A Lightweight Network With Embedded Soft Constraints on Approximate Spectral Features for Real-Time Water Body Segmentation in Remote Sensing Images

Qingqing Cao, Boya Zhao, Zijin Li, Fangfang Zhang, and Yuanfeng Wu, *Senior Member, IEEE*

***Abstract*—Real-time extraction of water bodies from remote sensing images captured by unmanned aerial vehicles (UAVs) or satellites is challenging due to the difficulty in performing precise atmospheric correction and other preprocessing steps. In this article, we propose a lightweight network with embedded soft constraints on approximate spectral features for real-time water body segmentation in remote sensing images. First, we introduce approximate spectral feature indices as auxiliary interband feature information. A lightweight pseudo-siamese feature extraction network (LPSE) is designed to separately extract features from visible bands and the approximate spectral indices. Second, we develop an approximate spectral feature soft constraint fusion mechanism (ASFC) that utilizes spatial attention to selectively fuse effective target features from the visible bands and approximate spectral indices. Third, we incorporate an atrous spatial pyramid pooling module for edge feature enhancement within a self-distillation edge-aware lightweight decoder. Finally, the proposed network is accelerated and quantized using TensorRT and deployed on the embedded device Jetson Orin NX. Experimental results show that the model achieves an intersection over union accuracy of 70.74% on the FloodNet dataset and 92.26% on the GF-FloodNet dataset. With only 0.22 million parameters and a computational cost of 0.32 GFLOPs, the inference time per image is 6.45 ms. The proposed method demonstrates significant advantages in segmentation accuracy and computational efficiency, making it highly promising for real-time water body segmentation on edge computing platforms, such as UAVs or satellites.**

***Index Terms*—Lightweight, real-time, self-distillation, soft constraints on approximate spectral features, water body segmentation.**

Received 24 January 2025; revised 1 March 2025; accepted 13 March 2025. Date of publication 19 March 2025; date of current version 29 April 2025. This work was supported in part by the National Key R&D Program of China under Grant 2021YFA0715203, and in part by the Science and Disruptive Technology Program, AIRCAS under Grant 2024-AIRCAS-SDTP-12. *(Corresponding author: Yuanfeng Wu.)*

Qingqing Cao and Zijin Li are with the Key Laboratory of Computational Optical Imaging Technology, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China, and also with the School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: caoqingqing22@mails.ucas.ac.cn; lizijin21@mails.ucas.ac.cn).

Boya Zhao and Yuanfeng Wu are with the Key Laboratory of Computational Optical Imaging Technology, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China (e-mail: zhaoby@aircas.ac.cn; wuyf@aircas.ac.cn).

Fangfang Zhang is with the Key Laboratory of Digital Earth Science, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China.

Digital Object Identifier 10.1109/JSTARS.2025.3552783

I. Introduction

With the advancement of remote sensing techniques for Earth observation missions, the timely extraction of information from remote sensing images captured by unmanned aerial vehicles (UAVs) or satellites is challenging, especially in response to sudden Earth surface anomalies, such as flood disasters, watershed environmental pollution incidents, earthquake-induced barrier lakes, landslides and debris flows, and forest and grassland fires [1]. In the wake of such events, the traditional process—where users passively request Earth observation data, the ground control center uploads observation instructions, and raw data are transmitted back to users for processing and analysis after imaging—fails to meet the stringent timeliness requirements of these applications. Real-time processing of sensor-acquired remote sensing images, which involves performing preprocessing and extracting target information on edge computing platforms aboard UAVs and satellites, represents a novel edge computing scenario. This approach enables the instantaneous conversion of remote sensing images into effective thematic information immediately after sensor imaging, transforming the "Big Data" of raw remote sensing images into the "small data" of specific targets. Such real-time processing significantly reduces data transmission burdens and provides immediate remote sensing information to support decision-making in time-sensitive application scenarios [2]. For example, during flood disasters, real-time extraction of water body distribution facilitates the spatiotemporal characterization of inundated areas, which is crucial for promptly understanding flood dynamics and enhancing emergency response capabilities.

This article investigates real-time methods for extracting surface water bodies. In remote sensing images, water bodies typically represent weak information sources; the signals received by sensors aboard UAVs and satellites are mainly composed of atmospheric reflection, water surface reflection, and water body radiation, with the latter accounting for only about 10% of the total signal. Moreover, the morphological characteristics of surface water bodies are complex and highly variable. The spectral characteristics of water bodies vary across regions and types [3] due to differing constituents, such as silt, phytoplankton, and suspended sediments, present in various local areas within large scenes, leading to spectral variability. In addition, the irregularity and complexity of water body boundaries further increase the difficulty of semantic segmentation tasks in remote sensing images [4]. These boundaries are often not clear lines but transitional zones, making the accurate extraction of water body boundaries during segmentation even more challenging.

In remote sensing imagery, different surface features exhibit distinct reflection and absorption characteristics across various electromagnetic wavelengths, forming unique spectral signatures [5]. Water indices are commonly employed to extract water regions from such images [6], including the normalized difference water index (NDWI) [7], the modified normalized difference water index (MNDWI), and the automated water extraction index (AWEI) [8]. While these band math methods are simple and effective, applying them to real-time onboard processing tasks on UAVs or satellites presents significant limitations. First, they rely on accurate calculations of water spectral reflectance; however, atmospheric correction for remote sensing images is highly complex and requires the simultaneous acquisition of various atmospheric parameters. Second, threshold selection lacks universality; different thresholds must be adjusted when binarizing water bodies across different scenes.

Deep learning methods have been applied to the processing of remote sensing images [9] [10]. Semantic segmentation models, such as UNet [11], PSPNet [12], and DeepLabv3+ [13], employ an encoder–decoder architecture, extract deep semantic features from input images through downsampling and then progressively reconstruct high-resolution feature maps via upsampling from the encoded features. However, these semantic segmentation models tend to underutilize the implicit spectral characteristics of water bodies across different bands in remote sensing images. Moreover, their high computational complexity and large number of parameters make real-time segmentation challenging on edge computing platforms aboard UAVs or satellites.

To tackle these challenges, we propose a lightweight network for real-time water body segmentation in remote sensing images, which incorporates soft constraints derived from approximate spectral features (ASFC-LNet), specifically designed for real-time water extraction tasks. The ASFC-LNet operates on raw digital number (DN) or top-of-atmosphere (TOA) radiance images without the need for atmospheric correction. Approximate spectral feature indices are derived from band math and integrate these spectral features as soft constraints during deep feature extraction. Furthermore, a self-distillation mechanism is introduced [14], the method mitigates interference from pixel values of nonwater objects resulting from the lack of atmospheric correction. This approach enables more precise real-time segmentation of water bodies in complex remote sensing scenarios.

The ASFC-LNet model encompasses the following aspects.

1) We introduce approximate spectral feature indices as auxiliary inter-band feature information. A lightweight pseudo-siamese feature extraction network (LPSE) is designed to separately extract features from visible bands and the approximate spectral indices.
2) We develop an approximate spectral feature soft constraint fusion mechanism (ASFC) that utilizes spatial attention to selectively fuse effective target features from the visible bands and approximate spectral indices.
3) We incorporate an atrous spatial pyramid pooling (ASPP) module for edge feature enhancement within a self-distillation edge-aware lightweight decoder. The decoder dynamically generates self-distillation edge labels, which guide the edge decoder to progressively learn edge features without additional computational cost during inference.
4) Finally, for model deployment and real-time inference, the proposed ASFC-LNet is accelerated and quantized using TensorRT and deployed on the embedded device Jetson Orin NX.

The rest of this article is organized as follows. Section II introduces the related work. Section III presents the proposed method. Section IV describes the experimental data, computational environment, and results. Finally, Section V concludes this article.

II. Related Work

A. Surface Water Extraction From Remote Sensing Images Using Spectral Features

Extracting water bodies from remote sensing images based on spectral features primarily employs thresholding techniques applied to water indices. Appropriate spectral bands are selected to construct water index models. These models are then used to compute single-band spectral feature grayscale maps. The grayscale distributions of these images are analyzed to determine suitable thresholds for binarization. By applying these thresholds to binarize the grayscale images, the spatial distribution of water bodies is effectively obtained.

McFeeters [7] introduced the NDWI, which exploits the low reflectance of water bodies and the high reflectance of vegetation in the near-infrared (NIR) band. By calculating the normalized difference between the green band and the NIR band, NDWI effectively suppresses vegetation signals to extract water information. Klemenjak et al. [15] proposed the RE-NDWI, replacing the NIR band in the NDWI formula with the red-edge band, applied on RapidEye satellite images. This substitution enhances the suppression of background information such as vegetation and soil, improving water body extraction. Wang et al. [16] developed the EWI, introducing a weighting factor into the denominator of the MNDWI. This adjustment accentuates water features and prevents anomalous results when the normalized difference vegetation index values of water pixels are zero or negative.

Shadow in remote images is also a challenging problem, and many publications investigated shadow removal of remote images [17]. To mitigate the interference caused by shadows [18] in water body extraction, Feyisa et al.AWEI [8] introduced the. AWEI employs five spectral bands from the Landsat 5 and maximizes the separability between water and nonwater pixels by differencing and summing these bands with specific coefficients. Yao et al. [19] presented the high-resolution water index (HRWI), which utilizes the red and green visible bands along with the NIR band. The optimal coefficients for HRWI using a support vector machine method. By incorporating building shadow detection techniques, they suppressed interference from building shadows, enabling the automatic extraction of urban water bodies. Wu et al. [20] developed the two-step urban water index (UWI), which combines the UWI and the urban shadow index. This composite index effectively distinguishes water bodies from shadows in urban environments. Xie et al. [21] proposed the NDWI-morphological shadow index (MSI) index, which integrates the NDWI with the MSI, applied to WorldView-2 satellite images. This integration enhances the delineation of water bodies while suppressing shadow regions. Li et al. [22] extended the MNDWI by proposing two new indices: the contrast difference water index (CDWI) and the shadow difference water index (SDWI). The CDWI effectively enhances water features by incorporating each pixel's maximum and minimum reflectance information into the MNDWI. The SDWI efficiently suppresses shadows by adding blue and green band reflectance information into the MNDWI. By applying a background regularizer to locally weigh the CDWI and SDWI, they derived the background difference water index for extracting surface water in complex backgrounds.

B. Water Extraction From Remote Sensing Images Using Semantic Segmentation Networks

Convolutional neural network (CNN)-based semantic segmentation has been extensively applied to processing remote sensing images [23]. The fully convolutional network (FCN) [24] is a pioneering end-to-end semantic segmentation model that leverages deep learning techniques. Building upon the foundation established by FCN, UNet [11] introduced an encoder–decoder architecture for image semantic segmentation. The encoder consists of multiple convolutional and pooling layers that extract features and reduce spatial resolution, while the decoder progressively restores the original resolution through upsampling operations. SegNet [25] enhances this approach by utilizing max-pooling indices to preserve spatial location information within feature maps, enabling the decoder to more accurately recover fine details during upsampling. PSPNet [12] incorporates a pyramid pooling module that performs pooling operations on high-level features at multiple scales, generating subregions of different sizes. Transformer architectures have further improved segmentation tasks. Swin Transformer [26] applies a hierarchical Transformer structure to image segmentation, efficiently capturing global contextual information. SegFormer [27] employs a Transformer-based encoder in conjunction with a lightweight multilayer perceptron decoder, achieving high precision and efficiency in semantic segmentation.

