applied sciences

MDPI

Article

An Enhanced Dual-Stream Network Using Multi-Source Remote Sensing Imagery for Water Body Segmentation

Xiaoyong Zhang [1], Miaomiao Geng [1,2], Xuan Yang [3] and Cong Li [2,*]

[1] Beijing Key Laboratory of High Dynamic Navigation, Beijing Information Science and Technology University, Beijing 100101, China; zhangxy@bistu.edu.cn (X.Z.); 2021020341@bistu.edu.cn (M.G.)
[2] State Key Laboratory of Remote Sensing Science, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China
[3] China Remote Sensing Satellite Ground Station, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China; yangxuan@radi.ac.cn
* Correspondence: licong@aircas.ac.cn

Abstract: Accurate surface water mapping is crucial for rationalizing water resource utilization and maintaining ecosystem sustainability. However, the diverse shapes and scales of water bodies pose challenges in automatically extracting them from remote sensing images. Existing methods suffer from inaccurate lake boundary extraction, inconsistent results, and failure to detect small rivers. In this study, we propose a dual-stream parallel feature aggregation network to address these limitations. Our network effectively combines global information interaction from the Swin Transformer network with deep local information integration from Convolutional Neural Networks (CNNs). Moreover, we introduce a deformable convolution-based attention mechanism module (D-CBAM) that adaptively adjusts receptive field size and shape, highlights important channels in feature maps automatically, and enhances the expressive ability of our network. Additionally, we incorporate a Feature Pyramid Attention (FPA) module during the advanced coding stage for multi-scale feature learning to improve segmentation accuracy for small water bodies. To verify the effectiveness of our method, we chose the Yellow River Basin in China as the research area and used Sentinel-2 and Sentinel-1 satellite images as well as manually labelling samples to construct a dataset. On this dataset, our method achieves a 93.7% F1 score, which is a significant improvement compared with other methods. Finally, we use the proposed method to map the seasonal and permanent water bodies in the Yellow River Basin in 2021 and compare it with existing water bodies. The results show that our method has certain advantages in mapping large-scale water bodies, which not only ensures the overall integrity but also retains local details.

Keywords: attention mechanism; swin transformer; convolutional neural networks; feature pyramid attention; Yellow River Basin

Citation: Zhang, X.; Geng, M.; Yang, X.; Li, C. An Enhanced Dual-Stream Network Using Multi-Source Remote Sensing Imagery for Water Body Segmentation. *Appl. Sci.* **2024**, *14*, 178. https://doi.org/10.3390/app14010178

Academic Editor: Francesco Zirilli

Received: 28 November 2023
Revised: 17 December 2023
Accepted: 22 December 2023
Published: 25 December 2023

1. Introduction

Water is a fundamental natural element and an invaluable resource for various human endeavors [1,2]. Contemporary shifts in global climate, coupled with intensified human activities, have precipitated pronounced spatial and temporal variations in water resources, commonly manifesting as reductions in both area and volume [3]. These problems pose substantial challenges to human livelihoods and agricultural production, underscoring the imperative for meticulous extraction and dynamic monitoring of surface water bodies. Remote sensing technology has become an effective tool for such monitoring, attributed to its accessibility, extensive coverage, frequent updating, and immediacy [4,5]. In recent years, the growing water use conflict in the Yellow River Basin has seriously affected human welfare and livelihoods, highlighting the critical importance of surveilling the dynamic fluctuations of Surface Water Area (SWA) within this region.

In recent years, deep learning has been widely used in tasks such as image classification [6], object detection [7], and semantic segmentation [8]. Water bodies extraction is a typical semantic segmentation task, where methods rooted in deep learning significantly improve the accuracy of traditional extraction techniques [9]. However, the application of deep learning in water bodies extraction faces several challenges, including diverse morphologies, complex scenarios (e.g., rivers, lakes, urban water bodies, and aquaculture), and variations in dimensions, with particular difficulty presented by small water bodies [8]. Convolutional Neural Network (CNN) have emerged as a prevalent strategy in water bodies extraction research, owing to their proficiency in semantic information of remote sensing images through convolutional operations processes [8,10–12]. Despite their merits, CNNs are constrained by their receptive field dimensions, leading to a predominant focus on local features while ignoring the spatial correlation within a water body. This limitation often culminates in inaccurate edges of lake boundaries [10,11]. Furthermore, the progressive enlargement of the receptive field amidst continuous down-sampling can result in the neglect of smaller water bodies' features, leading to fragmented extraction outcomes [12]. The advent of Fully Convolutional Network (FCN) marked a pivotal transition by introducing an encoder-decoder structure for image segmentation, facilitating a shift from image-level classification to a more fine-grained, pixel-level classification [13,14]. Li et al. [15] used a FCN model for water bodies extraction from Very High Resolution (VHR) images, demonstrating the efficacy of pixel-level classification. However, this method suffers from the inaccurate extraction of small-size water bodies due to an inherent information attrition during the convolutional operations.

To address the limitations of CNNs, particularly the loss of information through continuous down-sampling, researchers have integrated attention mechanisms and dilated convolution modules into existing methodologies. Zhong et al. [16] introduced a two-way channel attention mechanism and a depth-expanded residual structure forming MIE-Net network, aiming to improve lake segmentation accuracy. Gosula [17] demonstrated the superior performance of a DeeplabV3+ model with an Atrous Spatial Pyramid Pooling (ASPP) module in extracting various water body sizes compared to its counterpart without ASPP structure. Yuan et al. [18] proposed an Enhanced ASPP (EASPP) module, adept at maintaining the prominence of high-dimensional features and improving the recognition of small-size water bodies. However, the primary focus remains on expanding the receptive field, rather than truly encoding contextual information. This limitation tends to overshadow the intricate local traits of small-sized water bodies, leading to inconsistencies in the extraction outcomes for rivers and lakes [19]. The emergence of a Transformer model, a sequence-to-sequence architecture based on the attention mechanism, heralds new prospects. Unlike traditional CNNs, a Transformer leverages multi-head attention to establish remote dependencies, thereby facilitating a more holistic feature extraction from images [20]. This technique is gaining traction in water bodies extraction [11,21]. Zhao et al. [11] introduced Convolutional Block Attention Module (CBAM) and substituted the conventional convolutional layers in the original U-Net model encoder with Vision Transformer (Vi-T), enhancing the continuous spatial relationship interpretation, and thereby improving the segmentation of lake water bodies on the Qinghai-Tibet Plateau. However, Transformer architecture, which compresses the image into one-dimensional tokens, interpreted as sequences, has its limitations. Specifically, this approach can result in the loss of internal and local image information, potentially impeding the recovery of fine-grained details during the decoding stage [21]. The fusion of optical and microwave images emerges as a robust solution, providing complementary visuals and fortifying the database requisite for feature identification [22–24]. Initial research efforts primarily focused on the amalgamation of multispectral and radar data through shallow feature fusion, as demonstrated in studies by Filsa [25]. These classical classifiers rely on feature engineering and exhibited limited generalization proficiencies. Inspired by deep learning in Computer Vision (CV), the powerful feature extraction capability of Convolutional

Neural Networks (CNNs) has been successfully applied in feature classification, with contemporary research gravitating towards multimodal learning.

In general, we note that there are still some problems when conducting fine extraction and dynamic monitoring of large-scale water bodies: (1) For water bodies with variable shapes and complex scenes, existing deep learning studies use convolution kernels to encode key features, which lack global information between pixels, resulting in inconsistent water extraction results and inaccurate lake extraction edges. (2) Continuous down-sampling of convolution operations will lead to information loss, and the extraction results of small water bodies will be off-flow.

In order to solve the above problems, the contributions of this paper are as follows:

(1) Using the parallel coding structure of Swin-T and CNN, while adaptively enhancing the local features of the image, the global association between pixels is strengthened, improving the accuracy of water boundary extraction and the consistency of the results.
(2) An attention mechanism module is proposed, which automatically highlight the important channels of the feature map according to the characteristics of the image feature map and avoids the problem of excessive attention of convolution on irrelevant local features.
(3) In the advanced feature coding stage, the FPA module is used to construct the feature pyramid to obtain the advanced semantic features of different scales of the image and improve the segmentation ability of multi-scale water bodies. We construct a water extraction network based on multi-source data, and use the network to extract water bodies, achieving high segmentation accuracy. In addition, we also mapped the seasonal water body and permanent water body of the Yellow River basin in 2021 and compared them with existing water body products.

2. Related Work

2.1. CNN-Based Semantic Segmentation for Water Body Extraction

In the field of water body extraction, CNN has been the dominant method and significant progress has been made. Recently, WaterHRNet [1] is a high-performance fully convolutional network that focuses on enhancing the results and detail representation of water body segmentation. It achieves this by focusing on features in the foreground region from a global perspective. This is accomplished through parallel channel-specific attention and spatial-specific attention, which enhance the differentiation between the foreground and background. MSAFNet [26] is a network that employs multiple attentional modules to extract both low-level and high-level local features. These multi-level features are then aggregated through the feature fusion module (FFM) to improve the mapping of water body edges, directly enhancing the segmentation of water bodies. To accurately fuse multi-scale contextual information, MSNANet [27] utilizes the MSNA module, establishing remote feature dependencies between channels to improve the accuracy and mapping capability of the model. In related work, DUPNet [28], a network structure with a U-shaped encoder and decoder, uses dense blocks to extract image semantic features and obtain highly abstract feature maps. Instead of employing the maximum pooling layer, it adopts Atrous Separable Convolution during down-sampling to increase the perceptual field of view of the feature maps, thereby enhancing the robustness of the network model. The aforementioned study demonstrates the paramount importance of integrating multi-scale features in the task of water body extraction.

2.2. CNN and Transformer-Based Semantic Segmentation for Water Body Extraction

CNN architectures have been widely used for water body segmentation tasks, but they often suffer from induction bias and a lack of long-range dependencies. To address this limitation, researchers have explored Transformer-based architectures that leverage the self-attention mechanism to encode global information. By integrating CNN and Transformer, these architectures can consider both spatial location information and local details of the

water body, effectively combining global and local features. One such improvement is seen in Sgformer [29], which utilizes the SGU module to integrate local details and contextual semantic information. It fuses the features from both the CNN and Transformer branches, effectively addressing the problem of semantic divergence at different levels and reducing the likelihood of misdetection and omission of water bodies. Another study by Zhang et al. [30] employs the Swin Transformer as the backbone network and integrates multi-scale contextual information through the SASPP module, reducing the chances of false detection and omission of water bodies. They also use CNN to reconstruct features in the decoding stage and redistribute the weights of multiresolution feature maps using the SE module, enhancing the capability of water body boundary segmentation. Peng et al. [31] introduce Conditional Position Encoding (CPE) as a replacement for traditional position encoding, enabling efficient interaction between local and global information in the encoding stage of the ViT model to enhance glacier segmentation. Meanwhile, ERSSeg [32] develops a more efficient self-attentive transformer backbone network based on Segformer. It introduces the Feature Fusion Module (FFM) and location-free coding FFN to enhance the decoder's ability to reconstruct the image and capture spatial and positional information independently of the input size. These improved approaches provide strong support for water body segmentation tasks by overcoming the limitations of CNN architectures.

