A Learning Framework With Multispectral Band-Differentiated Encoding for Remote Sensing Water Body Detection

Debin Wei, Hongji Xie, Pinru Li, and Yongqiang Xu

***Abstract*—Classic deep convolutional neural network (DCNN) models have demonstrated notable efficacy in segmenting remote sensing images. However, their ability to enhance the precision of water body detection, particularly for smaller ones amid intricate backgrounds, remains challenging. This article proposes the negative Laplacian filter (NLF) method as a solution, enhancing regional color contrast during preprocessing to capture more intricate details effectively. Furthermore, a novel approach employs a differential dual-encoding structure that encodes diverse spectra based on their spectral attributes. Lastly, leveraging prior insights from remote sensing, we introduce the weak label weight adjustment operation for refining predicted images in postprocessing stages. The proposed model significantly outperforms the comparison models on our remote sensing water body dataset.**

***Index Terms*—Multispectral remote sensing, negative Laplacian filtering (NLF), semantic segmentation.**

I. Introduction

DEEP learning-based remote sensing image detection is of practical significance in urban planning, agricultural development, and natural disaster warning. However, compared to tasks such as remote sensing road extraction and building detection, much less research is focused on remote sensing water body detection. This scarcity can be attributed to the challenges associated with preparing datasets and the difficulty in achieving high accuracy in identifying remote sensing water bodies.

Satellites like Landsat-8, Sentinel-2, and Gaofen-2 provide convenient dataset support for remote sensing image detection, including the detection of water bodies. However, the manual cost can significantly increase when dealing with raw, unprocessed materials, and annotating objects as water bodies. The irregular and rough shape of water bodies and the scattered distribution of small water patches due to various factors have posed challenges in pixel-wise labeling. Although the semisupervised learning-based deep neural network (DNN) training method proposed by Protopapadakis et al. [1] greatly reduces the workload of annotating remote sensing datasets, it has not yet attracted the attention of researchers in the field of remote sensing water detection because it does not directly bring convenience to the dataset.

Detecting large lakes, reservoirs, and relatively regular-shaped main watercourses with smooth boundaries is straightforward in remote sensing water body detection. However, automatic water body detection challenges are wider than such strong backgrounds. It becomes more challenging when dealing with targets like narrow rivers and water-filled quarries, which have finer lines or shapes that make extracting features within a smaller receptive field difficult. In addition, different solutes in the water can change its absorption characteristics, causing the water to appear in different colors. Relying solely on visible light RGB band images makes it challenging to reliably distinguish stagnant water with high algae content from surrounding green vegetation and bare rock surfaces on mountains from possible streams flowing over them.

Therefore, to leverage water's absorption and reflection characteristics in different spectral bands, neural networks must learn the features of water bodies in these various spectral bands. While considering the use of information from spectral bands beyond RGB is an excellent approach to complement the spectral characteristics of water bodies, the advantage of multispectral methods over RGB-only approaches appears relatively weak based on the utilization of multiple bands, as seen in [2]. Yuan et al. [3] suggests that this may be because different bands have different resolutions, and the near-infrared (NIR) and short-wave infrared (SWIR) bands need to be upsampled to the exact resolution as the panchromatic (PAN)/RGB bands. This process could potentially compromise the information of the original features. Specific multispectral statistical index methods like the Normalized Difference Water Index (NDWI) [4] have been applied to some extent. However, they tend to perform effectively in coarser-grained scenarios with lower spatial resolutions. When dealing with complex water body scenes or higher

Manuscript received 28 January 2024; revised 9 April 2024 and 24 April 2024; accepted 7 May 2024. Date of publication 10 May 2024; date of current version 23 May 2024. This work was developed by the IEEE Publication Technology Department. This work is distributed under the LaTeX Project Public License (LPPL) (http://www.latex-project.org/) version 1.3. A copy of the LPPL, version 1.3, is included in the base LaTeX documentation of all distributions of LaTeX released 2003/12/01 or later. This work was supported in part by Dalian Youth Science and Technology Star Program under Grant 2023RQ014 and in part by the Interdisciplinary Project of Dalian University under Grant DLUXK-2023-QN-016. *(Corresponding author: Debin Wei.)*

Debin Wei is with the Communication and Network Laboratory, Dalian University, Dalian 116622, China (e-mail: weidebin@163.com).

Hongji Xie is with the Communication and Network Laboratory, Dalian University, Dalian 116622, China (e-mail: lshy030407@163.com).

Pinru Li is with the Communication and Network Laboratory, Dalian University, Dalian 116622, China (e-mail: lipinru1998@163.com).