Chen [28] significantly advanced semantic segmentation with DeepLab series of models. In DeepLab v1, they introduced conditional random fields as a postprocessing step to smooth segmentation results, thereby reducing boundary inaccuracies and eliminating small erroneous regions. DeepLab v2 [29] removed certain pooling operations within the network and replaced standard convolutions with atrous (dilated) convolutions. This modification enabled dense feature extraction and expanded the receptive field without increasing computational cost. DeepLab v3 [30] further enhanced the ASPP module by optimizing the atrous rate settings, which improved the model's ability to capture multiscale contextual information effectively. In DeepLab v3+ [13], the Xception architecture was employed as the backbone network, depthwise separable convolutions were utilized to reduce model parameters, and max-pooling operations were substituted with convolutions, collectively enhancing segmentation performance. HRNet [31] maintains high-resolution representations by connecting multiple resolution branches in parallel and performing multiscale feature fusion, thereby preserving detailed spatial information throughout the network.

Semantic segmentation methods have been applied to water body extraction in remote sensing images. Li et al. [32] introduced the dense-local feature compression network, which extracts water bodies from high-resolution remote sensing imagery. Guo et al. [33] proposed the multiscale water extraction network, where encoder feature maps are fed into dilated convolutions with varying dilation rates to capture multiscale features. Wang et al. [34] developed a water body extraction method for remote sensing images, employing a dual attention module to enhance global dependencies across both spatial and channel dimensions. Safavi and Rahnemoonfar [35] compared the performance of various semantic segmentation networks on aerial imagery, evaluating them using the FloodNet dataset. Duan and Hu [36] proposed a multiscale refinement network for water body segmentation, leveraging multiscale features to achieve more precise results.

To address the challenge of extracting water body boundaries, Chen et al. [37] proposed a hybrid semantic segmentation method based on K-Net, achieving high-precision lake water extraction through iterative refinement of feature information. Freitas et al. [38] utilized PlanetScope Dove satellite imagery with the mask R-CNN model, analyzing performance at different confidence thresholds to select the optimal threshold that balances precision and recall. Xiang et al. [39] introduced the dense pyramid pooling module (DensePPM) to mitigate discontinuities in water body predictions caused by outliers in aerial imagery. Liu et al. [40] proposed a multiscale feature extraction network for water body segmentation, employing contrastive learning to reduce the requirement for large sample sizes. Miao et al. [41] presented the RRF DeconvNet, which combines DeconvNet, residual units, and an edge-weighted loss function to make the network more sensitive to water body boundaries.

Moreover, spectral features have been incorporated into deep neural networks for semantic segmentation. In [42], a lake reservoir extraction method was proposed, which combines the NDWI with thresholding in the NIR band. Ma et al. [43] introduced a water extraction network that integrates water indices with the Swin Transformer. Li et al. [44] proposed a spectral index-driven, weakly supervised method for water body extraction. Broni-Bediako et al. [45] improved water body segmentation accuracy by incorporating NIR band features alongside RGB image bands.

C. Remote Sensing Image Segmentation Using Lightweight Networks

Lightweight neural networks hold significant potential in resource-constrained environments, such as UAVs and satellites. Iandola [46] proposed SqueezeNet, a lightweight and efficient

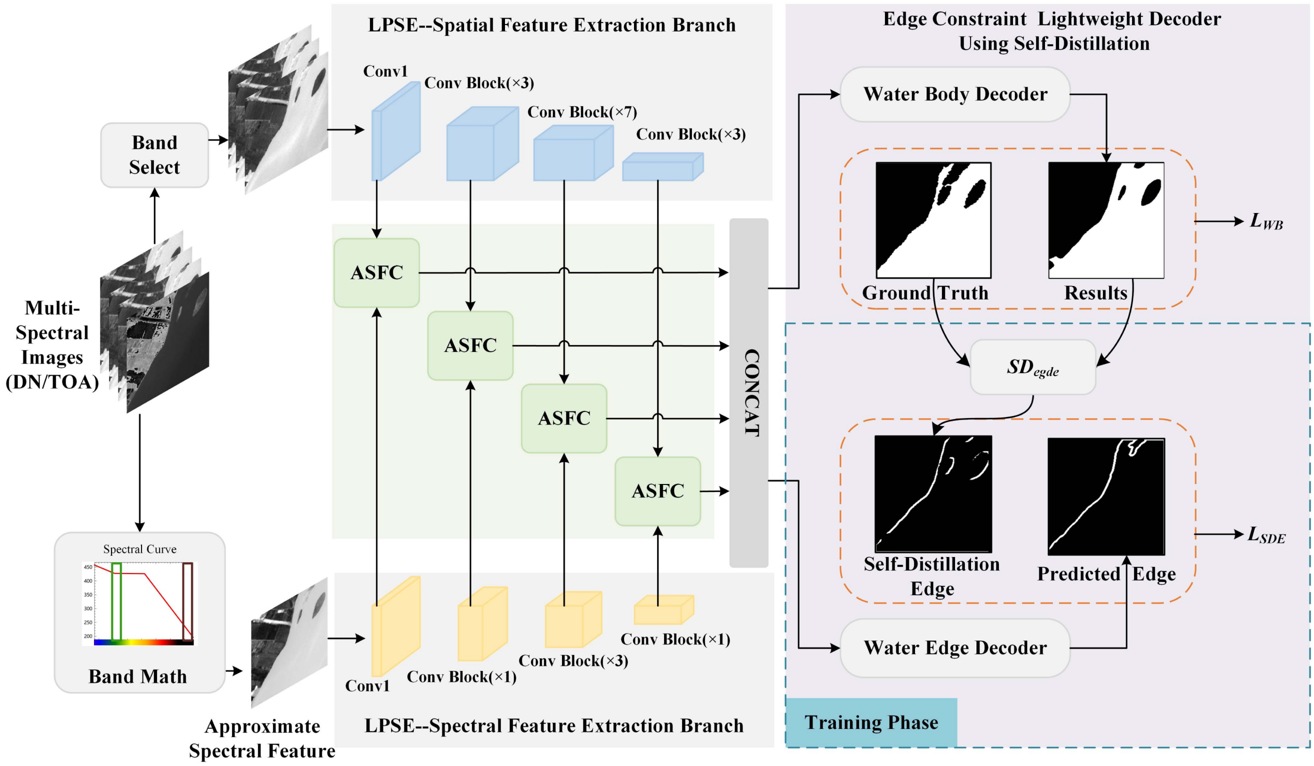


Fig. 1. Overall architecture of the proposed ASFC-LNet network.

CNN with a compressed model size of 0.48 MB. The authors in [47], [48], and [49] introduced the MobileNet series, designed for mobile devices and embedded systems. ResNet and DenseNet have demonstrated that reusing features can effectively improve network performance and accelerate convergence. MobileNetv2 introduces inverted residual structures with linear bottlenecks to enhance efficiency. Zhang et al. [50] proposed ShuffleNet, which employs group convolution and channel shuffling to reduce computational complexity. Face++ introduced ShuffleNet v2 [51], incorporating pointwise group convolution and channel reordering, achieving a model with 1.8 million parameters and a computational complexity of 146 MFLOPs while maintaining high efficiency.

Siam et al. [52] explored combinations of encoders and decoders to construct frameworks for lightweight segmentation networks. For example, VGG16 and ResNet18, as well as MobileNet and ShuffleNet, were utilized as feature extraction encoders, while networks like UNet served as decoders. Paszke et al. proposed ENet [53], designed for mobile applications, achieving a parameter count of only 0.36 million and a computational complexity of 3.8 GFLOPs. Zhao et al. [54] introduced ICNet, which employs a multiresolution cascade structure; low-resolution branches capture coarse semantic features, while high-resolution branches recover and refine detailed features. Romera et al. [55] presented the Efficient ConvNet model, which combines residual connections and factorized convolutions by decomposing 3 × 3 convolutional kernels into sequential 3 × 1 and 1 × 3 convolutions, significantly reducing computational cost. EDANet [56] utilizes asymmetric convolutional dense modules that decompose n × n convolution kernels into n × 1 and 1 × n kernels for enhancing segmentation efficiency. ESPNet [57] employs a modular design that combines spatial pyramids of dilated convolutions and pointwise convolutions, achieving a 0.36 million parameters and an inference speed of approximately 112 FPS. ESPNetv2 [58] introduces grouped pointwise convolutions to enhance interchannel information exchange.

III. Methods

A. Overview

We propose a lightweight network (ASFC-LNet) for real-time segmentation of water bodies in remote sensing images, integrating soft constraints based on approximate spectral features. The ASFC-LNet operates on raw DN or TOA radiance images without the need for atmospheric correction. The proposed network primarily comprises three modules: a lightweight pseudo-siamese feature extraction network (LPSE), an approximate spectral feature soft constraint for multi-scale fusion (ASFC), and an edge constraint lightweight decoder using self-distillation. The overall architecture of the proposed network is illustrated in Fig. 1.

The lightweight pseudo-siamese feature extraction network is designed to separately extract spatial and spectral features from visible bands and the approximate spectral indices. Specifically, the approximate spectral indices are calculated by band math. The features are fused selectively in the feature fusion module

TABLE I
STRUCTURE OF THE SPATIAL FEATURE EXTRACTION BRANCH

Stage	Layer	Stride	Repeat	Output	Params
Input				512×512×3	
	Conv2d	2	1	256×256×24	696
	Max Pooling	2	1	128×128×24	
Stage1	DownSampling Block	2	1	64×64×48	6936
	Basic Block	1	3	64×64×48	
Satge2	DownSampling Block	2	1	32×32×96	45552
	Basic Block	1	7	32×32×96	
Satge3	DownSampling Block	2	1	16×16×192	89952
	Basic Block	1	3	16×16×192	
Sum					143136

TABLE II
STRUCTURE OF THE SPECTRAL FEATURE EXTRACTION BRANCH

Stage	Layer	Stride	Repeat	Output	Params
Input				512×512×3	
	Conv2d	2	1	256×256×24	696
	Max Pooling	2	1	256×256×24	
Stage1	DownSampling Block	2	1	64×64×48	3912
	Basic Block	1	1	64×64×48	
Satge2	DownSampling Block	2	1	32×32×96	24240
	Basic Block	1	3	32×32×96	
Satge3	DownSampling Block	2	1	16×16×192	50208
	Basic Block	1	1	16×16×192	
Sum					79156

ASFC. The fused features are then fed into a water body decoder and an edge decoder to predict the water body and its edges. During training phase, self-supervised edge labels are dynamically generated based on the predictions from the water body decoder and the ground truth, guiding the model to focus more on edge. During inference phase, the edge decoder and self-supervised process are omitted and the complexity of the model is not influenced.

B. Lightweight Pseudo-siamese Feature Extraction Network

Both branches of the pseudosiamese feature extraction network are designed with a streamlined architecture that retains only channels highly relevant to the target features, effectively reducing redundancy.

