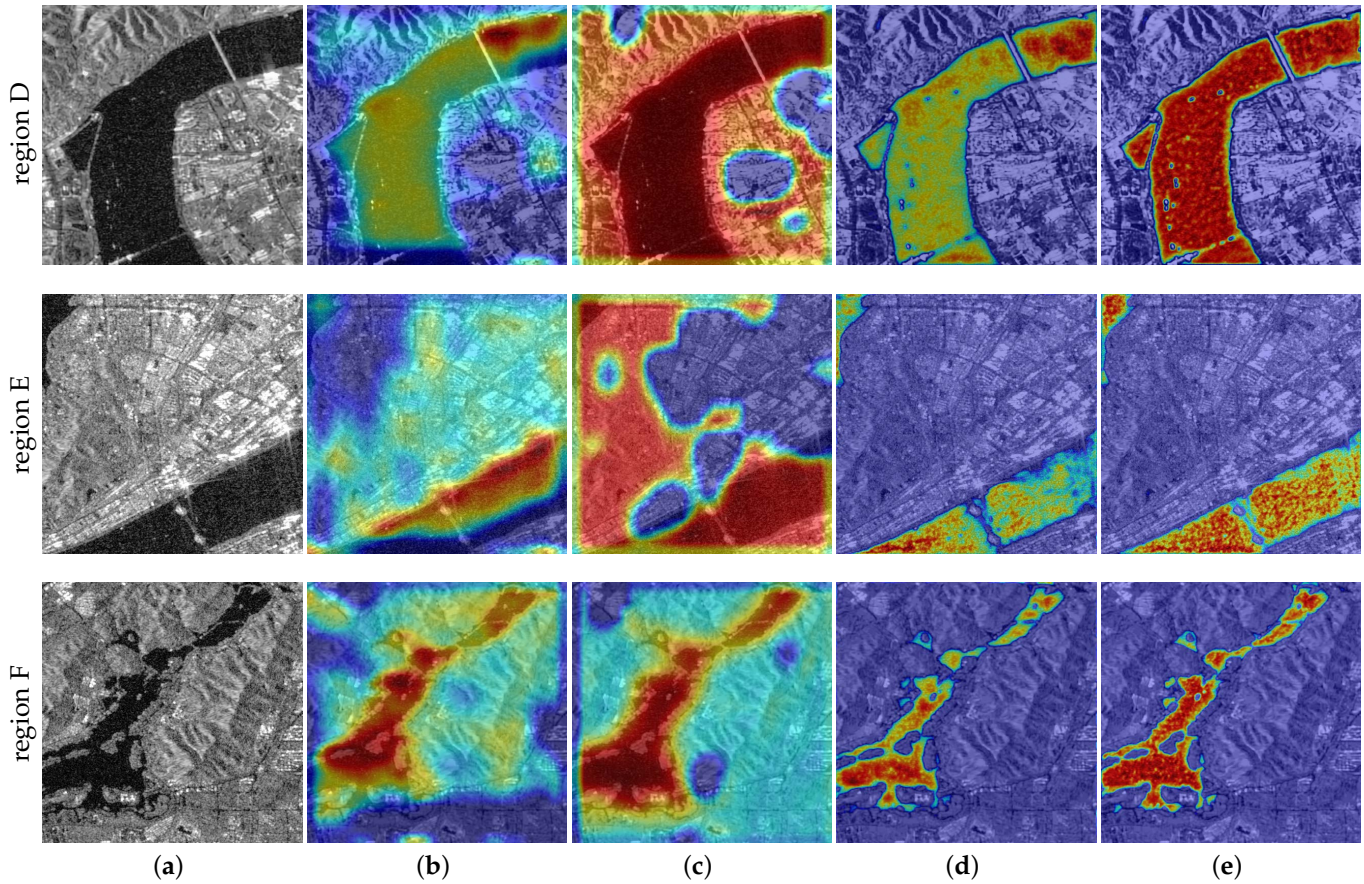

Figure 11. Grad-CAM visualization results. (**a**) SAR image related to regions D-F in Figure 9a. (**b**) The attention map of $feat_5$ in original Unet. (**c**) The attention map of $feat_5$ in the proposed HA-Unet. (**d**) The attention map of up_1 in original Unet. (**e**) The attention map of up_1 in the proposed HA-Unet.