Yongqiang Xu is with the Communication and Network Laboratory, Dalian University, Dalian 116622, China (e-mail: xyqiang1125@163.com).

Digital Object Identifier 10.1109/JSTARS.2024.3399600

spatial resolutions, methods like NDWI may struggle to achieve efficient segmentation.

To fully exploit the multispectral [5] characteristics of water bodies, we select a multispectral water dataset captured by Sentinel-2, which includes RGB-NIR-SWIR channels. We apply differential filtering to the different channels of the multispectral data within a custom-designed preprocessing module. Furthermore, to mitigate the loss of fine-grained features during the downsampling process in classical CNN networks, we design two distinct encoding pathways and fuse the feature maps from different layers. The fusion results incorporated spatial information from all spectral bands and the relative importance of these bands. As anticipated, our experimental results demonstrate the effectiveness of adapting network models with domain-specific prior knowledge. The contributions of this article are as follows.

1) We introduce an innovative filtering approach known as the negative Laplace filter (NLF). Designed to densely aggregate similar pixels and increase the distance between dissimilar pixel groups, the NLF, particularly when applied as a preprocessing step in low spatial resolution remote sensing imagery, significantly enhances image sharpness and color contrast.
2) We employ separate encoding schemes for RGB color images and NIR-SWIR grayscale images. Through this approach of separate encoding, we are able to better capture additional information of water bodies in the nonvisible light spectrum, thereby obtaining richer detailed features.
3) A weak label weight adjustment (WLWA) module is proposed to refine the preliminary predictions of DNN by seamlessly integrating the a priori distribution patterns of water bodies with deep learning technologies, thereby enhancing the prediction accuracy and reliability of the model.
4) Depth-to-Space (D2S) operation and convolutional block attention module (CBAM) are introduced to fuse information from different channels during the upsampling process effectively. This enhancement contributes to improving model performance and achieving more precise water body segmentation.

The rest of this article is organized as follows. Section II discusses related work, briefly overviewing various research approaches in remote sensing water body detection. Section III describes the employed dataset, data augmentation, and data preprocessing steps. Section IV elaborates on our proposed model, with its effectiveness demonstrated and further discussed in the experimental results presented in Section V, which includes an ablation study. Finally, Section VI concludes this article.

II. Related Work

Before integrating DNNs with remote sensing for water body detection, extensive scientific research had already explored water detection from various angles, including spectral indices and machine learning [6]. However, methods based on deep learning for water body detection offer the capability of end-to-end automated detection with high precision, versatility, and robustness.

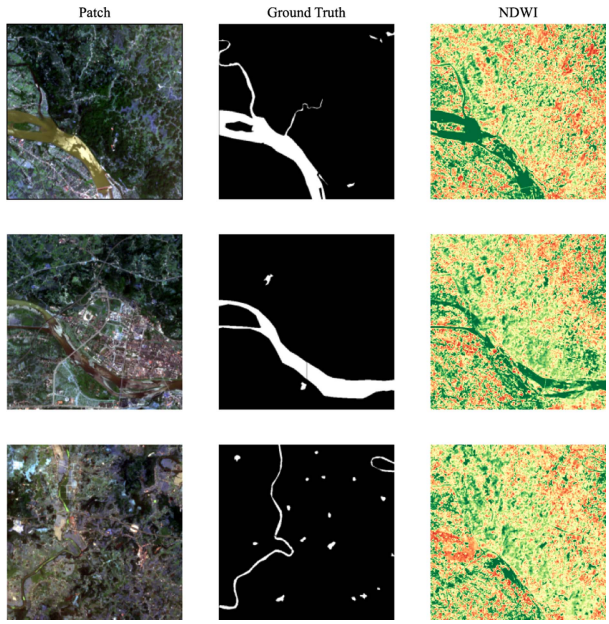


Fig. 1. Heatmaps of NDWI water detection performance under different background intensities. NDWI can basically distinguish water bodies from background objects in the first row. In the second row, NDWI starts to confuse vegetation with water bodies. And in the third row, NDWI can no longer distinguish the shape or contour of water bodies.