1) Spatial feature extraction branch: We adopt a progressive feature extraction architecture consisting of an initial convolutional layer followed by three stages. By gradually reducing the resolution of the feature maps, the network effectively extracts multiscale spatial information at each stage, capturing features ranging from local details to global structures. Each stage comprises an efficient downsampling block and multiple basic blocks.

DownSampling block: Downsampling is achieved using depthwise convolution with a stride of 2. In both branches, a 1 × 1 convolution is employed to adjust the number of channels. After merging via concatenation, the spatial dimensions of the feature map are halved, and the number of channels is doubled.

Basic block: The input feature channels are split into two branches using channel splitting. Each branch extracts features using 1 × 1 and 3 × 3 convolutions, respectively. After merging through concatenation, a channel shuffle operation is applied. As a result, both the spatial dimensions of the feature map and the number of channels remain unchanged.

Table I shows the details of the spatial feature extraction branch, which adopts the structure of stage=[4, 8, 4]. The outputs of the max pooling layer and the three stages are selected as multiscale features, with feature map sizes of [128 × 128 × 24], [64 × 64 × 48], [32 × 32 × 96], and [16 × 16 × 192].

2) Spectral feature extraction branch: By reducing the number of basic blocks in each stage, the lightweight approximate spectral feature extraction network decreases network depth, complexity, and parameter count, while preserving the ability to perform multiscale feature extraction.

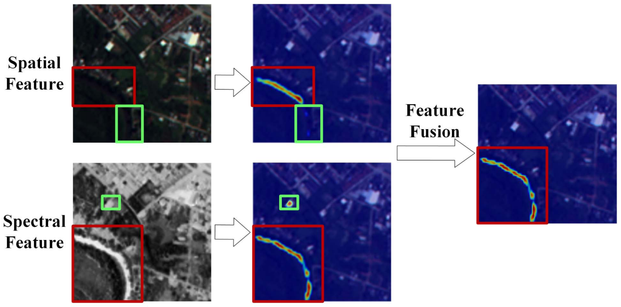


Fig. 2. Visualization of the fused heat map resulting from the combination of spatial and spectral feature extractions.

Spectral feature extraction branch is designed by reducing the number of basic blocks in each stage, thereby decreasing the network depth, complexity, and number of parameters while still retaining the ability to extract multiscale features.

Specifically, the number of basic blocks is reduced by 2, 4, and 2 for the three Stages based on spatial feature extraction branch, resulting in a stage configuration of [2, 4, 2]. To ensure effective fusion of spectral and spatial features, the spectral feature maps are at the same scale as the spatial feature maps, which include the outputs from the max pooling layer and three stages. As shown in Table II, the size of output multiscale feature maps are [128 × 128 × 24], [64 × 64 × 48], [32 × 32 × 96], and [16 × 16 × 192].

C. Approximate Spectral Feature Soft Constraint for Multiscale Fusion

As shown in Fig. 2. exclusively on spatial features can lead to missed extraction of small water bodies, while depending solely on spectral features may introduce interference from spectral variability due to imprecise atmospheric correction. To fully exploit the complementary advantages of both feature types and mitigate the impact of interference in spectral data, we introduce an attention mechanism. This mechanism generates fusion weights based on attention maps, enabling multiscale feature fusion with soft constraints provided by approximate spectral features.

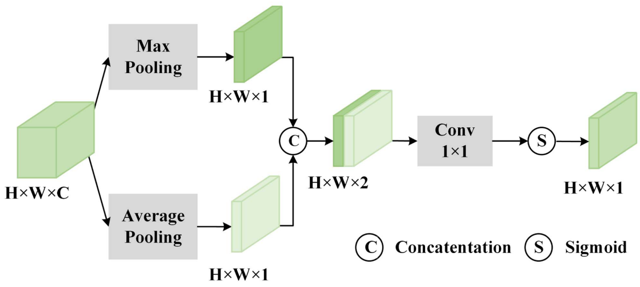


Fig. 3. Architecture of spatial attention mechanism.

The spatial attention mechanism aims to focus on the important regions of the feature map along the spatial dimension. The structure of the spatial attention mechanism is illustrated in Fig. 3. In order to obtain the overall distribution and locally spatial feature, the input feature map $X \in R^{H\times W\times C}$ is fed into the global average pooling and the global maximum pooling. The input feature maps are aggregated along the channel dimension to obtain two single-channel feature maps. The two single-channel feature maps are concatenated along the channel dimension to form a feature map containing two kinds of spatial information. Further, the two kinds of spatial feature are fused by the convolution, and the values of the spatial weight map are mapped to the range [0, 1] by the Sigmoid function. Finally, the output is the spatial attention weight map. This weight map represents the importance of pixels. The larger the value, the more significant of the pixel. The spatial attention is defined as follows:

$$X_{\max} = \mathrm{MaxPooling}(X) \tag{1}$$

$$X_{\mathrm{avg}} = \mathrm{AveragePooling}(X) \tag{2}$$

$$X_{\mathrm{concat}} = \mathrm{Concat}\left[X_{\max}, X_{\mathrm{avg}}\right] \tag{3}$$

$$M_s = \sigma\left(\mathrm{Conv}_{1\times 1}\left(X_{\mathrm{concat}}\right)\right) \tag{4}$$

where $X_{\max}$ and X_{avg} denote the feature maps after maximal and average pooling, $X_{\max}, X_{avg} \in R^{H\times W\times 1}$, X_{Concat} denotes the feature map containing two kinds of pooling,$X_{\mathrm{concat}} \in R^{H\times W\times 2}$, and M_s the spatial attention weight map, σ denotes the Sigmoid function, and $M_s \in R^{H\times W\times 1}$.

Furthermore, we propose a soft-constrained spectral fusion module for approximate spectral features, which jointly considers the spatial attention of both feature types to extract key spatial information and generate shared fusion weights. Specifically the following holds.

When discrepancies exist in the target attention features between the visible bands spatial features and the spectral features, the weight generation phase fuses the spatial information from both to achieve a comprehensive and accurate representation of the target region.

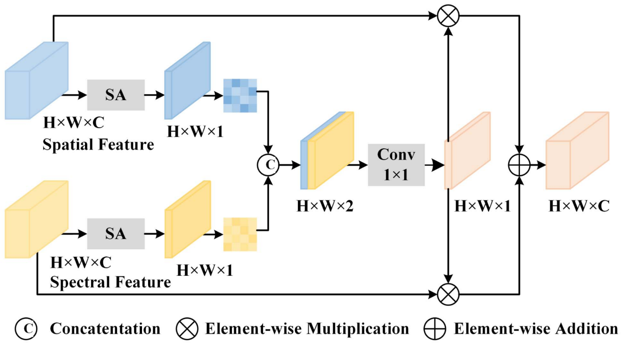


Fig. 4. Architecture of soft-constrained spectral feature fusion module.

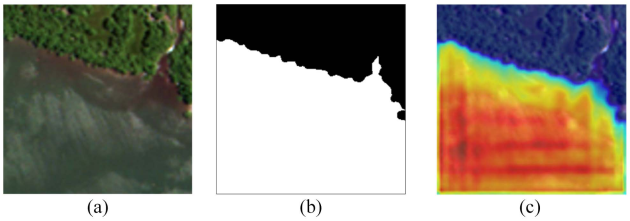

(a) (b) (c)

Fig. 5. Heat map distributions for water feature extraction: (a) original remote sensing image, (b) ground truth water labels, and (c) heat map of water features.

When the spatial features and spectral features exhibit high spatial consistency, with similar salient regions and target distributions, sharing a common spatial weight map effectively reduces redundant computations [59]. This approach ensures that the spatial information in both feature branches remains consistent, avoiding issues of inconsistent feature representations that may arise from independent weights.

Fig. 4 shows the architecture of soft-constrained spectral feature Fusion module. Algorithm 1 outlines the process of soft constrained spectral feature fusion. The two attention maps generated from spatial and spectral features are concatenated. The feature maps $X \in R^{H\times W\times 2}$ contain the target attention information of the two features. A convolution is applied to the concatenated feature map, which extracts effective target attention. The fusion feature weight is generated using the Sigmoid function, where each value of the weight map represents the importance of the feature at the corresponding pixel in the feature map. The higher the value of the weight map, the more significant the feature at that pixel, and the greater its impact on the final segmentation result. The adaptive weight feature fusion is defined by the (5) and (6) as follows:

$$\begin{aligned} \mathrm{Weight} = \sigma\,(f\,(\mathrm{SA}\,(\mathrm{Feature}_{\mathrm{Spatial}}) \\ + \mathrm{SA}(\mathrm{Feature}_{\mathrm{Spectral}}))) \end{aligned} \tag{5}$$

$$\begin{aligned} F_{\mathrm{fusion}} = \mathrm{Feature}_{\mathrm{Spatial}} \times \mathrm{Weight} \\ + \mathrm{Feature}_{\mathrm{Spectral}} \times \mathrm{Weight} \end{aligned} \tag{6}$$

where $\mathrm{Feature}_{\mathrm{Spectral}}$ denotes spectral features, $\mathrm{Feature}_{\mathrm{Spatial}}$ denotes spatial features, SA denotes spatial attention computation process, f denotes convolutional layer, σ denotes Sigmoid function layer, and Weight denotes shared feature fusion weight, and F_{fusion} denotes the feature after fusion.

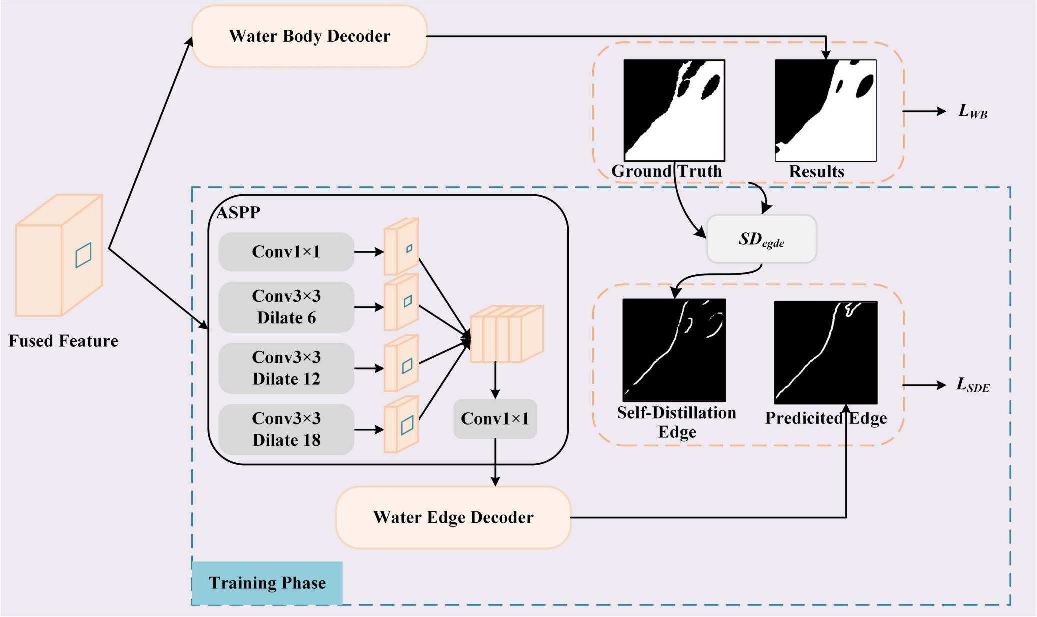


Fig. 6. Architecture of edge constraint lightweight decoder using self-distillation.

```
Algorithm 1: ASFC Module.
  Input: Spatial features $Feature_{Spatial}$ and spectral
   features $Feature_{Spectral}$
  1: Spatial attention computation: Compute the attention
     maps of $Feature_{Spatial}$ and $Feature_{Spectral}$.
         $M_{s\text{-}Spatial} = SA\left(Feature_{Spatial}\right)$
         $M_{s\text{-}Spectral} = SA\left(Feature_{Spectral}\right)$
  2: Spatial attention feature stacking: Concatenate the
     attention maps generated from spatial and spectral
     features.
         $M_{concat} = \text{Concat}[M_{s\text{-}Spatial}, M_{s\text{-}Spectral}]$
  3: Shared attention weight generation:
         $Weight = \sigma\left(f\left(\left(M_{concat}\right)\right)\right)$
  4: Spatial feature fusion weighting:
         $F_{fusion} = Feature_{Spatial} \times Weight$
                 $+ Feature_{Spectral} \times Weight$
```