3. Methodology

3.1. Proposed Deep Learning Architecture

This study introduces a novel Dual-Stream feature extraction network, termed DSST-UNet, which synergizes principles from CNN and Transformer models, specifically Swin-Unet, to enhance water bodies' extraction using fused multimodal data, including optical and SAR imagery. Inputs to the network are distinct patches from both optical and SAR images. The comprehensive DSST-UNet architecture is depicted in Figure 1: 1(a) outlines the entire network, 1(b) details CNN branch (DCB-ConvNet), and part 1(c) elucidates the elements within 1(a).

DSST-UNet encoder, dedicated to feature extraction and fusion, cooperates through three main steps:

(1) Pseudo-Siamese feature extraction: Illustrated in Figure 2, this phase involves the parallel processing of optical and SAR image patches through two distinct sub-networks within Pseudo-Siamese feature extractor, each functioning independently without shared weights. Each sub-network yields four hierarchical feature maps, simultaneously decreasing in spatial dimensions and increasing in channel depth. Specifically, the optical images produce feature maps {O1, O2, O3, O4}, and SAR images yield {S1, S2, S3, S4}, with corresponding spatial resolutions of {128, 64, 32, 16} and channel counts of {96, 192, 384, 768}. Subsequently, a fusion process combines corresponding feature maps from the optical and SAR pathways, resulting in an integrated feature set denoted as {C1, C2, C3, C4}.

(2) Dual-Stream feature encoding: As outlined in Figures 3 and 4, the encoder's initial three stages involve a parallel configuration of a Transformer branch and a CNN branch (DCB-ConvNet). The Transformer branch employs a self-attention mechanism to assimilate the global context across varying image locations, thereby securing comprehensive semantic insights. Conversely, the CNN branch (DCB-ConvNet) harnesses an enhanced D-CBAM to meticulously discern local attributes such as texture, contours, and shapes, augmenting the system's proficiency in local detail recognition of diverse water body forms. These branches operate concurrently, generating a feature map that amalgamates both global and local image information. During this phase, the merged features from each optical image stage are labeled {O1, O2, O3}, and those from each SAR image stage are {S1, S2, S3}. Subsequent to this, a fusion procedure combines the stage-specific optical image feature maps with those of SAR image to create multimodal features. This fusion involves channel-dimension splicing of O1 with S1 to form a novel feature tensor, followed by a dimension-reduction technique (using 1×1 convolution) that rescales the spliced tensor down to C1. This method of splicing and dimension reduction is consistently applied to O2 with S2 and

O3 with S3, yielding the compounded features C2 and C3. This dual-stream encoding strategy effectively merges the multimodal data from the optical and SAR images, resulting in integrated feature maps from the first three stages {C1, C2, C3}.

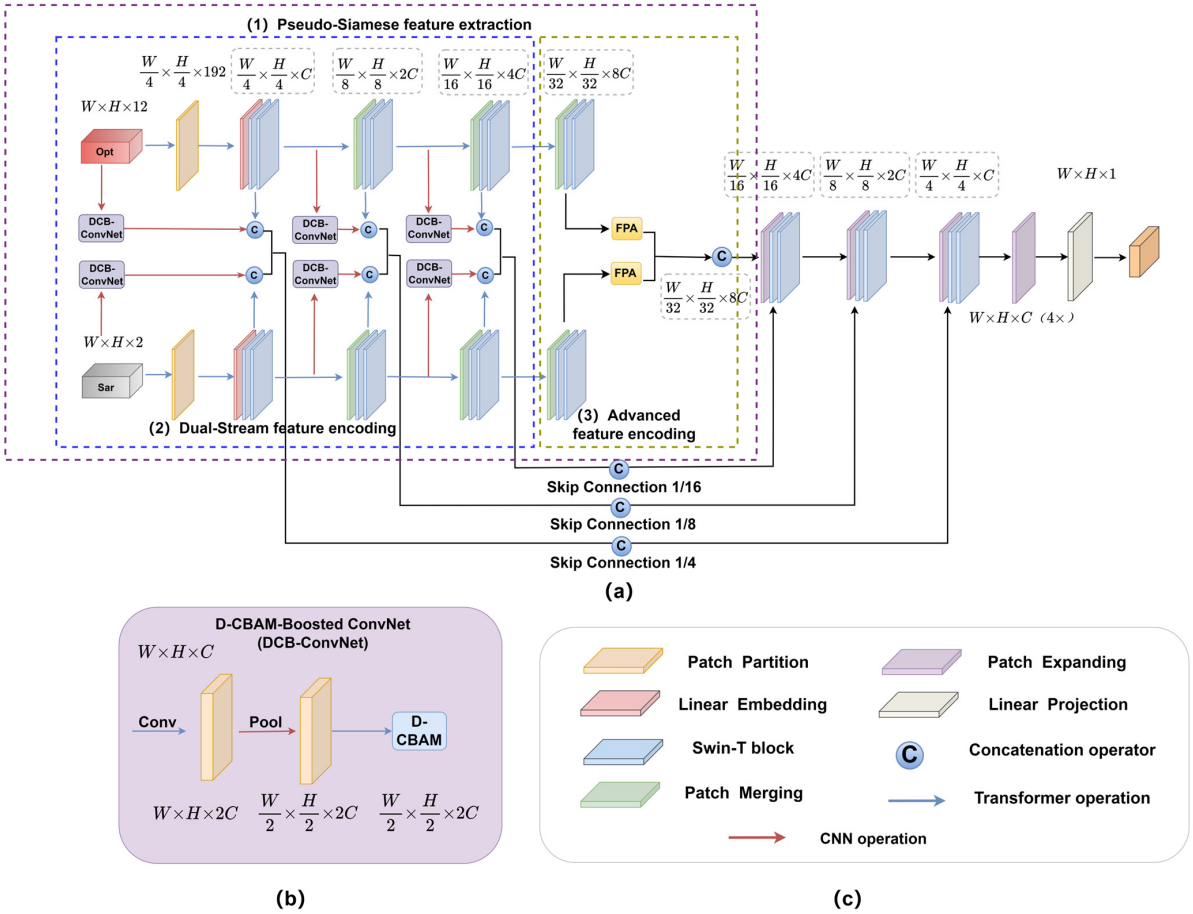

Figure 1. Overview of the network structure. Figure 1: (**a**) outlines the entire network, (**b**) details CNN branch (DCB-ConvNet), and part (**c**) elucidates the elements within (a). Figure 2 corresponds to (1); Figure 3 corresponds to (2); Figure 4 corresponds to (3); Figure 5 corresponds to the process of fusion of multi-source data features at each stage of the process in (2) and (3); Figure 6 corresponds to D-CBAM; Figure 8 corresponds to FPA.

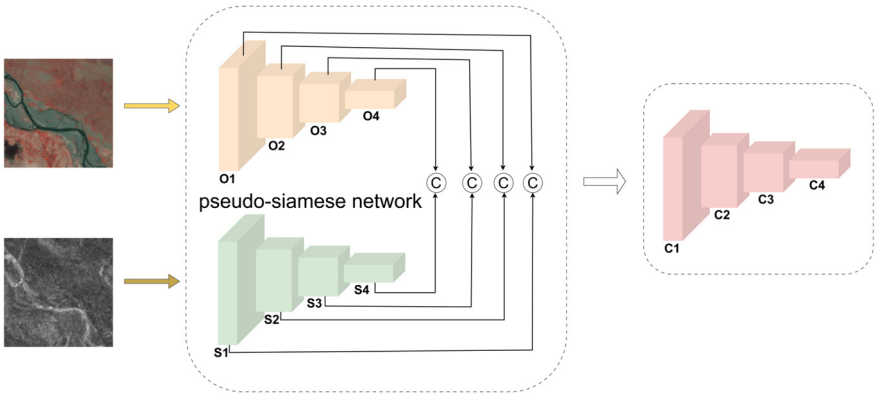

Figure 2. Structure of Pseudo-Siamese feature extraction.

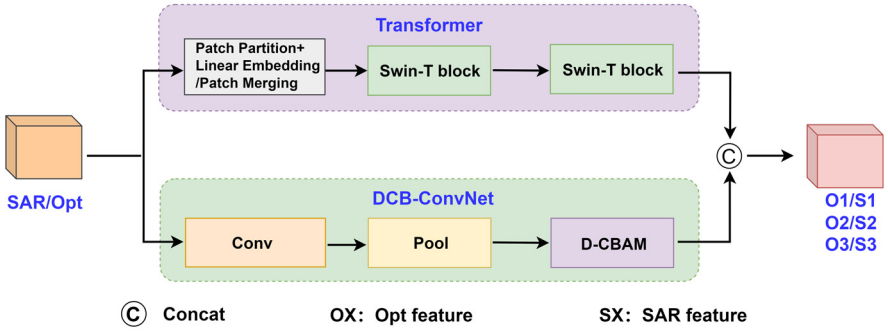

Figure 3. Structure of dual-stream feature encoding process (refer to Figure 6 for D-CBAM structure).

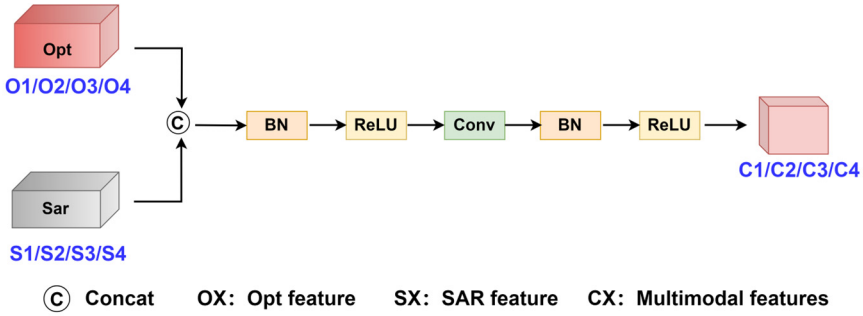

Figure 4. Structure of multimodal feature fusion process.

(3) Advanced feature encoding: As depicted in Figure 5, the final segment of the encoder phase involves the intricate integration of the Transformer branch with a FPA module. Within this process, the Transformer branch operates independently on the designated optical and SAR image segments, dedicated to nuanced feature extraction. Following this, a FPA module [33] is applied to the preliminary extraction outcomes, producing refined outputs O4 and S4. The FPA module plays a pivotal role in extracting profound semantic information from elevated-level feature maps, simultaneously reinstating precision in pixel localization. Throughout this stage, the module expands the receptive field comprehensively across multiple scales, thereby recovering intricate pixel localization details. Such augmentation is crucial for the recognition of water bodies on multiple scales. The terminal part of this phase encompasses the multimodal feature fusion, previously elaborated, resulting in the composite output referred to as C4. This fusion signifies the termination of the encoding section, yielding an array of multimodal feature maps {C1, C2, C3, C4}, consistently throughout the four integral stages of this phase.

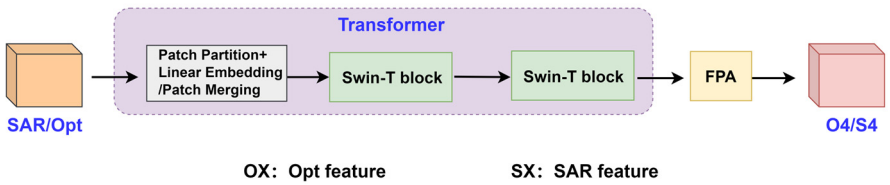

Figure 5. Structure of advanced feature extraction process (refer to Figure 8 for FPA structure).