6. Discussion

In this work, the study of urban water extraction in S1A SAR images is carried out. Good results are achieved by the proposed HA-Unet based on the hybrid attention. The proposed HA-Unet is structurally innovative, employing the channel and spatial attention mechanism and the multi-head self-attention mechanism at the same time. First, the low semantic features are enhanced by CSAM to improve the expression of water features. Then, the deep feature maps after MSAB can capture more feature references for local predictions. The urban water extraction results and the quantitative indexes indicate that the proposed HA-Unet is more effective than the original Unet as well as DeeplabV3+ in urban water extraction. Additionally, three enlarged regions A, B and C intuitively show that HA-Unet with the hybrid attention has fewer omission errors and commission errors even in complex scenes. Furthermore, the ablation experiment and visualization results vividly show the important role of HA-Unet in urban water extraction. In comparing the two attention modules, either of the two modules can improve the performance of original Unet gradually in urban water extraction. Thanks to the identifiable water features emphasized by CSAM and the long-range dependency, the proposed HA-Unet can better understand the characteristics of water boundaries, locationsm, and shapes to improve the urban water extraction accuracy.

Of course, there are still limitations and shortcomings to our work. Since MSAB can model long-range dependencies of SAR images, the massive number of parameters in the multi-head self-attention mechanism makes it difficult to meet the requirements of real-time inferencing, in high-resolution SAR images. In this work, MSAB is performed only in the last two stages in Resnet50 to achieve a balance between efficiency and accuracy. In the

future, further research will be continued to explore a more efficient attention mechanism to improve the efficiency of the water extraction algorithm.

7. Conclusions

In this paper, HA-Unet is proposed for urban water extraction in SAR images based on hybrid attention. Considering the feature redundancy in feature maps of standard convolution layers, CSAM based on local attention is adopted to emphasize water features and filter out non-semantic features at early stage of the encoder. In order to compensate for the insufficient global information extraction in the standard convolution layer, MSAB is also introduced to capture global information and long-range interactions of SAR images at late stage of the encoder. In addition, feature maps of the last two stages of the encoder are aggregated to construct high-resolution feature maps for further richer contextual interactions and spatial information. The quantitative evaluations and visualization results of attention maps both indicate that the proposed HA-Unet can extract urban water accurately and effectively. Therefore, the proposed method has great potential in urban water extraction.

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

Funding: This research was funded jointly by National Natural Science Foundation of China grant number 62101167, 61901149, 62101169 and 12071104, National Natural Science Foundation of Zhejiang Province grant number LQ20D010007, LQ20F010007 and LQ22D040001.

Acknowledgments: The authors would like to thank the ESA for providing the time series of Sentinel-1 SAR data.