D. Edge Constraint Lightweight Decoder Using Self-Distillation

In the scenario illustrated in Fig. 5, the surface water body exhibits intricate boundaries, and the color of the water near these edge regions differs from that of the central area. This inconsistency leads to weaker model responses at the edges, making it difficult to accurately capture the water body boundaries. To address this issue, we design a lightweight edge-aware decoder based on self-distillation. As shown in Fig. 6, this decoder incorporates a multiscale context feature extraction module, specifically the ASPP, to extract multiscale information from the edge regions. Dynamic self-distillation signals are generated based on the predictions of the water body decoder and the ground truth labels, constraining the output of the edge decoder.

During the training phase, the water edge decoder and the water body decoder share the feature extraction network. By utilizing dynamic labels and a self-distillation edge loss, we guide the shared layer features to focus more on the edge regions. In the inference phase, the edge decoder is removed, and the model completes the segmentation task using only the water body decoder. This approach enhances boundary segmentation accuracy while ensuring computational efficiency.

The water body decoder processes the fused features to produce the segmentation results for water bodies. This decoder utilizes a 1 × 1 convolution layer to generate the probability map of the target class. The design of the 1 × 1 convolution not only keeps the network lightweight but also allows for efficient adjustment of the number of channels. Then, the output is upsampled to the original resolution of the input image using bilinear interpolation. This results in a segmentation prediction map that matches the input dimensions, enabling precise identification of water body regions.

To optimize the output of the water body segmentation decoder, ground truth labels are used to supervise the segmentation predictions. For the primary segmentation task, the binary cross-entropy loss function is employed to enhance the model's performance by minimizing the discrepancy between the predicted probability distribution and the ground truth distribution. The loss function is defined as follows:

$$L_{WB} = -\sum_{(x,y)} \left[G(x,y) \log P(x,y) + (1 - G(x,y)) \log\left(1 - P(x,y)\right)\right] \tag{7}$$

where $G(x,y)$ denotes the ground truth, $P(x,y)$ is the probability map, and (x,y) represents the position of the pixel in the images.

For edge constraint, a multiscale context feature extraction module (ASPP) is introduced to capture multiscale information in boundary regions. The ASPP module significantly expands the receptive field for feature extraction by employing convolution operations with various dilation rates (6, 12, and 18), thus balancing local details and global context information at different scales. Convolutions with smaller dilation rates focus on local features near edges, such as texture details and subtle transitions, while larger dilation rates capture broader background information and the global structural characteristics of boundaries. Finally, the convolution outputs from each branch are fused through a concatenation operation, followed by a 1×1 convolution to reduce dimensionality, generating a feature map enriched with detailed edge information.

The ASPP module generates a feature map enriched with detailed edge information, which is then passed to the water edge decoder for further processing. The water edge decoder mirrors the structure of the water body decoder, employing 1×1 convolutions and upsampling to progressively decode deep edge features into edge prediction results.

To effectively guide the training of the edge features, we introduce a dynamic self-distillation mechanism. This mechanism generates self-distilled edge labels by combining the predictions from the water body decoder with the ground truth boundary labels. It is defined by the following formula:

$$\mathrm{SD}_{edge} = G(x,y) \times Edge\_Extract\left(R(x,y)\right) \tag{8}$$

where SD_{edge} denotes the dynamically generated self-distillation edge label; $G(x,y) \in \{0,1\}$ denotes the ground truth of pixel(x,y); $R(x,y)$ denotes the binarized result of the output of the water body decoder; $Edge\_Extract()$ denotes the edge extraction.

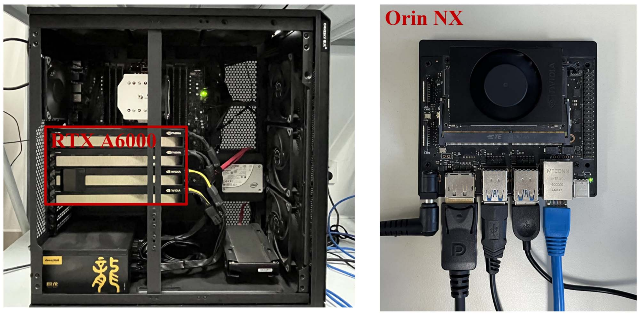
To compute the self-distillation edge loss, we compare the predicted outputs of the water edge decoder with the dynamically generated edge labels. The loss function is defined as follows:

$$L_{SDE} = 1 - \frac{2\sum_{(x,y)} P_{edge}(x,y) SD_{edge}(x,y)}{\sum_{(x,y)} P_{edge}(x,y)^2 + \sum_{(x,y)} SD_{edge}(x,y)^2} \tag{9}$$

where $P_{edge}(x,y)$ is the probability map of edge. The dynamic edge label SD_{edge} is derived by combining the ground truth with the edge prediction results of the water body decoder. It guides the network to focus on learning the correctly predicted areas of the water body decoder. This process gradually expands to include low-confidence or complex edge, achieving optimization of edge from simple to complex.

Therefore, the final loss function is defined as the sum of these two terms

$$L = L_{\mathrm{WB}} + L_{\mathrm{SDE}}. \tag{10}$$



(a) (b)

Fig. 7. High-performance computing platform for ground-based training (left); Embedded computing platform for real-time inference on UAVs or satellites (right).

TABLE III
GPU Specifications of the Training and Inference Platforms

Parameter	NVIDIA RTX A6000	NVIDIA Jetson Orin NX 16GB
CUDA Cores	10752	1024
Tensor Cores	336	32
Memory	48G 384-bit GDDR6	16GB 128-bit LPDDR5
Memory Bandwidth	768GB/s	102.4GB/s
Power	300W	10 –25W
Performance	309.7 TFLOPS	100 TOPS

IV. Experiments and Results

A. Description of Computing Facilities

1) High-performance computing platform for ground-based training:

For model training, we utilize high-performance NVIDIA RTX A6000 GPUs (48 GB) and computers equipped with Intel Xeon Gold 5320 CPUs (2.20 GHz) and 256 GB of memory, as shown on the left side of Fig. 7.

2) Embedded computing platform for real-time inference on UAVs or satellites:

To evaluate the real-time inference capabilities of the water segmentation models on edge computing platforms, such as UAVs and satellites, we conducted inference tests using a Jetson Orin NX embedded computing board, as shown on the right side of Fig. 7. The GPU specifications of the training and inference platforms are listed in Table III.

B. Parameter Settings for Semantic Segmentation Models

We implemented the models using the PyTorch framework and trained it on a NVIDIA A6000 GPU. The model was trained for 50 epochs with a batch size of 64 using the Adam optimizer. The initial learning rate was set to 0.0005 and decayed to a minimum of 0.000005 following a cosine annealing schedule; a weight decay of 0.0001 was also applied.

To facilitate model quantization and embedded deployment on the NVIDIA Jetson Orin NX platform, we converted the trained model to ONNX format. TensorRT 8.5.4.2 was employed

TABLE IV
SEMANTIC CLASSES AND LABEL ENCODING OF THE FLOODNET DATASET

Semantic class	Original label	Merged label
Background	0	0
Flooded Buildings	1	0
Nonflooded Buildings	2	0
Flooded Roads	3	1
Nonflooded Roads	4	0
Water	5	1
Trees	6	0
Vehicles	7	0
Pools	8	1
Grassland	9	0

for quantization and deployment. The inference time after deployment was measured using the built-in trtexec tool.

C. Datasets

1) UAV Remote Sensing Dataset: FloodNet [60]: FloodNet is a UAV-based aerial remote sensing dataset designed for semantic segmentation with pixel-level annotations. It consists of 2343 aerial images at a size of 3000 × 4000 pixels, each containing RGB three-band data. The dataset is partitioned into a training set of 1445 images, a validation set of 450 images, and a test set of 448 images. It encompasses ten different semantic categories, as listed in Table IV. The original images have not undergone atmospheric correction and are stored in JPEG format, while the semantic label images are stored in PNG format.

To evaluate the effectiveness of various methods for water body extraction in remote sensing images, we processed the original FloodNet dataset by merging the classes. Flooded roads are defined as areas where the road surface is directly covered by water. In contrast, flooded buildings refer to buildings surrounded by flood waters but where the building surfaces themselves are not submerged. Consequently, the three classes—flooded roads, water, and swimming pools—were merged into a single class labeled water (label 1). The remaining classes, including flooded buildings, were combined into a background class (label 0).

2) High-resolution satellite remote sensing dataset: GF-FloodNet [61]: The GF-FloodNet dataset comprises 13 388 multispectral images of size 256 × 256 pixels captured by China's Gaofen-2 satellite. The images have not undergone atmospheric correction, and the spatial resolutions include 1.5, 2.5, and 4 m. The dataset contains two classes: water bodies (label 1) and background (label 0). It encompasses four spectral bands: red, green, blue (RGB), and NIR. The wavelength information for each spectral band is provided in Table V.

The GF-FloodNet dataset covers multiple regions worldwide. Variations in background across different regions lead to an imbalanced data distribution. If the training set includes samples from only certain regions, the model may not learn features representative of other regions, resulting in poor generalization in those areas. Therefore, when partitioning the dataset, samples from all regions were included in both the training and validation sets, maintaining a training-to-validation ratio of 1:1.

TABLE V
BAND AND WAVELENGTH INFORMATION CONTAINED IN THE GF-FLOODNET DATASET

Band	Wavelength (nm)
Blue	450˜520
Green	520˜590
Red	630˜690
INR	770˜890

D. Evaluation Metrics

In remote sensing image processing tasks involving water body segmentation, the number of pixels representing water is often significantly smaller than that of nonwater background pixels, leading to a data imbalance between target and background classes. This imbalance can adversely affect evaluation metrics commonly used in traditional semantic segmentation, such as OA and mean intersection over union (IoU), rendering them incapable of accurately reflecting the model's performance in segmenting water bodies. To address the bias in accuracy assessment caused by the predominance of background pixels, this study employs metrics, such as IoU, precision, recall, and F1, to evaluate the segmentation accuracy of water bodies in remote sensing images. The evaluation metrics utilized in this experiment are as follows.