The decoder serves to restore spatial dimensions and refine object specifics [34], converting encoded feature maps into unified semantic segmentation maps for both optical and SAR images. Within DSST-UNet, the decoder employs Swin-Unet's structure. This involves a Patch Expanding mechanism and successive Swin-T modules, collectively constituting an up-sampling layer. Feature information undergoes consolidation into a novel feature map via skip-connections, integrating sub-high-level data from the corresponding down-

sampling layer. The process culminates with the generation of a water body segmentation map, accomplished through a Linear Projection layer.

3.2. D-CBAM

While CNNs adeptly discern local features and spatial information via convolutional operations, they also address certain Transformer limitations, such as accuracy in localization and extraction of local features. During network training, prioritizing more salient features is desirable, making the integration of an attention mechanism within convolutional operations a potent strategy [35,36]. CBAM, a streamlined attention module, sequentially employs channel and spatial attention, effectively interpreting channel and spatial information within feature maps. To enhance the expressive capacity of the convolutional attention module, the study introduces an advanced form of CBAM, termed D-CBAM, which optimizes the original structure. The architecture of D-CBAM is depicted in Figure 6. The structure of D-CBAM is shown in Figure 6.

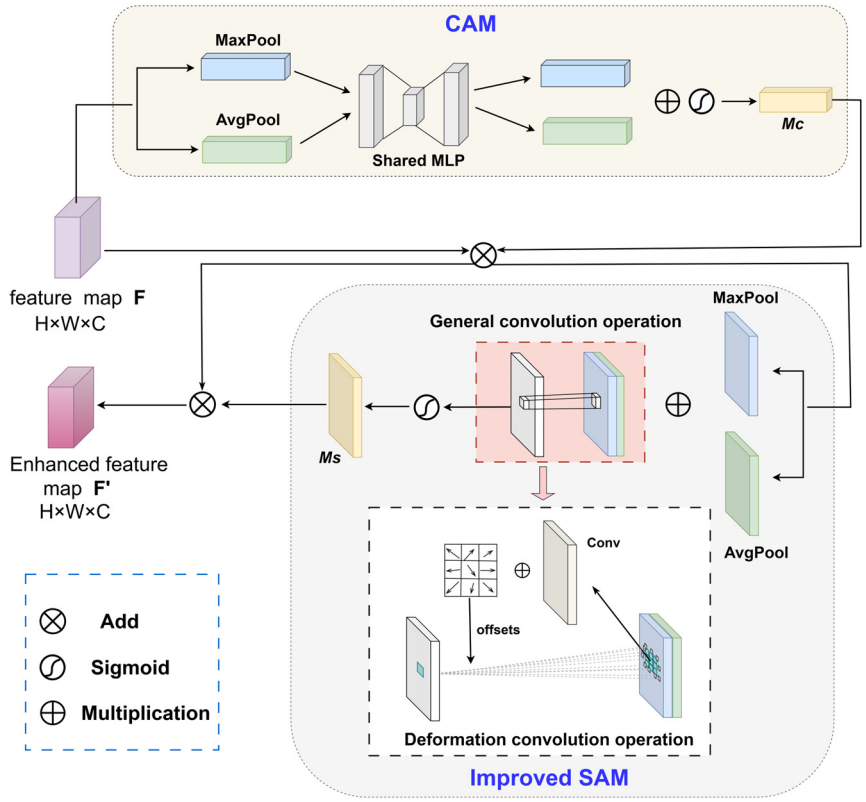

Figure 6. Structure of D-CBAM.

Channel attention facilitates the allocation of distinct attention levels to different channels at each position within an image, as delineated in Equation (1):

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

where M_c denotes the channel attention map; σ denotes the Sigmoid function; MLP denotes the multi-layer perceptron; $MaxPool$ denotes the global maxinum pooling; $AvgPool$ denotes the global average pooling.

The deformable convolutional layer, with its inherent ability to learn offsets, can dynamically alter the shape and size of the receptive field in accordance with the specific characteristics of water bodies within an image. This advanced feature substantially augments the network's proficiency in pinpointing the accurate spatial attributes of water bodies. Consequently, this study introduces an enhancement to the spatial attention module

within CBAM by substituting traditional convolution, which conducts sampling of a fixed size and is thus termed regular convolution, with deformable convolution. Deformable convolution distinguishes itself by incorporating a learnable offset for each sampling point within the receptive field, providing a more adaptable approach to feature extraction [37]. The mathematical representation of deformable convolution is articulated in Equation (2):

$$y(p_0) = \sum_{p_n \epsilon R} w(p_n) \times x(p_0 + p_n + \Delta p_n) \quad (2)$$

where $w(p_n)$ denotes Convolution checks the weight of the position; Δp_n denotes offsets; $x(p_0 + p_n + \Delta p_n)$ denotes the offset element value at $p_0 + p_n$ position on the input feature map.

In the enhanced spatial attention module, feature set F′, refined by the channel attention mechanism, undergoes global average pooling and max pooling. These procedures, executed across the channel dimension, convert F′, originally with dimensions H × W × C, into two feature tensors, each with dimensions 1 × 1 × C. These tensors are then concatenated along the channel dimension, resulting in a single 1 × 1 × 2C feature tensor. This tensor is subsequently processed through a 3 × 3 deformable convolutional layer, emphasizing the adaptability to the image's spatial characteristics. The Sigmoid activation function is then applied to this convoluted output, yielding weight matrices that signify the relative importance of various spatial components within the image. These weight matrices are element-wise multiplied with the original feature set F′, assigning distinctive weights to different spatial elements in the image. This process is mathematically captured in Equation (3), with the deformable convolution-based spatial attention module illustrated in Figure 7:

$$M_s(F) = \sigma\left(f_d^{3\times3}([AvgPool(F); MaxPool(F)])\right) \quad (3)$$

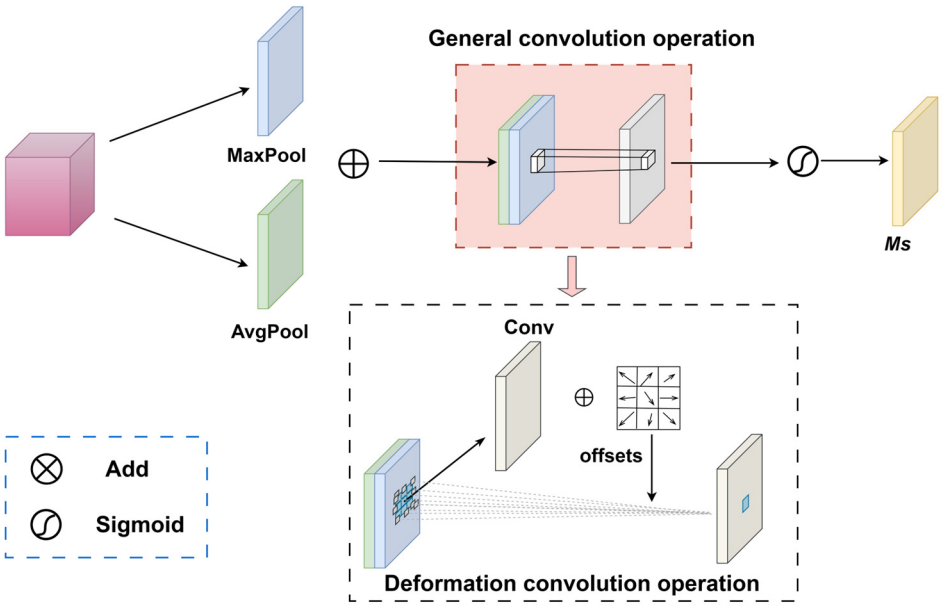

Figure 7. Structure of improved SAM.

3.3. FPA

In the last level of DSST-UNet encoder, convolutions of varying scales (7 × 7, 5 × 5, 3 × 3) are applied to high-level features to extract multi-scale features. These features are then up-sampled, elevating lower-level details to higher-level features, thereby imbuing them with pixel-level accuracy. This process ensures that features crucial for water body identification receive higher priority. Following this, refined features, processed through

a 1 × 1 convolution, are integrated with pyramid attention features via element-wise multiplication, resulting in an encompassing enhancement of the feature set. Finally, these composite features are merged with the global pooling branch, culminating in the comprehensive pyramid attention features. The advanced feature extraction FPA module shown in Figure 8.

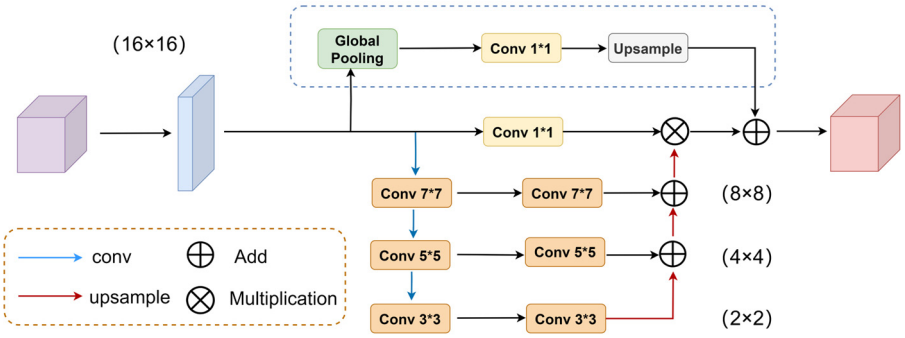

Figure 8. Structure of FPA.

3.4. Loss Function

In the network model, the chosen loss function integrates Binary Cross Entropy (BCE) [38] and Dice coefficients [39]. The task of water body delineation falls under binary classification within semantic segmentation, targeting the discernment of water bodies from their surrounding terrain in images. BCE, a standard in binary classification scenarios, is adopted as the foundational loss metric. This loss criterion quantifies the discrepancies at the pixel level between the predicted outputs and actual labels, employing natural logarithm calculations as detailed in Equation (4):

$$L_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}(y_i log\hat{y}_i + (1-y_i)(1-log\hat{y}_i)) \quad (4)$$

where N denotes the number of samples; y_i denotes the true label value; $y_i \in \{0, 1\}$; and $\hat{y}_i$ denotes the predicted probability of a particular area being classified as a water body.

However, BCE loss focuses on pixel-level classification accuracy, posing challenges when subjects are minimal or sparse, resulting in the loss value being disproportionately influenced by the background imagery. Particularly in water body identification tasks, images from mountainous or urban regions frequently feature limited water areas. To address this imbalance, the Dice Loss function is incorporated, computing the congruence between actual labels and predictive outputs from the model. Dice Loss alleviates the detrimental effects arising from foreground-background area discrepancies within the samples, as delineated in Equation (5):

$$L_{DICE} = 1 - \frac{2|Y \cap P|}{|Y| + |P|} \quad (5)$$

where Y denotes the true labeled value; P denotes the model predicted value; and $|\cdot|$ denotes the number of elements in the set.

The overall loss function of the network model can be expressed as in Equation (6):

$$Loss = \lambda_1 \times L_{BCE} + \lambda_2 \times L_{DICE} \quad (6)$$

where $\lambda_1 + \lambda_2 = 1$, the performance of the model is optimized through the calibration of the weights attributed to the two loss functions during the training process.

3.5. Methods for Analyzing Seasonal Water Bodies

Generally, Water Inundation Frequency (WIF) is described as the proportion of instances a specific image location is identified as water over a certain duration, relative to the total count of suitable observations (unobstructed by clouds, cloud shadows, or snow and ice) [40]. This is mathematically represented as in Equation (7):

$$WF = \frac{\sum_{i=1}^{N} w}{N} \times 100\% \tag{7}$$

where N represents the total number of times that have been effectively observed; $w \in \{0, 1\}$, $w = 1$ means the pixel is a water body, and $w = 0$ means the pixel is not a water body.