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

References

1. Ai, J.; Tian, R.; Luo, Q.; Jin, J.; Tang, B. Multi-Scale Rotation-Invariant Haar-Like Feature Integrated CNN-Based Ship Detection Algorithm of Multiple-Target Environment in SAR Imagery. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 10070–10087. [CrossRef]
2. Ai, J.; Mao, Y.; Luo, Q.; Xing, M.; Jiang, K.; Jia, L.; Yang, X. Robust CFAR Ship Detector Based on Bilateral-Trimmed-Statistics of Complex Ocean Scenes in SAR Imagery: A Closed-Form Solution. *IEEE Trans. Aerosp. Electron. Syst.* **2021**, *57*, 1872–1890. [CrossRef]
3. Ai, J.; Mao, Y.; Luo, Q.; Jia, L.; Xing, M. SAR Target Classification Using the Multikernel-Size Feature Fusion-Based Convolutional Neural Network. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–13. [CrossRef]
4. Ai, J.; Luo, Q.; Yang, X.; Yin, Z.; Xu, H. Outliers-Robust CFAR Detector of Gaussian Clutter Based on the Truncated-Maximum-Likelihood- Estimator in SAR Imagery. *IEEE Trans. Aerosp. Electron. Syst.* **2020**, *21*, 2039–2049. [CrossRef]
5. Zhang, J.; Xing, M.; Sun, G.-C.; Chen, J.; Li, M.; Hu, Y.; Bao, Z. Water Body Detection in High-Resolution SAR Images With Cascaded Fully-Convolutional Network and Variable Focal Loss. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 316–332. [CrossRef]
6. Wang, Y.; Li, Z.; Zeng, C.; Xia, G.-S.; Shen, H. An Urban Water Extraction Method Combining Deep Learning and Google Earth Engine. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 769–782. [CrossRef]
7. Li, Y.; Yang, Y.; Zhao, Q. Urban Riverway Extraction from High-Resolution SAR Image Based on Blocking Segmentation and Discontinuity Connection. *Remote Sens.* **2020**, *12*, 4014. [CrossRef]
8. Bao, L.; Lv, X.; Yao, J. Water Extraction in SAR Images Using Features Analysis and Dual-Threshold Graph Cut Model. *Remote Sens.* **2021**, *13*, 3465. [CrossRef]
9. Huang, X.; Xie, C.; Fang, X.; Zhang, L. Combining Pixel- and Object-Based Machine Learning for Identification of Water-Body Types From Urban High-Resolution Remote-Sensing Imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *8*, 2097–2110. [CrossRef]
10. Zeng, C.; Wang, J.; Huang, X.; Bird, S.; Luce, J.J. Urban Water Body Detection from the Combination of High-Resolution Optical and SAR Images. In Proceedings of the 2015 Joint Urban Remote Sensing Event, Lausanne, Switzerland, 30 March–1 April 2015; pp. 1–4.
11. Shi, Q.; Liu, M.; Li, S.; Liu, X.; Wang, F.; Zhang, L. A Deeply Supervised Attention Metric-Based Network and an Open Aerial Image Dataset for Remote Sensing Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–16. [CrossRef]
12. Liao, H.-Y.; Wen, T.-H. Extracting Urban Water Bodies from High-Resolution Radar Images: Measuring the Urban Surface Morphology to Control for Radar's Double-Bounce Effect. *Int. J. Appl. Earth Obs. Geoinf.* **2020**, *85*, 102003. [CrossRef]