1) Precision: It is used to measure the accuracy of the model when predicting a specific category. It represents the proportion of correctly predicted samples for a category out of all samples predicted as that category. The formula is

$$\text{Precision} = \frac{TP}{TP + FP}. \tag{11}$$

2) Recall: It is used to measure the model's capacity to identify samples that actually belong to a specific category. It represents the proportion of correctly predicted samples for a category out of all samples that actually belong to that category. The formula is

$$\text{Recall} = \frac{TP}{TP + FN}. \tag{12}$$

3) IoU: It represents the ratio of the intersection to the union of the predicted and true target areas. The value ranges from 0 to 1, with a higher value indicating better model performance. The formula is

$$\text{IoU} = \frac{TP}{TP + FN + FP}. \tag{13}$$

4) F1: It is the harmonic mean of precision and recall, used to evaluate the model's precision and recall ability simultaneously. The formula is

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{14}$$

TABLE VI
EXPERIMENTAL RESULTS ON THE FLOODNET DATASET

Method	IoU Water (%)	Precision (%)	Recall (%)	F1 (%)	Params (M)	GFLOPs	Inference Time (ms)
Deeplabv3+ [13]	62.95	84.58	80.01	82.23	54.85	236.54	80.23
UNet [11]	70.67	85.37	80.40	82.81	24.89	225.85	30.56
PSPnet [12]	70.18	**85.68**	79.50	82.47	48.96	61.63	25.58
HRnet [31]	70.19	83.61	81.39	82.48	9.64	18.66	36.61
SegNet [25]	70.42	84.42	80.93	82.63	29.44	160.68	–
ABCNet [62]	69.82	84.58	80.01	82.23	13.33	15.36	23.89
UNetFormer [63]	61.34	80.92	76.23	78.51	11.3	46.9	50.43
SegFormer [27]	67.14	83.56	78.91	81.17	3.72	6.41	38.47
MCCANet [64]	68.88	83.35	81.29	82.30	42.29	104	69.21
Ewas [65]	70.06	83.98	80.35	82.12	44.68	50.50	33.69
ASFC-LNet	**70.74**	84.11	**81.65**	**82.86**	**0.22**	**0.32**	**6.45**

The bold values indicate the best performance among the methods.

where TP denotes true positives,; FP denotes false positives; and FN denotes false negatives

5) Parameters: It refers to the total number of parameters that need to be trained in the model, reflecting the model's complexity. The number of parameters in a convolutional layer is calculated as

$$\text{Params} = k_h \times k_w \times C_{\text{in}} \times C_{\text{out}} \tag{15}$$

6) Floating Point Operations (FLOPs): It refers to the total number of FLOPs required by the model to process a single input. The FLOPs of a convolutional layer is calculated as

$$\begin{aligned} \text{FLOPs} = & \left[(C_{\text{in}} \times k_w \times k_h) + (C_{in} \times k_w \times k_h - 1) \right] \\ & \times C_{\text{out}} \times W \times H \end{aligned} \tag{16}$$

where k_h is the height of the convolution kernel, k_w is the width of the convolution kernel, C_{in} is the number of input feature map channels, and $C_{\text{out}}$is the number of output feature map channels.

7) Inference time: It refers to the time required for the trained model to make predictions on new data, reflecting the real-time performance of the model.

E. Results

To validate the effectiveness of the proposed ASFC-LNet water body segmentation method, we compared it with ten existing semantic segmentation algorithms on the UAV dataset FloodNet and the satellite dataset GF-FloodNet. The algorithms used for comparison include Deeplabv3+, UNet, PSPNet, HRNet, SegNet, ABCNet, UNetFormer, SegFormer, MCCANet, and Ewas. All experiments were conducted using the computing facilities described in Section IV-A.

Table VI presents the experimental results on the FloodNet dataset, where bold values indicate the best performance among the methods evaluated. Compared to general deep segmentation models, such as DeepLabv3+, UNet, and PSPNet, the proposed ASFC-LNet method outperforms these models in terms of IoU, Recall, and F1, while also demonstrating higher efficiency in model parameter size and inference time. Specifically, compared to DeepLabv3+, which has the largest number of parameters, our method achieves a 7.79% higher IoU, reduces the parameter count by approximately 54 million, and decreases inference time by over 70 ms. Compared to PSPNet, our method improves IoU, Recall, and F1 by 0.56%, 2.15%, and 0.39% respectively, reduces the parameter count by 48.74 million, and shortens inference time by 19.13 ms.

Compared with other lightweight models such as UNetFormer, SegFormer, and ABCNet, the proposed ASFC-LNet method also demonstrates significant advantages in IoU, Recall, and F1, while achieving higher efficiency in terms of model parameter size and inference time. Specifically, compared to SegFormer, which has the smallest parameter count, our method achieves a 3.60% higher IoU and a 1.69% higher F1, and reduces inference time by 32.02 ms. Compared to ABCNet, a lightweight model with relatively good accuracy, our method improves IoU, Recall, and F1 by 0.92%, 1.64%, and 0.63%, respectively, reduces the parameter count by 13.11 million, and decreases inference time by 17.44 ms.

Table VII presents the experimental results based on the GF-FloodNet dataset, where values in bold indicate the best performance among the methods evaluated. Compared with general deep segmentation models, such as DeepLabv3+, UNet, and PSPNet, the proposed ASFC-LNet method outperforms these models on key metrics including IoU, Recall, and F1, while significantly reducing the number of parameters and inference time. Specifically, compared with DeepLabv3+, which has the largest parameter count, our method achieves 0.17% and 1.50% higher IoU and Recall, respectively, while reducing the parameter count by 54.63 million and shortening inference time by 73.78 ms. Compared with HRNet, our method improves IoU by 0.95%, F1 by 0.51%, and reduces inference time by 30.16 ms.

Compared with other lightweight models, such as UNetFormer, SegFormer, and ABCNet, the proposed ASFC-LNet method also demonstrates significant advantages in IoU, Recall, and F1, while achieving higher efficiency in terms of model parameter size and inference time. Specifically, compared with SegFormer, which has the smallest number of parameters, our

TABLE VII
EXPERIMENTAL RESULTS ON THE GF-FLOODNET DATASET

Method	IoU Water (%)	Precision (%)	Recall (%)	F1 (%)	Params (M)	GFLOPs	Inference Time (ms)
Deeplabv3+ [13]	92.09	**98.06**	93.79	95.88	54.85	236.54	80.23
UNet [11]	90.99	96.14	94.44	95.28	24.89	225.85	30.56
PSPnet [12]	90.70	96.57	93.28	94.90	48.96	61.63	25.58
HRnet [31]	91.31	96.52	94.42	95.46	9.64	18.66	36.61
SegNet [25]	90.95	95.31	95.22	95.26	29.44	160.68	–
ABCNet [62]	89.81	95.83	93.46	94.63	13.33	15.36	23.89
UNetFormer [63]	88.86	94.11	94.09	94.10	11.30	46.90	50.43
SegFormer [27]	90.25	95.89	94.60	95.24	3.72	6.41	38.47
MCCANet [64]	89.96	95.42	94.78	95.10	42.29	104	69.21
Ewas [65]	91.08	96.12	94.59	95.35	44.68	50.50	33.69
ASFC-LNet	**92.26**	96.66	**95.29**	**95.97**	**0.22**	**0.32**	**6.45**

The bold values indicate the best performance among the methods.

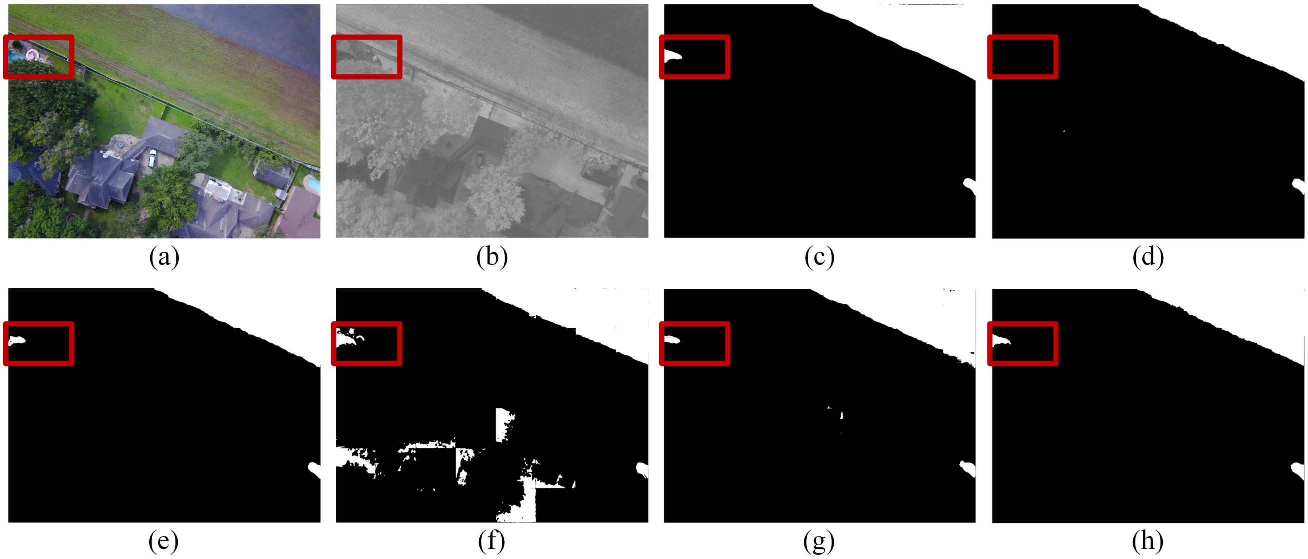


Fig. 8. Visual comparison of segmentation results using different models on Scene 1 of the FloodNet dataset. (a) Original UAV remote sensing image. (b) Approximate spectral feature map. (c) Ground truth labels. (d) DeepLabv3+ segmentation result. (e) UNet segmentation result. (f) PSPNet segmentation result. (g) Ewas segmentation result. (h) Segmentation result of the proposed ASFC-LNet method.

method improves IoU and F1 by 1.99% and 0.73%, respectively, and reduces inference time by 32.02 ms. Compared with ABCNet, a lightweight model with relatively good accuracy, our method increases IoU and F1 by 2.45% and 1.34%, respectively, while reducing the parameter count by 13.11 million and decreasing inference time by 17.44 ms.

To summarize the above analysis, based on the quantitative evaluation metrics, the ASFC-LNet method proposed in this study exhibits significant lightweight advantages compared to models such as DeepLabv3+, UNet, and PSPNet. Utilizing only 0.22 million parameters and 0.32 GFLOPs, it achieves an IoU of 70.74% on the FloodNet dataset and 92.26% on the GF-FloodNet dataset. This method significantly reduces the number of model parameters and computational complexity while maintaining excellent segmentation performance. These results demonstrate its great potential for application in real-time water body segmentation on edge computing platforms, such as UAVs and satellites.

To further evaluate the visual effectiveness of water body segmentation, we conduct a detailed comparison of the proposed ASFC-LNet method with other approaches, particularly emphasizing its capability to accurately segment small water bodies and preserve edge details.