In this study, the 2021 monthly water bodies extraction results for the Yellow River Basin are analyzed through WIF, defined at each pixel as the ratio of the number of times water is detected to the total number of months. Utilizing WIF values, water bodies are classified as permanent (WIF $\in$ (75%, 100%]), seasonal (WIF $\in$ (25%, 75%]), or non-existent (WIF $\in$ (0, 25%]). Essentially, locations registering water presence over nine times annually are deemed permanent, those observed three to nine times are categorized as seasonal, and those under three times are designated as non-water bodies. This analysis focuses on seasonal and permanent categories, collectively referred to as year-round effective water bodies, forming the basis for subsequent product comparisons.

3.6. Evaluation Metrics

The model's efficacy is assessed using key performance metrics, including accuracy, Recall, F_1 Score, and Intersection Over Union (IoU). The F_1 Score, a harmonized mean of accuracy and Recall, encapsulates the model's accuracy and inclusiveness. Conversely, IoU offers an extensive evaluation by quantifying the congruence between predicted and actual water body areas. The respective computations for these evaluation metrics are as follows:

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Reacll} \tag{10}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{11}$$

where TP is a pixel correctly recognized as a water body; FP is a background pixel incorrectly recognized as a water body; and FN is a water body pixel incorrectly classified as a background.

3.7. Training Details

In this research, all neural network models were constructed utilizing the PyTorch framework. The training phase employed the complete array of training and validation samples, with computations facilitated by four NVIDIA TITAN XP GPUs, each equipped with 12 GB of RAM. The training protocol stipulated a batch size of 12, utilizing AdamW optimizer [41]. An initial learning rate of 0.001 was established, with a comprehensive iteration count of 200. This learning rate was subject to dynamic modification contingent on F1-score observations from the validation set. Specifically, a stagnation in the improvement of F1-score over 20 epochs triggered a reduction of the learning rate by 50%. This rate adjustment could be enacted up to three times or until the prescribed iteration limit was attained. Pertaining to the loss function's hyperparameters, the values were fixed at 0.7

for λ_1 and 0.3 for λ_2. Furthermore, batch normalization [42] techniques were applied to expedite model convergence.

4. Experiment

4.1. Study Area

The Yellow River Basin, illustrated in Figure 9, lies in northern China, extending from 32° to 42° N and 96° to 119° E. Its geographical stretch covers around 1100 km from north to south and approximately 1900 km from east to west, with a total area of 795,000 km^2, making it China's second-longest river [4]. The basin's topography is intricate, characterized by a decline from west to east, with western regions marked by plateaus and mountains, central areas by plains and basins, and eastern locales by plains and hills. The basin hosts a dense urban population, predominantly in its central and lower segments. Water body morphology within the Yellow River Basin presents substantial complexity and variety. The upper basin, known for its high altitude, features numerous perennially frozen lakes due to its colder climate. The middle basin contends with significant sedimentation from Loess Plateau, resulting in rivers with high sand content. Conversely, the lower basin lacks prominent large-scale tributaries feeding into the river, creating a distinct fluvial environment.

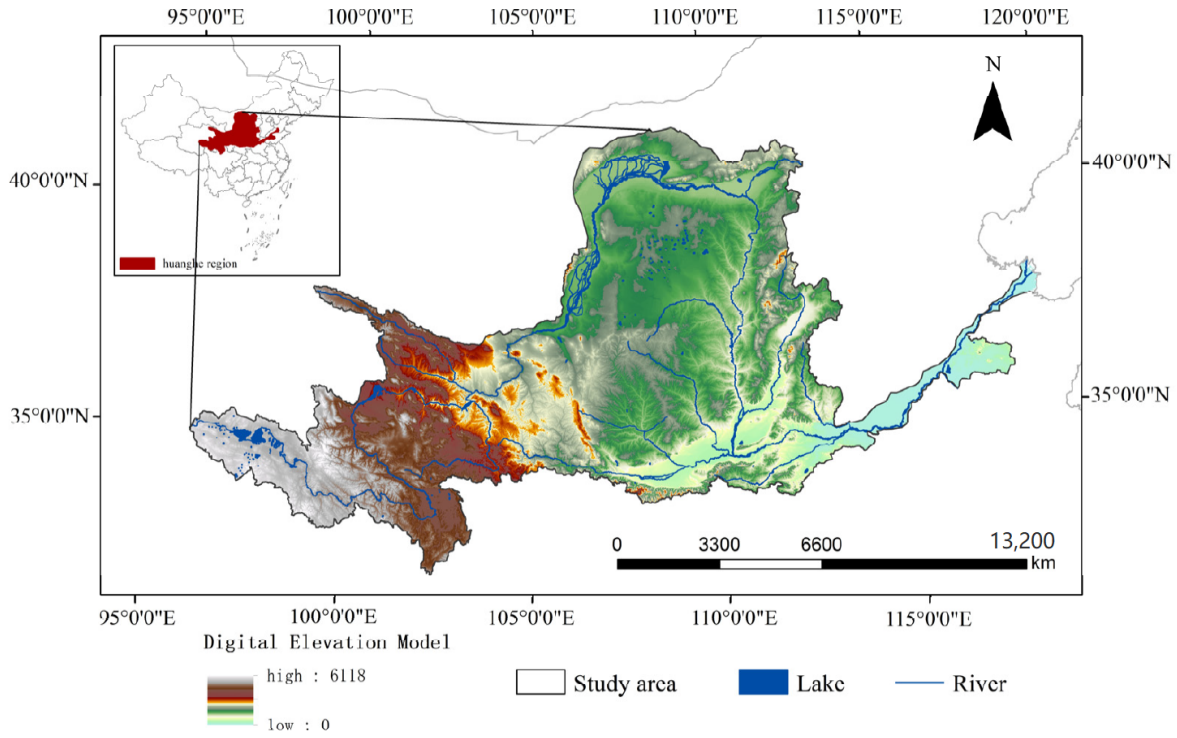

Figure 9. Geographical location of the study area.

4.2. Data Sources and Preprocessing

4.2.1. Remote Sensing Data

This study used high-resolution data from the multispectral imaging satellite Sentinel-2 and SAR satellite Sentinel-1. Comprising Sentinel-2A and Sentinel-2B, the Sentinel-2 system operates in a collaborative fashion, offering a 5-day revisit period. For the Yellow River Basin in 2021, surface reflectivity products (L2A) were sourced from the European Space Agency's (ESA) Copernicus Data Center (available at https://scihub.copernicus.eu/, (accessed on 12 November 2023) Date: from January to December 2021). The dataset, delivered in 12 monthly segments, encompasses a total of 1768 image captures. These images incorporate 12 spectral bands, featuring spatial resolutions of 10 m, 20 m, and 60 m. For compatibility with deep learning methodologies, bands of lower resolution were

resampled to a uniform 10 m, culminating in the integration of bands to produce optical image data with consistent resolution.

Sentinel-1, part of ESA's Copernicus Earth Observation Project, is a high-resolution SAR satellite featuring a six-day revisit period. It offers Level 1 Ground Range Detected (GRD) products across four modes: SM, IW, EW, and WV. This research focuses on high-resolution GRD data from the IW mode, boasting a 10-m resolution and supporting dual polarizations: vertical (VV) and cross-polarization (VH). Data was obtained through Google Earth Engine and synthesized via averaging techniques into monthly installments, which amounted to 12 editions for the year, readily usable for deep learning endeavors.

Furthermore, this study integrates auxiliary data for a comprehensive analysis. This includes ESA's 10 m surface coverage product [43] and the Joint Research Centre's (JRC) global water body data [44]. ESA product is instrumental in creating sample data, thereby refining the deep learning model's accuracy and reliability. Concurrently, JRC and ESA datasets facilitate the evaluation of the model's accuracy and a comparative study of the results.

4.2.2. Reference Samples

This study categorizes water bodies into four distinct types based on their morphology and distribution contexts: rivers, lakes, urban water bodies, and aquaculture, collectively referred to as complex water bodies. To enhance the accuracy of complex water bodies identification, a specialized sample library encompassing these diverse water body types was developed. Illustrative examples are provided (Figure 10), where "a" denotes lakes, "b" rivers, "c" aquaculture, and "d" urban water bodies. The sample labeling process, crucial for accurate water body identification, was executed in three strategic phases. Initially, conventional thresholding techniques were employed to derive preliminary water body outlines. Subsequently, integration of ESA land cover products and Digital Elevation Model (DEM) data [45] was undertaken to eliminate features potentially mistaken for water bodies, such as mountain shadows. The final phase involved generating a binary water body sample dataset through meticulous interpretation and accuracy labeling of concurrent spatio-temporal images sourced from Google Maps.

For model training, 26 images, each measuring 16,053 $\times$ 16,054 pixels, were chosen. These images are evenly distributed across the Yellow River Basin, encompassing all water body types identified in the small sample library. The images were segmented into 1024 $\times$ 1024 pixel slices with a 128-pixel overlap, followed by data augmentation on these segments. The slices were then randomly allocated into training, validation, and test sets at an 8:1:1 ratio. Throughout the training phase, the model was built using the training set, while the validation set facilitated hyperparameter tuning. Post-training, the model's generalization capabilities were assessed using the test set. This process resulted in 6231 training images, 1021 validation images, and an equal number of test images.