13. Giustarini, L.; Hostache, R.; Matgen, P.; Schumann, G.J.-P.; Bates, P.D.; Mason, D.C. A Change Detection Approach to Flood Mapping in Urban Areas Using TerraSAR-X. *IEEE Trans. Geosci. Remote Sens.* **2013**, *51*, 2417–2430. [CrossRef]
14. Guo, Z.; Wu, L.; Huang, Y.; Guo, Z.; Zhao, J.; Li, N. Water-Body Segmentation for SAR Images: Past, Current, and Future. *Remote Sens.* **2022**, *14*, 1752. [CrossRef]
15. Ai, J.; Wang, F.; Mao, Y.; Luo, Q.; Yao, B.; Yan, H.; Xing, M.; Wu, Y. A Fine PolSAR Terrain Classification Algorithm Using the Texture Feature Fusion-Based Improved Convolutional Autoencoder. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14. [CrossRef]
16. Kim, M.U.; Oh, H.; Lee, S.-J.; Choi, Y.; Han, S. Deep Learning Based Water Segmentation Using KOMPSAT-5 SAR Images. In Proceedings of the 2021 IEEE International Geoscience and Remote Sensing Symposium, Honolulu, HI, USA, 25–30 July 2021.
17. Mason, D.C.; Giustarini, L.; Garcia-Pintado, J.; Cloke, H.L. Detection of Flooded Urban Areas in High Resolution Synthetic Aperture Radar Images Using Double Scattering. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *28*, 150–159. [CrossRef]
18. Denbina, M.; Towfic, Z.J.; Thill, M.; Bue, B.; Kasraee, N.; Peacock, A.; Lou, Y. Flood Mapping Using UAVSAR and Convolutional Neural Networks. In Proceedings of the 2020 IEEE International Geoscience and Remote Sensing Symposium, Waikoloa, HI, USA, 26 September–2 October 2020; pp. 3247–3250.
19. Xue, W.; Yang, H.; Wu, Y.; Kong, P.; Xu, H.; Wu, P.; Ma, X. Water Body Automated Extraction in Polarization SAR Images With Dense-Coordinate-Feature-Concatenate Network. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 12073–12087. [CrossRef]
20. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-Local Neural Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7794–7803.
21. Geng, J.; Wang, H.; Fan, J.; Ma, X. SAR Image Classification via Deep Recurrent Encoding Neural Networks. *IEEE Trans. Geosci. Remote Sens.* **2021**, *56*, 2255–2269. [CrossRef]
22. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the 2015 Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241.
23. Chen, F. Comparing Methods for Segmenting Supra-Glacial Lakes and Surface Features in the Mount Everest Region of the Himalayas Using Chinese GaoFen-3 SAR Images. *Remote Sens.* **2021**, *13*, 2429. [CrossRef]
24. Wang, J.; Wang, S.; Wang, F.; Zhou, Y.; Ji, J.; Xiong, Y. Flood Inundation Region Extraction Method Based on Sentinel-1 SAR Data. *J. Catastrophol.* **2021**, *36*, 214–220.
25. Pai, M.; Mehrotra, V.; Aiyar, S.; Verma, U.; Pai, R. Automatic Segmentation of River and Land in SAR Images: A Deep Learning Approach. In Proceedings of the 2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering, Sardinia, Italy, 3–5 June 2019; pp. 15–20.
26. Lalchhanhima, R.; Saha, G.; Sur, S.; Kandar, D. Water body segmentation of Synthetic Aperture Radar image using Deep Convolutional Neural Networks. *Microprocess. Microsyst.* **2021**, *87*, 104360. [CrossRef]
27. Li, J.; Wang, C.; Xu, L.; Wu, F.; Zhang, H.; Zhang, B. Multitemporal Water Extraction of Dongting Lake and Poyang Lake Based on an Automatic Water Extraction and Dynamic Monitoring Framework. *Remote Sens.* **2021**, *13*, 865. [CrossRef]
28. Ren, Y.; Li, X.; Yang, X.; Xu, H. Development of a Dual-Attention U-Net Model for Sea Ice and Open Water Classification on SAR Images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
29. Shamshiri, R.; Nahavandchi, H.; Motagh, M. Persistent Scatterer Analysis Using Dual-Polarization Sentinel-1 Data: Contribution From VH Channel. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 3105–3112. [CrossRef]
30. Di Martino, G.; Di Simone, A.; Iodice, A.; Poggi, G.; Riccio, D.; Verdoliva, L. Scattering-Based SARBM3D. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2016**, *9*, 2131–2144. [CrossRef]
31. Liu, Y.; Zhang, Y.; Wang, Y.; Hou, F.; Yuan, J.; Tian, J.; Zhang, Y.; Shi, Z.; Fan, J.; He, Z. A Survey of Visual Transformers. *arXiv* **2022**, arXiv:2111.06091.
32. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. *arXiv* **2015**, arXiv:1512.03385.
33. Zheng, Z.; Yang, C.; Zhao, J.; Feng, Y. Remote Sensing Geological Classification of Sea Islands and Reefs Based on Deeplabv3+. In Proceedings of the 2022 7th International Conference on Intelligent Computing and Signal Processing, Shenzhen, China, 27–29 May 2022; pp. 1907–1910.
34. Qiu, J.; Chen, C.; Liu, S.; Zeng, B. SlimConv: Reducing Channel Redundancy in Convolutional Neural Networks by Weights Flipping. *IEEE Trans. Image Process.* **2021**, *30*, 6434–6445. [CrossRef]
35. Garcia-Garcia, A.; Orts-Escolano, S.; Oprea, S.; Villena-Martinez, V.; Garcia-Rodriguez, J. A Review on Deep Learning Techniques Applied to Semantic Segmentation. *arXiv* **2017**, arXiv:1704.06857.
36. Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In Proceedings of the 2018 European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 833–851.
37. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In Proceedings of the 2017 IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.