Addressing water body segmentation scenarios involving significantly disparate area sizes, as illustrated in Fig. 8, we present the segmentation results of different models on the FloodNet dataset. Fig. 8(a) shows the original UAV remote sensing image, which contains two water bodies with different colors and shapes in the visible bands, one larger area in the upper right corner and a smaller area in the upper left corner. Fig. 8(b) shows

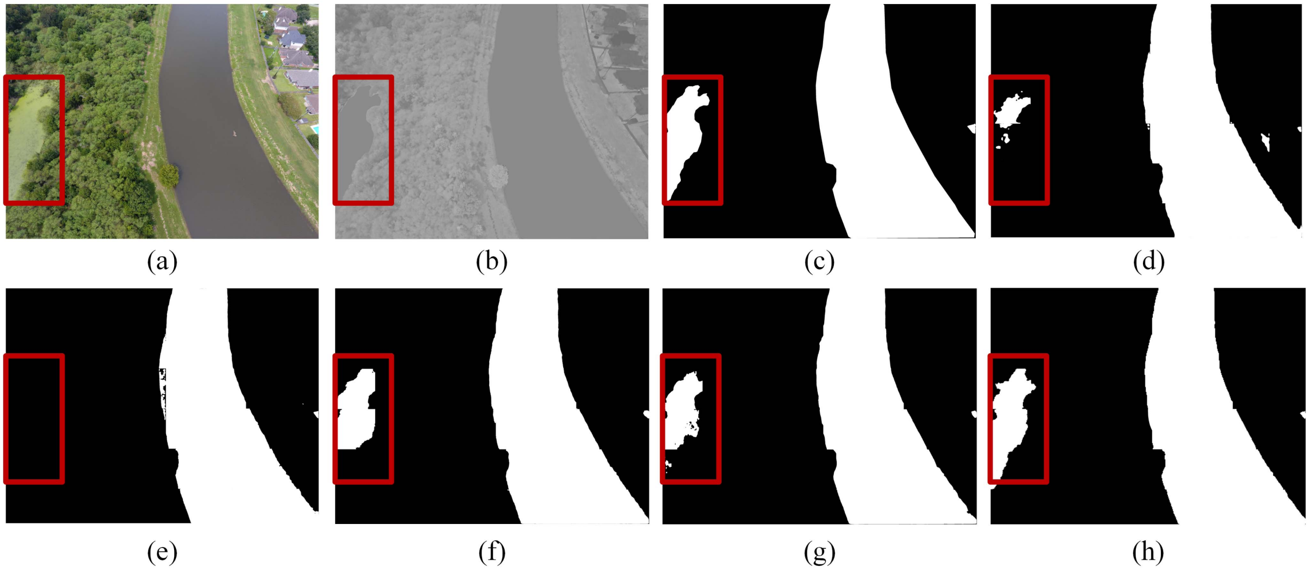


Fig. 9. Visual comparison of segmentation results using different models on Scene 2 of the FloodNet dataset. (a) Original UAV remote sensing image. (b) Approximate spectral feature map. (c) Ground truth labels. (d) DeepLabv3+ segmentation result. (e) UNet segmentation result. (f) PSPNet segmentation result. (g) Ewas segmentation result. (h) Segmentation result of the proposed ASFC-LNet method.

the approximate spectral features obtained through interband calculations, demonstrating that the two water bodies have consistent spectral characteristics. Fig. 8(c) provides the ground truth labels. Fig. 8(d) presents the result of the DeepLabv3+ method, which fails to detect the small blue water body in the upper left corner. Fig. 8(e) shows the result of the UNet method, and Fig. 8(g) shows the result of the Ewas method; both methods can detect the two water bodies but exhibit incomplete boundary detection of the small water body in the upper left corner. Fig. 8(f) illustrates the result of the PSPNet method, showing extensive false detections in the building areas. Fig. 8(h) presents the result of our proposed ASFC-LNet method, which demonstrates the best visual performance among all methods by accurately detecting both the large and small water body regions and precisely extracting their boundaries.

Fig. 9 presents the segmentation results of different models on another scene from the FloodNet dataset. Fig. 9(a) shows the original UAV remote sensing image, which contains two water bodies exhibiting different colors and shapes in the visible spectrum: a larger area on the right and a smaller area on the left with algal blooms on its surface. Fig. 9(b) displays the approximate spectral features obtained through inter-band calculations. Fig. 9(c) provides the ground truth labels. Fig. 9(d), (f), and (g) illustrates the results of the DeepLabv3+, PSPNet, and Ewas methods, respectively. While these three methods detect both water bodies, they fail to fully delineate the edges of the left-side water body with algal blooms. Fig. 9(e) shows the result of the UNet method, which misses the left-side water body. Fig. 9(h) presents the result of our proposed ASFC-LNet method. Among all methods, our approach yields the best visual results, accurately detecting both the left and right water regions and precisely extracting their boundaries, with almost no missed detections or false positives. The detection performance closely approximates the ground truth labels.

Fig. 10 illustrates the segmentation results of various models applied to Scene 3 of the FloodNet dataset, which features water bodies with complex and intertwined boundaries, such as those occluded by trees. Fig. 10(h) presents the results of the proposed ASFC-LNet method, which outperforms all other methods. Notably, small water bodies are completely detected, and the water body boundaries are clear and well-defined.

To evaluate the segmentation performance in the transition regions between water bodies and shorelines, Fig. 11 presents the results of various models applied to Scene 4 of the FloodNet dataset. Most of the compared methods fail to accurately delineate the boundaries of the water bodies. In contrast, Fig. 11(h) shows the results of our proposed ASFC-LNet method, which outperforms all other methods. It accurately segments the water body regions and effectively captures the fine details in the transition areas.

To address the challenge of water body segmentation in large-area and complex scenes, Fig. 12 presents the detection results on a full-scene remote sensing image of an Australian region captured by China's GF-2 satellite. The image has a size of 5376 × 2560 pixels. This scene contains various types of water bodies, including rivers, lakes, and reservoirs, with significant differences in color spectra and morphology features. In the approximate spectral map, the water bodies and the background display distinct values, which provide valuable segmentation information. The proposed ASFC-LNet method accurately detects the different types of water bodies and precisely extracts their boundaries.

F. Ablation Study

Table VIII shows the results of the ablation study on the Floodnet dataset. When only using the baseline method, where

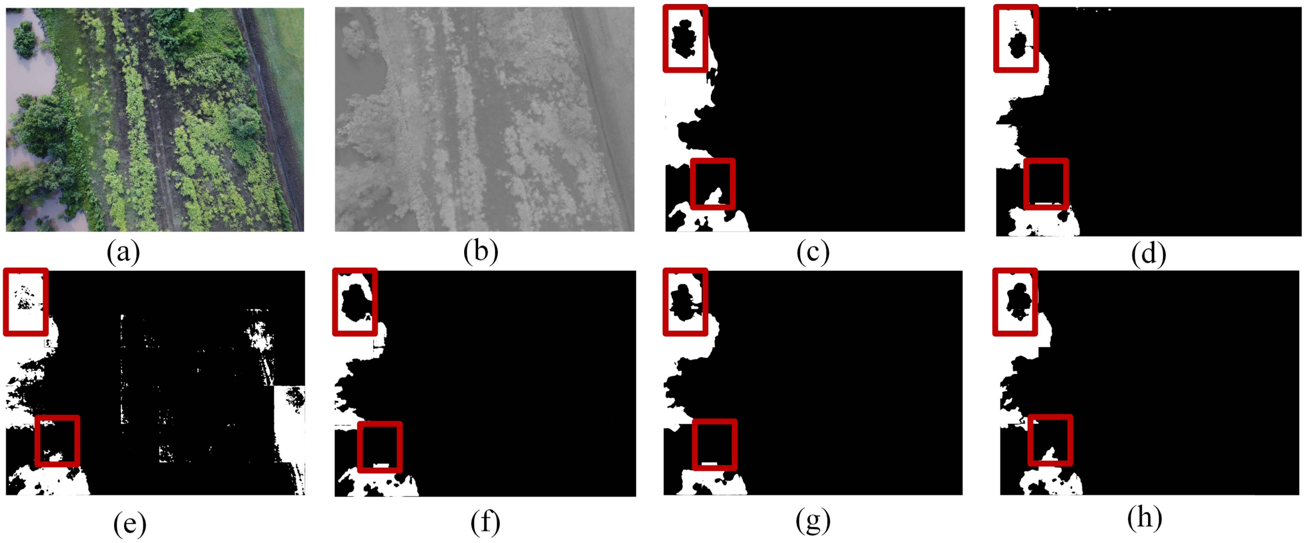


Fig. 10. Visual comparison of segmentation results using different models on Scene 3 of the FloodNet dataset. (a) Original UAV remote sensing image. (b) Approximate spectral feature map. (c) Ground truth labels. (d) DeepLabv3+ segmentation result. (e) UNet segmentation result. (f) PSPNet segmentation result. (g) Ewas segmentation result. (h) Segmentation result of the proposed ASFC-LNet method.

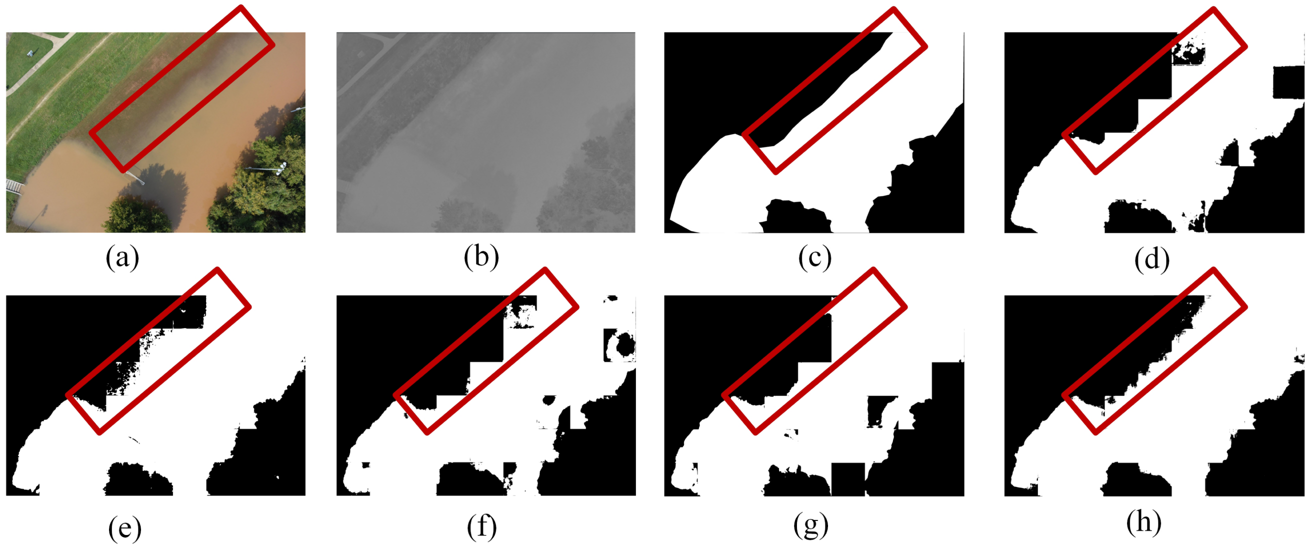


Fig. 11. Visual comparison of segmentation results using different models on Scene 4 of the FloodNet dataset. (a) Original UAV remote sensing image. (b) Approximate spectral feature map. (c) Ground truth labels. (d) DeepLabv3+ segmentation result. (e) UNet segmentation result. (f) PSPNet segmentation result. (g) Ewas segmentation result. (h) Segmentation result of the proposed ASFC-LNet method.