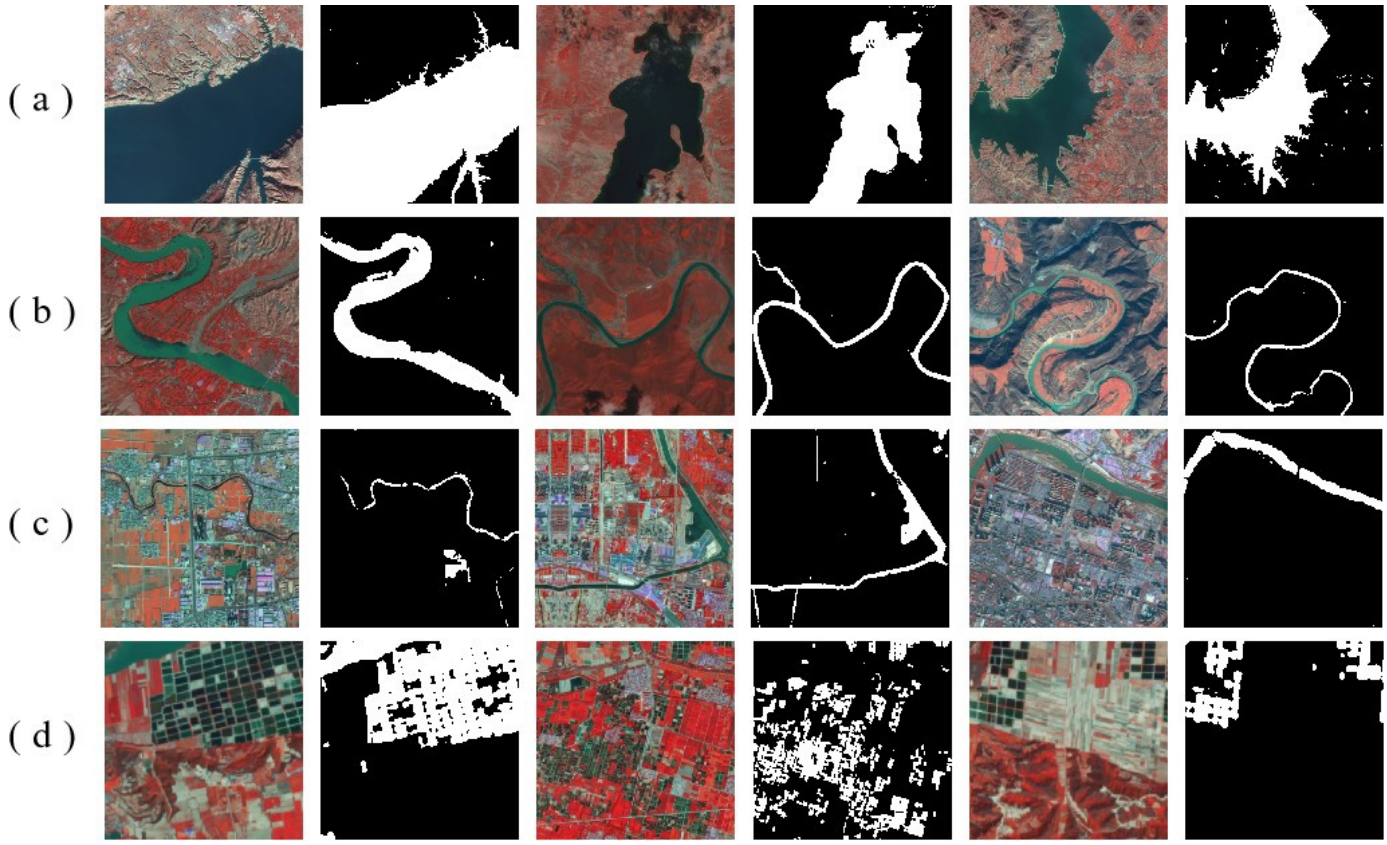

Figure 10. Sample repository mapping: (**a**) Lakes; (**b**) Rivers; (**c**) Urban water bodies; (**d**) Aquaculture. Presented in descending order.

4.3. Ablation Study

Ablation study dissects the performance contributions of various configurations within the Swin-UNet architecture as follows:

① Swin-Unet + Optical: This configuration employs U-shaped architecture of the Swin-UNet network, designated as "Swin-UNet-Opt". It exclusively utilizes optical image patches as input.

② Swin-Unet + Pixel-level fusion: Architecturally akin to "Swin-UNet-Opt", this variation, termed "Swin-UNet-Sum," also relies on U-shaped Swin-UNet framework. However, it diverges by inputting patches derived from a pixel-wise fusion of optical and SAR imagery.

③ SiamSwin-UNet + feature-level fusion: SiamSwin-UNet model is an evolution of Swin-UNet structure, characterized by a dual-branch design. These branches operate independently, maintaining distinct weights and input samples. The first branch processes optical image patches, while the second handles SAR image patches.

To validate the individual contributions of each component within DSST-UNet framework, this study conducted systematic ablation tests using the Yellow River Basin dataset. We established Method ① as the Baseline, utilizing only optical input. Method ② extends this by incorporating pixel-level fusion of multisource data, while Method ③ evolves the structure to a dual-branch SiamSwin-UNet, employing feature-level fusion. Subsequent iterations, labeled Methods ④ through ⑥, sequentially integrate DCB-ConvNet and FPA modules, either individually or combined, into the SiamSwin-UNet structure, thereby allowing an assessment of each configuration's impact on performance enhancement.

The ablation study results, detailed in Table 1, demonstrate that feature-level fusion of multi-source data outperforms pixel-level fusion in water bodies extraction, with the latter surpassing the use of solely optical data. Accordingly, feature-level fusion is adopted as the

standard for subsequent analysis. Employing just a simple feature-level fusion in SiamSwin-UNet yielded an F1 score of 91.62, indicating potential underutilization of features. The introduction of a DCB-ConvNet branch, which incorporates deformable convolutions in CNN operations, addressed this by enhancing local detail processing, resulting in a 2% F1 score increase to 93.58. Incorporation of a FPA module further improved the F1 score to 92.78, a 1% rise, and augmented Recall by 1%, diminishing detection omissions. This enhancement stems from FPA's ability to guide the network's focus towards salient high-level features, thereby mitigating detail loss. The concurrent application of both the DCB-ConvNet branch and the FPA module significantly bolstered extraction efficacy. The DCB-ConvNet branch fortified local feature learning, the deformable convolution diversified the adaptability to various water body types, and the FPA module minimized detail neglect. Collectively, this approach harmonized accuracy and Recall metrics for water bodies extraction, culminating in a 2% F1 score leap to 93.7.

Table 1. Ablation experiments for the network design.

Method	Metrics (%)			
	Precision	**Recall**	**IoU**	**F1**
Swin-Unet-Opt	91.42	90.34	83.27	90.87
Swin-Unet-Sum	92.89	89.60	83.85	91.22
SiamSwin-UNet	92.35	90.89	84.53	91.62
SiamSwin-UNet + DCB-ConvNet	95.07	92.13	87.94	93.58
SiamSwin-UNet + FPA	92.36	93.20	86.53	92.78
SiamSwin-UNet + DCB-ConvNet + FPA	94.82	92.79	88.31	93.79

In the illustrative outcomes depicted in Figure 11, the ablation findings are segmented into four categories. The initial two rows depict lakes with complex perimeters, followed by rivers, geometrically distinct aquaculture formations, and finally, urban water bodies.

The first row contains both expansive and diminutive lakes characterized by their intricate borders. These water bodies, set against a relatively non-complex topographical backdrop, are distinguished by their pronounced boundaries, geometric regularity, and contrast relative to adjacent areas, making them perceptibly distinct in color and texture. This result suggests that Transformer's self-attention mechanism is proficient in capturing the image's global context, thereby effectively segregating lakes from other elements. Nonetheless, when dealing with smaller lakes exhibiting complex contours, SiamSwin-UNet framework tends to generate numerous incorrect identifications. This limitation arises partly because the Transformer prioritizes overarching information, often at the expense of localized details, and partly due to the absence of multi-scale learning in processing high-level feature maps. The integration of both DCB-ConvNet and FPA modules markedly diminishes instances of overlooked lake boundaries, with the former enhancing the extraction of potent local features and the latter bolstering the comprehension of extensive, complex lake perimeters through multi-scale feature assimilation. However, while the FPA module enriches the model's capacity to discern complex lake borders, it concurrently elevates the rate of false detections even as it curtails instances of overlooked detections.

The second row encompasses extensive rivers and complex river networks, positioned in terrains akin to the first group but distinguished by their sinuous boundaries, complicating water bodies extraction processes. It is evident, as shown in the third row, that the baseline framework frequently fails to accurately identify segments of the dense river network, indicating a deficiency in the network's stability concerning river delineation tasks. The incorporation of either a DCB-ConvNet branch or a FPA module significantly mitigates instances of water bodies misidentification. Ultimately, the amalgamation of both the DCB-ConvNet branch and the FPA module results in river delineations that more accurately reflect actual labels, although certain inaccuracies persist.

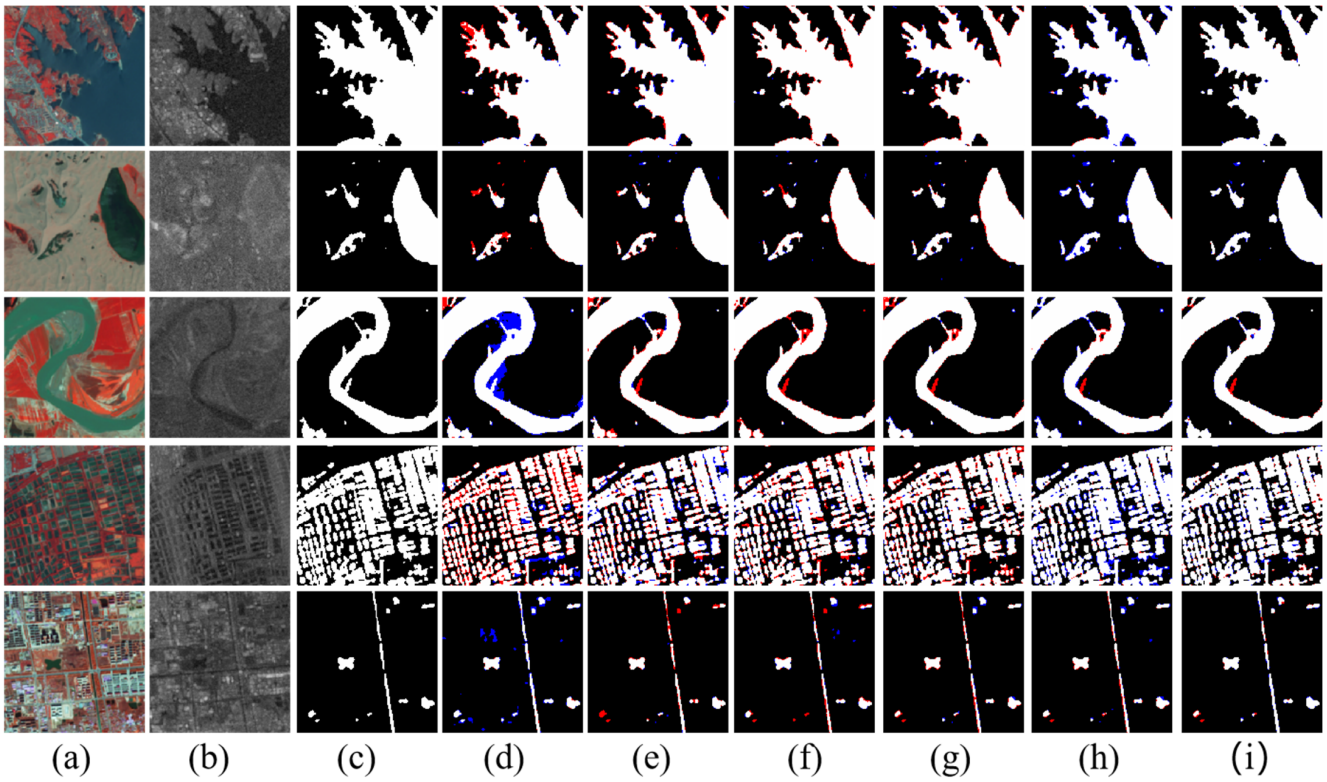

Figure 11. Visual assessment of ablation studies in water bodies extraction: White indicates correct detections (where ground truth is water bodies, and detection successful), red indicates missed detections (where ground truth is water bodies, but detection fails) and blue indicates false detections (where ground truth is not water bodies, but is incorrectly detected as such). From left to right: (**a**) Opt image; (**b**) SAR image; (**c**) Ground truth; (**d**) Swin-UNet-Opt; (**e**) Swin-UNet-Sum; (**f**) SiamSwin-UNet; (**g**) SiamSwin-UNet + DCB-ConvNet; (**h**) SiamSwin-UNet + FPA; and (**i**) SiamSwin-UNet + DCB-ConvNet + FPA.

The fourth row illustrates block-distributed aquaculture interspersed among cropland, leading the baseline network to misclassify certain agricultural plots as aquacultural zones. The introduction of the DCB-ConvNet branch and FPA module enhances aquaculture delineation compared to the baseline, but the intricate and fragmented nature of aquaculture distribution results in persistent inaccuracies and omissions in boundary identification against actual labels.

The fifth row displays urban water bodies environments, characterized by diverse forms including linear channels, slender streams, and block-like ponds. The baseline network, constrained by its multimodal feature fusion capabilities, exhibits limited efficacy in urban water body delineation. The DCB-ConvNet branch's convolution operations and FPA module's strategy for multi-scale learning on high-level feature maps contribute to enhanced accuracy in urban water bodies extraction. Nonetheless, a degree of misdetection persists, indicating a need for continued exploration in the realm of urban water bodies extraction methodologies.

4.4. Comparing Methods

To validate the efficacy of the proposed method, both qualitative and quantitative analyses were conducted against contemporary cutting-edge techniques using the Yellow River Basin water bodies dataset. The methodologies for comparison encompass FCN [13], PSPNet [46], PVT (Pyramid Vision Transformer) [47], and TransUnet [48]. The first two comparative methods are based on traditional CNN, and the latter two are based on the Transformer architecture. Specifically, FCN uses an inverse convolutional layer to

up-sample the feature map to achieve pixel-level classification of the image. PSPNet contains a pyramid pooling module that integrates global contextual information. PVT introduces progressively shrinking pyramids while employing finer-grained image chunks as inputs and always has a global receptive field. TransUnet employs a hybrid encoder of a CNN and a Transformer of hybrid encoders, where Transformer encodes labeled image chunks from Convolutional Neural Network (CNN) feature maps as an input sequence for extracting global context. In this case, TransUnet uses the serial structure of Transformer and CNN in the encoding stage, while our DSST-UNet uses the parallel mode. In addition, SegFormer [49] stands out for its integration of a decoder with a full MLP structure. This design choice allows SegFormer to have a larger and more efficient receptive field compared to regular CNN models, while still maintaining a lightweight architecture. Furthermore, our study incorporates FAN [50], which introduces a novel attentional channel design to the Vision Transformer (ViT) architecture. By dynamically reweighting channel selection based on content, FAN enhances the network's feature extraction capabilities. In our comparison, we utilize SegFormer with FAN Backbone as a comparative method, referred to as SegFormer-FAN. All of the above comparison methods use both optical and SAR images as inputs. Our comparison results are shown in Table 2. From Table 2, we can see that our method has obvious advantages in all indicators and compared with the highest accuracy of 92.64% of mainstream methods, our method can improve the accuracy by about 1%.

Table 2. Quantitative comparison of the proposed method with mainstream method.

Method	Metrics (%)			
	Precision	Recall	IoU	F1
FCN	87.34	90.29	79.84	88.79
PSPNet	90.54	88.74	81.21	89.63
PVT	92.35	90.89	84.53	91.62
TransUnet	93.06	92.84	86.83	92.95
SegFormer-FAN	93.53	93.75	88.04	93.64
Ours (DSST-UNet)	94.82	92.79	88.31	93.79

In the comparative results depicted in Figure 12, the analyses are segregated into four distinct categories. The initial row showcases large lakes with complex perimeters, followed by the second and third rows that illustrate riverine landscapes. The fourth row depicts aquaculture organized in block formations, while the fifth and final section focuses on urban water features.

The first row is a frozen lake, and the extraction results of different methods are somewhat misdetected and underdetected. Since FCN is only a simple convolution operation, it has a large number of missed detections on the frozen lake. The Transformer-based network architecture performs better in the face of this problem. Our method can adaptively adjust the receptive field size for local feature enhancement learning, so the extraction results are closer to the labels than these comparison methods. Farmland exists beside the fine rivers in the second row, and the fine tributaries in the large rivers in the third row have complex features. The pooling operation in PSPNet omits the local information of the image, which leads to the breakage of the extraction results of the fine water bodies. Compared with the results of the rivers extracted by other comparative methods, the extraction results of our method are more continuous.

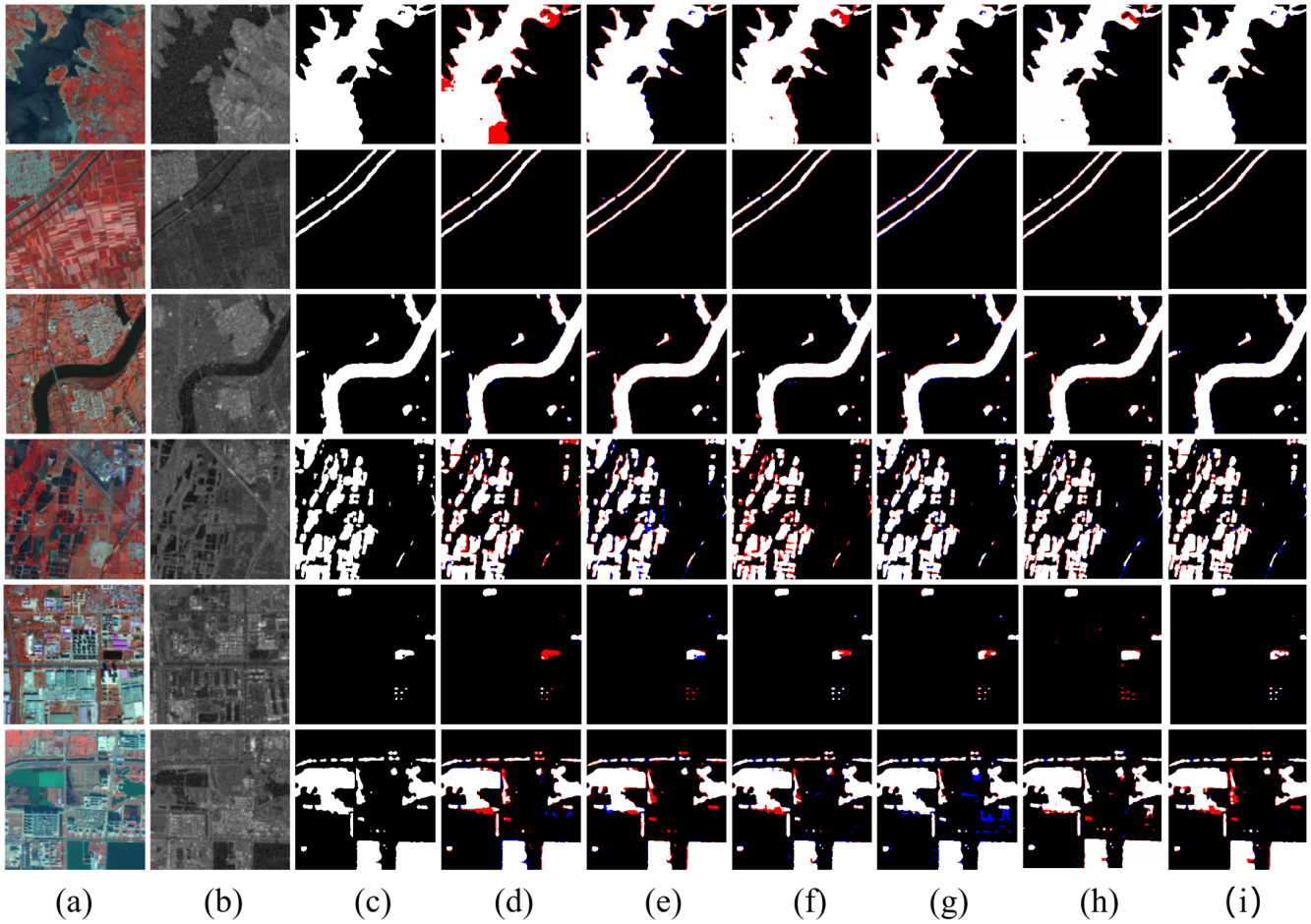

Figure 12. Comparative analysis of water bodies extraction approach: White indicates correct detections (where ground truth is water bodies, and detection successful), red indicates missed detections (where ground truth is water bodies, but detection fails) and blue indicates false detection (where ground truth is not water bodies, but is incorrectly detected as such). From left to right: (**a**) Opt image; (**b**) SAR image; (**c**) Ground truth; (**d**) FCN; (**e**) PSPNet; (**f**) PVT; (**g**)TransUnet; and (**h**) SegFormer-FAN; (**i**) Proposed method in this study.

The fourth row is the block distribution of aquaculture, aquaculture due to the influence of the terrain, in a fragmented and scattered distribution. Moreover, there are easily confused croplands between aquaculture, which makes classification more difficult. Overall, the network based on the Transformer architecture performs better than the CNN, thanks to the fact that the Transformer is good at capturing global information. For aquaculture, the latter performs better because PVT divides tokens under the role of pyramid, SegFormer-FAN is a pure transformer encoder, while TransUnet first performs convolutional operation and encodes the context through the Transformer. Additionally, our method performs parallel coding of CNN and the Transformer at each stage, so the extraction is the best. In the final row, focusing on linear ditches, established methods tend to produce fragmented results, interrupting the continuity of flow, whereas the proposed approach ensures more seamless extraction outcomes. Moreover, for landscape water bodies encompassing complex peripheries, the novel technique outperforms its counterparts, achieving heightened accuracy in boundary extraction.

5. Discussion

5.1. About the Model

The efficacy of extracting water bodies is intrinsically linked to the quality of feature information fed into the network. Historically, both optical and SAR imagery have been leveraged for this task. Nonetheless, SAR imagery often falls prey to noise interference, hampering the clarity of water body delineations. Concurrently, optical imagery acquisition becomes problematic during cloudy conditions, and the extraction process is often confounded by the 'same spectrum, different objects' dilemma. One strategy for amalgamating optical and SAR images involves their concatenation within the channel dimension. While this methodology endows the network with a richer feature set, the simplistic nature of pixel-level fusion inadvertently ushers in substantial noise during feature extraction, given the unselective sharing of weight information. In contrast, feature-level fusion operates on a multimodal platform, treating different satellite data individually based on their inherent properties. This approach involves a comparative analysis of labels to discern the contribution of each pixel to varying branches, thereby optimizing parameters for a more effective water bodies extraction process.

The efficacy of water bodies extraction is intricately tied to the network architecture employed. This study utilizes a hybrid network, drawing upon the strengths of both Transformer and CNNs, structured around the U-Net framework for enhanced multimodal water bodies extraction. During the single-modal feature encoding phase, an innovative connection is forged between the DCB-ConvNet and Transformer branches through systematic feature linkage in the channel dimension at each level. This structural confluence facilitates a robust extraction of global semantic information, augmented by detailed local features. An advanced D-CBAM module is integrated to steer the DCB-ConvNet branch, emulating human perceptual capabilities by automatically modulating the receptive field size in response to the diverse morphologies of complex water bodies, thereby bolstering local feature amplification. Within the Transformer branch, the self-attention mechanism's inclusion significantly mitigates the limitations traditionally associated with convolutional operations' ability to model long-range dependencies. This strategic amalgamation effectively addresses the global context insufficiency and extended semantic interactions typically lacking in CNNs. Furthermore, the introduction of a FPA module enhances the network's proficiency in discerning water bodies across various scales. Feature-level fusion is meticulously achieved by concatenating the features distilled from both the Transformer and DCB-ConvNet branches within the channel dimension. Progressing to the decoding stage, the study employs U-Net architecture's skip connection concept, leveraging high-level feature insights to navigate lower-level feature synthesis, culminating in a more nuanced extraction of edge detail features.