TABLE VIII
Ablation Experiments on the FloodNet Dataset

Method	ASFC	Edge Decoder	IoU Water (%)	Precision (%)	Recall (%)	F1 (%)	Params (M)	GFLOPs
Baseline			67.66	83.53	78.08	80.71	2.15	3.56
Baseline	✓		70.64	83.84	81.77	82.79	2.20	3.68
Baseline	✓	✓	**70.74**	**84.11**	**81.65**	**82.86**	**0.22**	**0.32**

The bold values indicate the best performance among the methods.

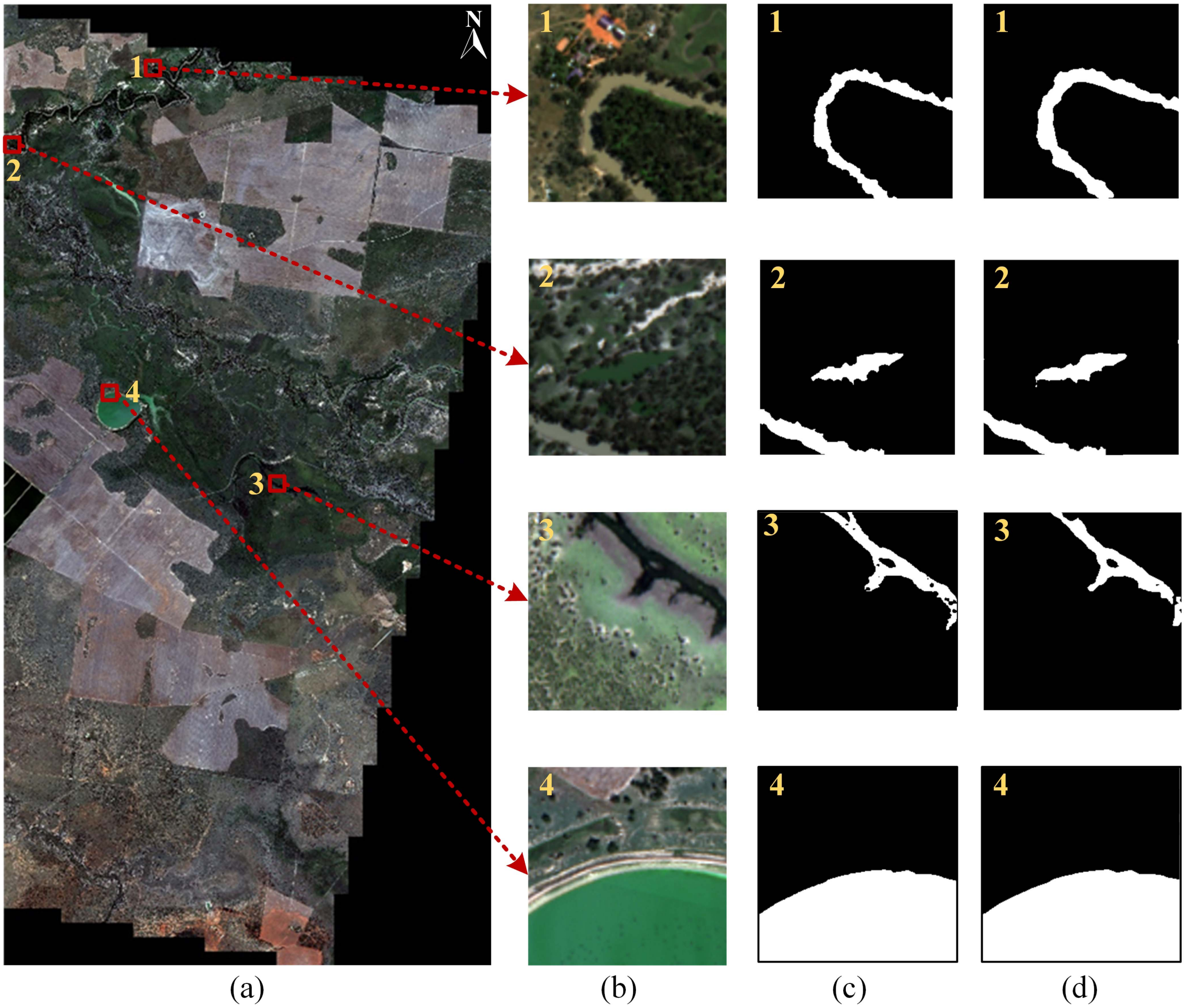


Fig. 12. Segmentation results on the full-scene GF-2 remote sensing image of the Australian region. (a) Original full-scene remote sensing image captured by the GF-2 satellite, without atmospheric correction. (b) Remote sensing images of four local regions. (c) Ground truth water body labels for these local regions. (d) Segmentation results produced by the proposed method.

the ASPP module is in the water body segmentation branch, the IoU reaches 67.66%, the Precision is 83.53%, the Recall is 78.08%, and the F1 is 80.71%. The number of parameters is 2.15 M, and the FLOPs are 3.56 G. After the introduction of the approximate spectral feature soft constraint for multiscale fusion, the IoU is increased to 70.64%, the Precision is increased to 83.84%, the Recall is significantly increased to 81.77%, the F1 is also increased to 82.79%, the number of parameters are increased to 2.2 M, and the FLOPs are increased to 3.68%. The ASPP module introduces a high number of parameters and computational complexity, and we optimized the architecture by relocating the ASPP module to the water edge decoder using self-distillation. This strategy significantly reduces the number of parameters and improves segmentation performance. The IoU reaches 70.74%, Precision is 84.11%, Recall is 81.65%, and F1 is 82.86%. The number of parameters are reduced to 0.22 M, and the FLOPs are reduced to 0.32 G.

The results of the ablation study on the GF-FloodNet dataset are shown in Table IX. When only using the baseline method, where the ASPP module is in the water body segmentation branch, the IoU of the model reaches 89.81%, the Precision is 95.83%, the Recall is 93.46%, the F1 is 94.63%, the number of parameters is 2.15 M, and the FLOPs are 3.56 G. After the introduction of the approximate spectral feature soft constraint for multiscale fusion, the IoU of the model is increased to 92.10%, Precision is increased to 96.32%, Recall is increased to 94.23%, F1 is increased to 95.26%, the number of parameters is increased to 2.2 M, and the number of FLOPs is increased to 3.68 G. The ASPP module introduces a high number of parameters and computational complexity, and we optimized the architecture by relocating the ASPP module to the water edge decoder using self-distillation. This strategy significantly reduces the number of parameters and improves segmentation performance. The IoU of the model reaches 92.26%, Precision is improved to 96.66%,

TABLE IX
ABLATION EXPERIMENTS ON THE GF-FLOODNET DATASET

Method	ASFC	Edge Decoder	IoU Water (%)	Precision (%)	Recall (%)	F1 (%)	Params (M)	GFLOPs
Baseline			89.81	95.83	93.46	94.63	2.15	3.56
Baseline	✓		92.10	96.32	94.23	95.26	2.20	3.68
Baseline	✓	✓	**92.26**	**96.66**	**95.29**	**95.97**	**0.22**	**0.32**

The bold values indicate the best performance among the methods.

Recall is increased to 95.29%, and F1 is improved to 95.97%, and at the same time, the number of parameters is reduced to 0.22 M, and FLOPs are reduced to 0.32 G.

The ablation experiments incorporating spectral adaptive fusion and boundary feature constraints significantly enhanced the model's detection accuracy. During the inference phase, structural optimization substantially reduced the number of parameters and computational complexity while maintaining accuracy.

V. CONCLUSION

To address the challenge of real-time water body extraction in UAV or satellite remote sensing imagery, we propose a lightweight real-time segmentation network that embeds soft constraints of approximate spectral features (ASFC-LNet). This method effectively leverages both the morphological characteristics and approximate spectral features of water bodies, overcoming limitations of traditional methods that rely on precise atmospheric correction and exhibit poor universality in threshold selection. The main contributions of this work are as follows: 1) A lightweight pseudo-siamese feature extraction network (LPSE) is designed to separately extract spatial morphological features and approximate spectral features. By employing a lightweight architecture, the model significantly reduces the number of parameters and computational complexity. 2) A multiscale feature fusion mechanism with soft constraints on approximate spectral features (ASFC) is introduced, enabling flexible fusion of spectral and spatial features. This approach dynamically adapts to the distribution of feature importance across different spatial locations. 3) An ASPP module for edge feature enhancement within a self-distillation edge-aware lightweight decoder is developed, which enhances learning in edge regions by generating dynamic self-edge labels. Experimental results on datasets, such as the UAV aerial remote sensing dataset FloodNet and satellite remote sensing dataset, GF-FloodNet demonstrate that the proposed method achieves optimal segmentation accuracy, boundary preservation, and inference speed without the need for strict atmospheric correction preprocessing. This work provides an effective solution for real-time water body extraction in UAV or satellite remote sensing applications.

REFERENCES

[1] N. Casagli, E. Intrieri, V. Tofani, G. Gigli, and F. Raspini, "Landslide detection, monitoring and prediction with remote-sensing techniques," *Nature Rev. Earth Environ.*, vol. 4, no. 1, pp. 51–64, 2023.

[2] B. Zhang et al., "Progress and challenges in intelligent remote sensing satellite systems," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1814–1822, 2022.

[3] Y. Li, B. Dang, Y. Zhang, and Z. Du, "Water body classification from high-resolution optical remote sensing imagery: Achievements and perspectives," *ISPRS J. Photogrammetry Remote Sens.*, vol. 187, pp. 306–327, 2022.

[4] M. Wieland, S. Martinis, R. Kiefl, and V. Gstaiger, "Semantic segmentation of water bodies in very high-resolution satellite and aerial images," *Remote Sens. Environ.*, vol. 287, 2023, Art. no. 113452.

[5] X. Kang, Z. Fei, P. Duan, and S. Li, "Fog model-based hyperspectral image defogging," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5512512.

[6] X. Yang et al., "Monthly estimation of the surface water extent in France at a 10-m resolution using Sentinel-2 data," *Remote Sens. Environ.*, vol. 244, 2020, Art. no. 111803.

[7] S. K. McFeeters, "The use of the normalized difference water index (NDWI) in the delineation of open water features," *Int. J. Remote Sens.*, vol. 17, no. 7, pp. 1425–1432, 1996.

[8] G. L. Feyisa, H. Meilby, R. Fensholt, and S. R. Proud, "Automated water extraction index: A new technique for surface water mapping using landsat imagery," *Remote Sens. Environ.*, vol. 140, pp. 23–35, 2014.

[9] J. Qu et al., "Progressive multi-iteration registration-fusion co-optimization network for unregistered hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5519814.

[10] J. Qu, W. Dong, Q. Du, Y. Yang, Y. Xu, and Y. Li, "Cyclic consistency constrained multiview graph matching network for unsupervised heterogeneous change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5605315.