5.2. Computational Efficiency

In order to comprehensively evaluate the complexity and performance of a deep learning model, it is customary to assess several key metrics, including model parameters, computational complexity, and inference time (FPS). FLOPs, which stands for Floating-Point Operations per Second, is primarily utilized to measure the computational speed and efficiency of the model. FPS refers to the number of image frames processed per second and is mainly used to gauge the real-time performance of the model. The number of parameters denotes the total count of trainable parameters in the model and serves as an indicator of its complexity and capacity. From Table 3, we can see that our model possesses the second highest number of parameters, trailing only TransUnet. This is attributed to the dual-stream structure and retention of two decoders. However, we do incur higher overhead in terms of inference time compared to other methods. On the other hand, SegFormer-FAN demonstrates superiority in balancing accuracy and efficiency, potentially due to its employment of lightweight ViT blocks and a simpler decoding structure. Therefore, future directions for improvement should involve exploring new backbone networks that

can reduce time and memory costs, while still retaining the dual-stream structure. This approach will likely further enhance the overall performance and efficiency of our model.

Table 3. Comparison of computational efficiency of different model.

Model	FLOPs (G)	Params (M)	FPS
FCN	55.26	23.85	35.63
PSPNet	62.83	48.90	28.72
PVT	82.35	63.75	29.83
TransUnet	105.30	103.23	16.64
SegFormer-FAN	67.63	58.92	19.48
Ours	92.85	83.26	8.65

5.3. Comparison Analysis of Water Body Classifications

The network architecture under discussion was applied to map diverse water bodies in the Yellow River Basin in 2021, using the WIF method. This mapping differentiated water bodies into non-water bodies, seasonal water bodies, and permanent water bodies, detailed in Figure 13. For an inclusive evaluation of these extensive mapping outcomes, products from ESA [43] and JRC [44] were brought into comparison. JRC product classifies observations into four types: invalid, non-water, seasonal water bodies, and permanent water bodies. In this analysis, the 'effective water bodies' term was coined to collectively describe both seasonal and permanent water bodies within the JRC framework. In contrast, ESA product presents 11 feature categories. This study narrows its scope to the water body category, aligning the comparison to concentrate solely on 'effective water bodies' for all referenced products.

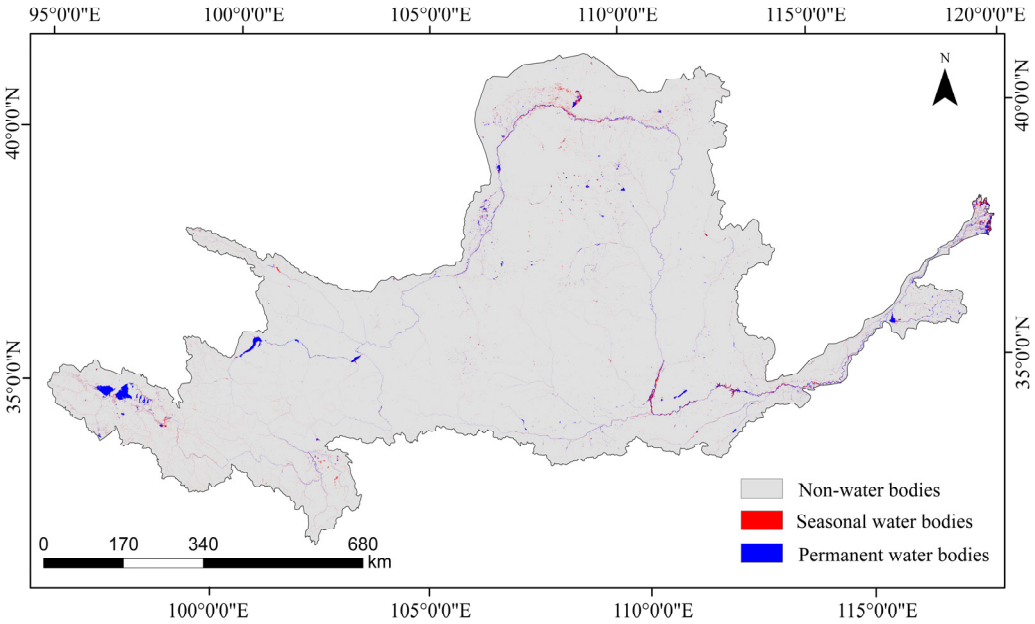

Figure 13. Mapping of 2021 water body distribution within Yellow River Basin.

This research undertook a comparative analysis using 2021 water body data from the JRC and the 2021 land-cover classification from ESA, focusing on case areas in Urad Front Banner in Inner Mongolia Autonomous Region, Maqu County, Zhongning County, and Dongping County. The discrepancies in the water body areas within the Urad Front Banner, as depicted in JRC's versus this study's findings and ESA's datasets, are noticeable (Figure 14). One plausible explanation for this variance could be the coarser resolution of JRC dataset, potentially leading to the erroneous classification of barren land as water bodies. However, apart from these deviations, a substantial congruence is observed among

the datasets. They all adeptly delineate the spatial distribution of water bodies in the Yellow River Basin, though this study indicates a higher number of water bodies compared to the figures in JRC and ESA datasets.

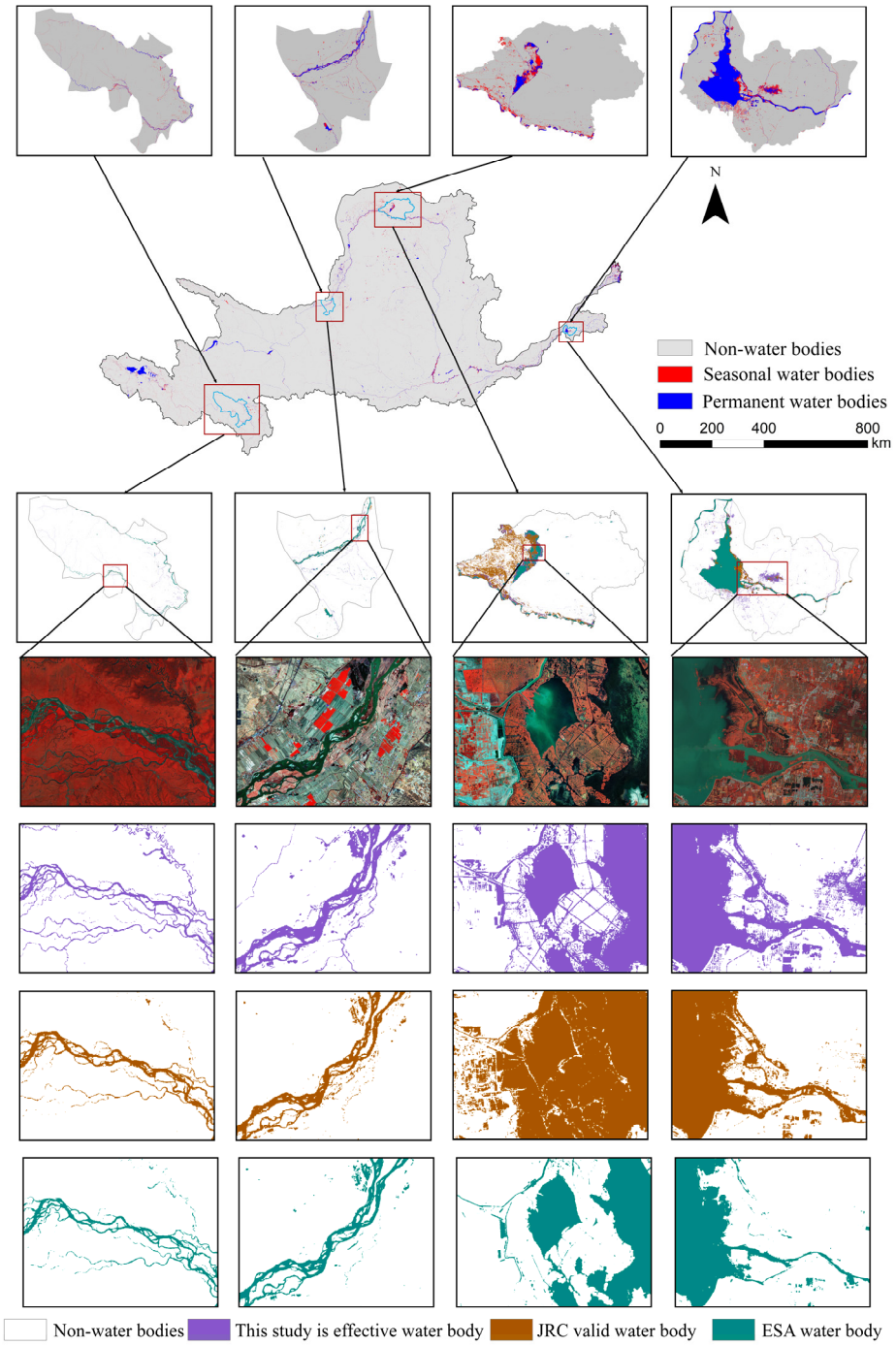

Figure 14. Comparative analysis of different water body detections in various products.

In the four case areas of Urad Front Banner, Maqu County, Zhongning County, and Dongping County, this study demonstrates superior accuracy in detecting both extensive and smaller water bodies when compared to JRC and ESA datasets. This is particularly evident in Maqu County and Zhongning County, regions with a dense river network comprising Yellow River's main course and numerous smaller tributaries. JRC's data, with a spatial resolution of 30 m, fails to capture many smaller water bodies. Furthermore, in areas like Urad Front Banner and Dongping County, known for Wuliangsu Lake and

Dongping Lake respectively, ESA's datasets offer static results, unable to account for seasonal water bodies and thereby omitting them. This research stands out by considering the seasonal dynamics of water bodies and employing higher-resolution data, which contribute to the enhanced accuracy in water body identification.

6. Conclusions

This research presents an end-to-end deep learning framework designed for the nuanced extraction of water bodies, employing a synergistic approach through the amalgamation of Transformer and CNN methodologies. This innovative model is meticulously configured to concurrently assimilate fine-grained local and comprehensive global information present in feature maps. One of the pivotal enhancements implemented within this framework is the refinement of CBAM in DCB-ConvNet segment. This is further complemented by the incorporation of deformable convolution techniques, significantly sharpening the system's acuity in identifying water body perimeters and augmenting its capability in detailed local information capture. Parallelly, the Transformer segment exploits its inherent self-attention mechanism, offering a solution to the inherent constraints of traditional convolution procedures, which exhibit deficiencies in managing extensive spatial dependencies. This strategic enhancement fortifies the model's capacity for comprehensive feature learning, thereby boosting the uniformity and reliability of outcomes in the context of water bodies extractions. Addressing the potential contamination of high-level features with irrelevant background constituents, the model integrates a FPA module. This critical inclusion serves to meticulously filter out non-water bodies elements, intensifying the focus on multi-scale water-related information, and substantially improving the accuracy of detecting smaller-scale water bodies. The model's accuracy and adaptability, as demonstrated by 93.7% F1 score on the challenging Yellow River Basin dataset, are further validated through rigorous ablation testing and comparative assessments, establishing its suitability for complex, large-scale applications. Its practical efficacy is additionally evidenced by successful, extensive mapping exercises conducted within the Yellow River Basin. Future initiatives aim to expand the utility of this cutting-edge model by undertaking comprehensive assessments of temporal fluctuations in the Yellow River Basin's water bodies over a triennial period, with a concentrated focus on discerning patterns within both seasonal and permanent water bodies. Concurrently, there is a commitment to enhancing the operational efficiency of the network, striving for a leaner configuration that retains, or potentially elevates, its analytical prowess.

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

Funding: This research was funded by The National Key Research and Development Program of China (No. 2021YFB3901300).

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to currently proprietary.

Acknowledgments: The authors are grateful to the editors and anonymous reviewers for their informative suggestions.