[11] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput.-Assisted Interv.: 18th Int. Conf.*, Munich, Germany, 2015, pp. 234–241.

[12] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 2881–2890, 2017.

[13] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[14] Q. Zhu, Z. Li, T. Song, L. Yao, Q. Guan, and L. Zhang, "Unrestricted region and scale: Deep self-supervised building mapping framework across different cities from five continents," *ISPRS J. Photogrammetry Remote Sens.*, vol. 209, pp. 344–367, 2024.

[15] S. Klemenjak, B. Waske, S. Valero, and J. Chanussot, "Unsupervised river detection in rapideye data," in *Proc. 2012 IEEE Int. Geosci. Remote Sens. Symp.*, 2012, pp. 6860–6863.

[16] S. Wang et al., "A simple enhanced water index (EWI) for percent surface water estimation using landsat data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 1, pp. 90–97, Jan. 2015.

[17] P. Duan, S. Hu, X. Kang, and S. Li, "Shadow removal of hyperspectral remote sensing images with multiexposure fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5537211.

[18] Y. Yang, M. Guo, Q. Zhu, L. Ran, and J. Pan, "KO-shadow: Knowledge-driven shadow progressive removal framework for very high spatial resolution remote-sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 4708914.

[19] F. Yao, C. Wang, D. Dong, J. Luo, Z. Shen, and K. Yang, "High-resolution mapping of urban surface water using ZY-3 multi-spectral imagery," *Remote Sens.*, vol. 7, no. 9, pp. 12336–12355, 2015.

[20] W. Wu, Q. Li, Y. Zhang, X. Du, and H. Wang, "Two-step urban water index (TSUWI): A new technique for high-resolution mapping of urban surface water," *Remote Sens.*, vol. 10, no. 11, 2018, Art. no. 1704.

[21] C. Xie, X. Huang, W. Zeng, and X. Fang, "A novel water index for urban high-resolution eight-band worldview-2 imagery," *Int. J. Digit. Earth*, vol. 9, no. 10, pp. 925–941, 2016.

[22] L. Li, H. Su, Q. Du, and T. Wu, "A novel surface water index using local background information for long term and large-scale landsat images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 172, pp. 59–78, 2021.

[23] Y. Yang, J. Qu, W. Dong, T. Zhang, S. Xiao, and Y. Li, "TMCFN: Text-supervised multidimensional contrastive fusion network for hyperspectral and Lidar classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5511015.

[24] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

[25] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[26] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022.

[27] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 12077–12090.

[28] L.-C. Chen, "Rethinking atrous convolution for semantic image segmentation," 2017, arXiv:1706.05587.

[29] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.

[30] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[31] K. Sun et al., "High-resolution representations for labeling pixels and regions," 2019, arXiv:1904.04514.

[32] M. Li, P. Wu, B. Wang, H. Park, H. Yang, and Y. Wu, "A deep learning method of water body extraction from high resolution remote sensing images with multisensors," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 3120–3132, 2021.

[33] H. Guo, G. He, W. Jiang, R. Yin, L. Yan, and W. Leng, "A multi-scale water extraction convolutional neural network (MWEN) method for gaofen-1 remote sensing images," *ISPRS Int. J. Geo- Inf.*, vol. 9, no. 4, 2020, Art. no. 189.

[34] B. Wang, Z. Chen, L. Wu, X. Yang, and Y. Zhou, "SADA-Net: A shape feature optimization and multiscale context information-based water body extraction method for high-resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1744–1759, 2022.

[35] F. Safavi and M. Rahnemoonfar, "Comparative study of real-time semantic segmentation networks in aerial images during flooding events," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 4–20, 2022.

[36] L. Duan and X. Hu, "Multiscale refinement network for water-body segmentation in high-resolution satellite imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 4, pp. 686–690, Apr. 2020.

[37] C. Chen, Y. Wang, S. Yang, X. Ji, and G. Wang, "A k-net-based hybrid semantic segmentation method for extracting lake water bodies," *Eng. Appl. Artif. Intell.*, vol. 126, 2023, Art. no. 106904.

[38] P. Freitas, G. Vieira, J. Canário, W. F. Vincent, P. Pina, and C. Mora, "A trained mask R-CNN model over planetscope imagery for very-high resolution surface water mapping in boreal forest-tundra," *Remote Sens. Environ.*, vol. 304, 2024, Art. no. 114047.

[39] D. Xiang, X. Zhang, W. Wu, and H. Liu, "DensePPMUNet-a: A robust deep learning network for segmenting water bodies from aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4202611.

[40] B. Liu, S. Du, L. Bai, S. Ouyang, H. Wang, and X. Zhang, "Water extraction from optical high-resolution remote sensing imagery: A multi-scale feature extraction network with contrastive learning," *Giscience Remote Sens.*, vol. 60, no. 1, 2023, Art. no. 2166396.

[41] Z. Miao, K. Fu, H. Sun, X. Sun, and M. Yan, "Automatic water-body segmentation from high-resolution satellite images via deep networks," *IEEE Geosci. remote Sens. Lett.*, vol. 15, no. 4, pp. 602–606, Apr. 2018.

[42] K. Yuan, X. Zhuang, G. Schaefer, J. Feng, L. Guan, and H. Fang, "Deep-learning-based multispectral satellite image segmentation for water body detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 7422–7434, 2021.

[43] D. Ma, L. Jiang, J. Li, and Y. Shi, "Water index and swin transformer ensemble (WISTE) for water body extraction from multispectral remote sensing images," *GIScience Remote Sens.*, vol. 60, no. 1, 2023, Art. no. 2251704.

[44] Z. Li, X. Zhang, and P. Xiao, "Spectral index-driven FCN model training for water extraction from multispectral imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 192, pp. 344–360, 2022.

[45] C. Broni-Bediako, J. Xia, and N. Yokoya, "Real-time semantic segmentation: A brief survey and comparative study in remote sensing," *IEEE Geosci. Remote Sens. Mag.*, vol. 11, no. 4, pp. 94–124, Dec. 2023.

[46] F. N. Iandola, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size," 2016, *arXiv:1602.07360*.

[47] A. G. Howard, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[48] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4510–4520.

[49] A. Howard et al., "Searching for mobilenetv3," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 1314–1324.

[50] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6848–6856.

[51] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "ShuffleNet V2: Practical guidelines for efficient CNN architecture design," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 116–131.

[52] M. Siam, M. Gamal, M. Abdel-Razek, S. Yogamani, M. Jagersand, and H. Zhang, "A comparative study of real-time semantic segmentation for autonomous driving," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 587–597.

[53] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "ENet: A deep neural network architecture for real-time semantic segmentation," 2016, *arXiv:1606.02147*.

[54] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia, "ICNet for real-time semantic segmentation on high-resolution images," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 405–420.

[55] E. Romera, J. M. Alvarez, L. M. Bergasa, and R. Arroyo, "Efficient convnet for real-time semantic segmentation," in *Proc. 2017 IEEE Intell. Veh. Symp. (IV)*, 2017, pp. 1789–1794.

[56] S.-Y. Lo, H.-M. Hang, S.-W. Chan, and J.-J. Lin, "Efficient dense modules of asymmetric convolution for real-time semantic segmentation," in *Proc. 1st ACM Int. Conf. Multimedia Asia*, 2019, pp. 1–6.

[57] S. Watanabe et al., "ESPnet: End-to-end speech processing toolkit," 2018, *arXiv:1804.00015*.

[58] S. Mehta, M. Rastegari, L. Shapiro, and H. Hajishirzi, "ESPNetv2: A light-weight, power efficient, and general purpose convolutional neural network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9190–9200.

[59] P. Duan, T. Shan, X. Kang, and S. Li, "Spectral super-resolution in frequency domain," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Oct. 29, 2024, doi: 10.1109/TNNLS.2024.3481060.

[60] M. Rahnemoonfar, T. Chowdhury, A. Sarkar, D. Varshney, M. Yari, and R. R. Murphy, "FloodNet: A high resolution aerial imagery dataset for post flood scene understanding," *IEEE Access*, vol. 9, pp. 89644–89654, 2021.

[61] Y. Zhang, P. Liu, L. Chen, M. Xu, X. Guo, and L. Zhao, "A new multi-source remote sensing image sample dataset with high resolution for flood area extraction: Gf-floodnet," *Int. J. Digit. Earth*, vol. 16, no. 1, pp. 2522–2554, 2023.

[62] R. Li, S. Zheng, C. Zhang, C. Duan, L. Wang, and P. M. Atkinson, "ABCNet: Attentive bilateral contextual network for efficient semantic segmentation of fine-resolution remotely sensed imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 181, pp. 84–98, 2021.

[63] L. Wang et al., "Unetformer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 190, pp. 196–214, 2022.

[64] J. Zheng, A. Shao, Y. Yan, J. Wu, and M. Zhang, "Remote sensing semantic segmentation via boundary supervision-aided multiscale channelwise cross attention network," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4405814.

[65] T. M. Pham, N. Do, H. T. Bui, and M. V. Hoang, "Enhanced deep learning-based water area segmentation for flood detection and monitoring," *Mach. Learn.: Sci. Technol.*, vol. 5, no. 4, 2024, Art. no. 045025.

Qingqing Cao received the B.S. degree in communication engineering from the College of Computer Science and Electronic Engineering, Hunan University, Changsha, China, in 2022. She is currently working toward the joint M.S. degree in electronic and information engineering with the University of Chinese Academy of Sciences, Beijing, China, and the Key Laboratory of Computational Optical Imaging Technology, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing.

Her research interests include real-time water body segmentation and remote sensing image processing.

Boya Zhao was born in 1990. He received the B.S. degree in electronic information engineering from the School of Electrical Engineering and Information, Hebei University of Technology, Tianjin, China, in 2013, and the Ph.D. degree in information and communication engineering from the School of Electrical and Information Engineering, Beijing Institute of Technology, Beijing, China, in 2019.

He is currently an Associate Professor with the Key Laboratory of Computational Optical Imaging Technology, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing. His research interests include object detection in complex background and onboard real-time information processing.

Zijin Li received the B.S. degree in geographic information science from Peking University, Beijing, China, in 2021, and the M.S. degree in signal and information processing from the University of Chinese Academy of Sciences, Beijing, China, in 2024. She is currently working toward the Ph.D. degree in signal and information processing with the Key Laboratory of Computational Optical Imaging Technology, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing.

Her research interests include hyperspectral image processing and object detection.

Fangfang Zhang received the Ph.D. degree in cartography and geographic information system from East China Normal University, Shanghai, China, in 2014.

He is an Associate Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. His research interests include water color remote sensing, development of estimation models for water quality parameters, and software for operational monitoring water quality.

Yuanfeng Wu (Senior Member, IEEE) received the B.S. and M.S. degrees in computer science from the China University of Mining and Technology, Beijing, China, in 2004 and 2007, respectively, and the Ph.D. degree in cartography and geographical information systems from the Graduate University of Chinese Academy of Sciences, Beijing, in 2010.

He is currently a Professor with the Key Laboratory of Computational Optical Imaging Technology, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing. His research interests include the development of onboard real-time algorithms, high-performance computing implementation, and computer software in hyperspectral image processing.