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

Abbreviations

The following abbreviations are used throughout this manuscript:

ASPP	atrous spatial pyramid pooling
BCE	binary cross entropy
CBAM	convolutional block attention module
CNN	convolutional neural network

CV	computer vision
DCB	dense convolutional block
D-CBAM	deformable convolutional block attention module
DCNN	deep convolutional neural networks
DEM	digital elevation model
EASPP	enhanced atrous spatial pyramid pooling
ESA	european space agency
FCN	fully convolutional network
FN	false negative
FP	false positive
FPA	feature pyramid attention
GRD	ground range detected
IoU	intersection over union
JRC	joint research centre
PAN	pyramid attention network
SAR	synthetic aperture radar
SWA	surface water area
TP	true positive
VHR	very high resolution
WIF	water inundation frequency

References

1. Yu, Y.; Huang, L.; Lu, W.; Guan, H.; Ma, L.; Jin, S.; Yu, C.; Zhang, Y.; Tang, P.; Liu, Z. WaterHRNet: A multibranch hierarchical attentive network for water body extraction with remote sensing images. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *115*, 103103. [CrossRef]
2. Amprako, J.L. The United Nations World Water Development Report 2015. *Future Food J. Food Agric. Soc.* **2016**, *4*, 64–65.
3. Mueller, N.; Lewis, A.; Roberts, D.; Ring, S.; Melrose, R.; Sixsmith, J.; Lymburner, L.; McIntyre, A.; Tan, P.; Curnow, S. Water observations from space: Mapping surface water from 25 years of Landsat imagery across Australia. *Remote Sens. Environ.* **2016**, *174*, 341–352. [CrossRef]
4. Hu, Q.; Li, C.; Wang, Z.; Liu, Y.; Liu, W. Continuous Monitoring of the Surface Water Area in the Yellow River Basin during 1986–2019 Using Available Landsat Imagery and the Google Earth Engine. *ISPRS Int. J. Geo-Inf.* **2022**, *11*, 305. [CrossRef]
5. Elhag, M.; Psilovikos, A.; Sakellariou-Makrantonaki, M. Land use changes and its impacts on water resources in Nile Delta region using remote sensing techniques. *Environ. Dev. Sustain.* **2013**, *15*, 1189–1204. [CrossRef]
6. Tang, P.; Liang, Q.; Yan, X.; Xiang, S.; Zhang, D. GP-CNN-DTEL: Global-part CNN model with data-transformed ensemble learning for skin lesion classification. *IEEE J. Biomed. Health Inform.* **2020**, *24*, 2870–2882. [CrossRef] [PubMed]
7. Fang, F.; Li, L.; Zhu, H.; Lim, J.-H. Combining faster R-CNN and model-driven clustering for elongated object detection. *IEEE Trans. Image Process.* **2019**, *29*, 2052–2065. [CrossRef]
8. Li, Y.; Dang, B.; Zhang, Y.; Du, Z. Water body classification from high-resolution optical remote sensing imagery: Achievements and perspectives. *ISPRS J. Photogramm. Remote Sens.* **2022**, *187*, 306–327. [CrossRef]
9. Lu, M.; Fang, L.; Li, M.; Zhang, B.; Zhang, Y.; Ghamisi, P. NFANet: A novel method for weakly supervised water extraction from high-resolution remote-sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5617114. [CrossRef]
10. Dong, Z.; Wang, G.; Amankwah, S.O.Y.; Wei, X.; Hu, Y.; Feng, A. Monitoring the summer flooding in the Poyang Lake area of China in 2020 based on Sentinel-1 data and multiple convolutional neural networks. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *102*, 102400. [CrossRef]
11. Zhao, X.; Wang, H.; Liu, L.; Zhang, Y.; Liu, J.; Qu, T.; Tian, H.; Lu, Y. A Method for Extracting Lake Water Using ViTenc-UNet: Taking Typical Lakes on the Qinghai-Tibet Plateau as Examples. *Remote Sens.* **2023**, *15*, 4047. [CrossRef]
12. Zhang, Y.; Lu, H.; Ma, G.; Zhao, H.; Xie, D.; Geng, S.; Tian, W.; Sian, K.T.C.L.K. MU-Net: Embedding MixFormer into Unet to Extract Water Bodies from Remote Sensing Images. *Remote Sens.* **2023**, *15*, 3559. [CrossRef]
13. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; pp. 3431–3440.
14. Pan, J.; Wei, Z.; Zhao, Y.; Zhou, Y.; Lin, X.; Zhang, W.; Tang, C. Enhanced FCN for farmland extraction from remote sensing image. *Multimed. Tools Appl.* **2022**, *81*, 38123–38150. [CrossRef]
15. Li, L.; Yan, Z.; Shen, Q.; Cheng, G.; Gao, L.; Zhang, B. Water body extraction from very high spatial resolution remote sensing data based on fully convolutional networks. *Remote Sens.* **2019**, *11*, 1162. [CrossRef]
16. Zhong, H.-F.; Sun, H.-M.; Han, D.-N.; Li, Z.-H.; Jia, R.-S. Lake water body extraction of optical remote sensing images based on semantic segmentation. *Appl. Intell.* **2022**, *52*, 17974–17989. [CrossRef]

17. Sunandini, G.; Sivanpillai, R.; Sowmya, V.; Variyar, V.S. Significance of Atrous Spatial Pyramid Pooling (ASPP) in Deeplabv3+ for Water Body Segmentation. In Proceedings of the 2023 10th International Conference on Signal Processing and Integrated Networks (SPIN), Noida, India, 23–24 March 2023; pp. 744–749.
18. Yuan, K.; Zhuang, X.; Schaefer, G.; Feng, J.; Guan, L.; Fang, H. Deep-learning-based multispectral satellite image segmentation for water body detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 7422–7434. [CrossRef]
19. He, X.; Zhou, Y.; Zhao, J.; Zhang, D.; Yao, R.; Xue, Y. Swin transformer embedding UNet for remote sensing image semantic segmentation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4408715. [CrossRef]
20. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
21. Guo, J.; Han, K.; Wu, H.; Tang, Y.; Chen, X.; Wang, Y.; Xu, C. Cmt: Convolutional neural networks meet vision transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 12175–12185.
22. Konapala, G.; Kumar, S.V.; Ahmad, S.K. Exploring Sentinel-1 and Sentinel-2 diversity for flood inundation mapping using deep learning. *ISPRS J. Photogramm. Remote Sens.* **2021**, *180*, 163–173. [CrossRef]
23. He, X.; Zhang, S.; Xue, B.; Zhao, T.; Wu, T. Cross-modal change detection flood extraction based on convolutional neural network. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *117*, 103197. [CrossRef]
24. Li, X.; Zhang, G.; Cui, H.; Hou, S.; Wang, S.; Li, X.; Chen, Y.; Li, Z.; Zhang, L. MCANet: A joint semantic segmentation framework of optical and SAR images for land use classification. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *106*, 102638. [CrossRef]
25. Bioresita, F.; Puissant, A.; Stumpf, A.; Malet, J.-P. Fusion of Sentinel-1 and Sentinel-2 image time series for permanent and temporary surface water mapping. *Int. J. Remote Sens.* **2019**, *40*, 9026–9049. [CrossRef]
26. Lyu, X.; Jiang, W.; Li, X.; Fang, Y.; Xu, Z.; Wang, X. MSAFNet: Multiscale Successive Attention Fusion Network for Water Body Extraction of Remote Sensing Images. *Remote Sens.* **2023**, *15*, 3121. [CrossRef]
27. Lyu, X.; Fang, Y.; Tong, B.; Li, X.; Zeng, T. Multiscale Normalization Attention Network for Water Body Extraction from Remote Sensing Imagery. *Remote Sens.* **2022**, *14*, 4983. [CrossRef]
28. Liu, Z.; Chen, X.; Zhou, S.; Yu, H.; Guo, J.; Liu, Y. DUPnet: Water Body Segmentation with Dense Block and Multi-Scale Spatial Pyramid Pooling for Remote Sensing Images. *Remote Sens.* **2022**, *14*, 5567. [CrossRef]
29. Weng, L.; Pang, K.; Xia, M.; Lin, H.; Qian, M.; Zhu, C. Sgformer: A local and global features coupling network for semantic segmentation of land cover. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 6812–6824. [CrossRef]
30. Zhang, C.; Jiang, W.; Zhang, Y.; Wang, W.; Zhao, Q.; Wang, C. Transformer and CNN hybrid deep neural network for semantic segmentation of very-high-resolution remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4408820. [CrossRef]
31. Peng, Y.; He, J.; Yuan, Q.; Wang, S.; Chu, X.; Zhang, L. Automated glacier extraction using a Transformer based deep learning approach from multi-sensor remote sensing imagery. *ISPRS J. Photogramm. Remote Sens.* **2023**, *202*, 303–313. [CrossRef]
32. Xiao, T.; Liu, Y.; Huang, Y.; Li, M.; Yang, G. Enhancing Multiscale Representations with Transformer for Remote Sensing Image Semantic Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5605116. [CrossRef]
33. Li, H.; Xiong, P.; An, J.; Wang, L. Pyramid attention network for semantic segmentation. *arXiv* **2018**, arXiv:1805.10180.
34. Luo, X.; Tong, X.; Hu, Z. An applicable and automatic method for earth surface water mapping based on multispectral images. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *103*, 102472. [CrossRef]
35. Li, X.; Lei, L.; Sun, Y.; Li, M.; Kuang, G. Multimodal bilinear fusion network with second-order attention-based channel selection for land cover classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 1011–1026. [CrossRef]
36. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
37. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 764–773.
38. Good, I.J. Rational decisions. *J. R. Stat. Soc. Ser. B (Methodol.)* **1952**, *14*, 107–114. [CrossRef]
39. Milletari, F.; Navab, N.; Ahmadi, S.-A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; pp. 565–571.
40. Zhang, Y.; Du, J.; Guo, L.; Fang, S.; Zhang, J.; Sun, B.; Mao, J.; Sheng, Z.; Li, L. Long-term detection and spatiotemporal variation analysis of open-surface water bodies in the Yellow River Basin from 1986 to 2020. *Sci. Total Environ.* **2022**, *845*, 157152. [CrossRef] [PubMed]
41. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv* **2017**, arXiv:1711.05101.
42. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 448–456.
43. Zanaga, D.; Van De Kerchove, R.; Daems, D.; De Keersmaecker, W.; Brockmann, C.; Kirches, G.; Wevers, J.; Cartus, O.; Santoro, M.; Fritz, S. ESA WorldCover 10 m 2021 v200. 2022. Available online: https://zenodo.org/records/7254221 (accessed on 22 December 2023).
44. Pekel, J.-F.; Cottam, A.; Gorelick, N.; Belward, A.S. High-resolution mapping of global surface water and its long-term changes. *Nature* **2016**, *540*, 418–422. [CrossRef]
45. Yang, L.; Meng, X.; Zhang, X. SRTM DEM and its application advances. *Int. J. Remote Sens.* **2011**, *32*, 3875–3896. [CrossRef]

46. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
47. Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 568–578.
48. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv* **2021**, arXiv:2102.04306.
49. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12077–12090.
50. Zhou, D.; Yu, Z.; Xie, E.; Xiao, C.; Anandkumar, A.; Feng, J.; Alvarez, J.M. Understanding the robustness in vision transformers. In Proceedings of the International Conference on Machine Learning, Baltimore, MD, USA, 17–23 July 2022; pp. 27378–27394.