A. Detection Methods Based on Spectral Indices

The NDVI was initially introduced by Rouse et al. in 1974 to distinguish between vegetation, soil, and other surface materials. Later, NDVI was also applied to identifying water bodies in some scenarios. The authors in [4] specifically proposed the NDWI for water body identification, based on the absorption and reflection characteristics of water bodies in the visible and infrared spectral bands. Subsequently, a range of improved spectral index methods emerged, including the Modified Normalized Difference Water Index, Weighted Normalized Difference Water Index [7], and Enhanced Water Index [8], aimed at enhancing the applicability of NDWI. Among them, [9] introduced six statistical indices, including the Normalized Difference Moisture Index, Shortwave Infrared Water Stress Index, Normalized Difference Infrared Index, Normalized Difference Shortwave Infrared, Normalized Difference Pond Index, and Normalized Difference Flood Index.

In addition to the practical value of water body detection, from the perspective of detection methods, the introduction and improvement of various statistical indices reflect continuous research and refinement of water body detection by scholars. However, simultaneously, spectral indices' detection methods have limitations, including suboptimal detection performance, limited generality, and sensitivity to the remote sensing image acquisition equipment. Fig. 1 illustrates the water body identification capability of the most representative NDWI among the mentioned statistical indices under different background intensities. It can be observed that as the background intensity becomes weaker, the detection difficulty increases, and NDWI gradually becomes less effective.

B. Detection Methods Based on Machine Learning

Several articles, such as [10], [11], and[12], have explored water body classification using the maximum likelihood method. The authors in [13] employed K-means clustering based on the pixel coordinates of RGB three-band values as 3-D coordinates. In addition, [14] applied Bayesian principles, and [15], [16], and [17] designed support vector machine solutions, all focusing on binary classification of water bodies against other land features. Solbo et al. [18] employed a primitive, fully connected artificial neural network approach for surface water mapping in SAR images. These machine learning-based methods, including but not limited to those mentioned, differ in data supervision levels, algorithm complexity, and processing flow complexity, resulting in varying detection accuracy and applicability.

Compared to purely statistical index-based methods, machine learning, when applied to water body detection, has the advantage of nonlinear modeling capability, allowing it to leverage image context information to improve detection accuracy. Consequently, it possesses specific reasoning capabilities.

C. Detection Methods Based on Deep Learning

Deep learning further enhances the advantages of machine learning in terms of nonlinear modeling, data-driven capabilities, and inference. It reduces the reliance on domain experts. Methods based on deep learning for water body detection, although characterized by more complex model structures and parameters, can be trained and fine-tuned automatically.

For instance, based on the U-Net model, Ch et al. [19] achieved high-precision detection while ensuring model robustness by incorporating a secure feature elliptical digital signature algorithm module to generate digital signatures for the predicted water regions. In [20], a framework based on Attention U-Net and LinkNet ingeniously combined prior knowledge generated by numerical simulators to predict the maximum water levels of floods and the terrain deformations caused by floods and debris flows. A polarization self-attention mechanism was incorporated into D-LinkNet by [21] to reduce information loss during dimensionality reduction, and its outstanding performance was validated on a dataset constructed from Gaofen-2 satellite remote sensing images. In addition to the classic neural networks, which have been repeatedly proven to have excellent scalability and enormous potential, Hong et al. [22] have integrated cutting-edge generative pretraining Transformer (GPT) technology with remote sensing detection. They proposed a powerful SpectralGPT universal large-scale model that effectively utilizes multispectral information. This has a positive impact on the continuous research of multispectral remote sensing water body detection.

III. Data and Data Preparation

A. Data Source

Chengdu City, located in Sichuan Province, China, was chosen as an ideal research area for automatic water body detection in remote sensing due to its abundant water bodies and complex terrain. K. Yuan [3] annotated satellite images covering Chengdu City and its surrounding area, totaling over 15 000 square kilometers, obtained from the Sentinel-2 satellite. They provided a set of RGB-NIR (bands 2, 3, 4, and 8) four-band raster images with dimensions of 20 982 × 20 982 pixels and a bit-depth of 16 b, along with a set of SWIR (band 12) single-band raster images of half the size with the same bit depth. As shown in Table I, the spatial resolution of the four-band images is 10 m, while the single-band images have a spatial resolution of 20 m. The table also includes information about recommended multispectral satellites and some of their parameters.

TABLE I
Spatial Resolutions of Multispectral Satellites, Including the Dataset Used in This Article

Satellite	No. of bands	Spatial resolution(m)
Sentinel-2	13	10 m(2–4&8)
		20 m(5–8,8a,11–12)
		60 m(1,9&10)
Landsat-8	11	30 m(PAN-15 m)
Gaofen-2	5	8 m(PAN-2 m)
World View 2	9	2.07 m(PAN-0.52 m)

B. General Preprocessing and Data Augmentation

1) *Image Segmentation and Edge Padding:* We adopt a sliding window segmentation strategy to consistently segment the RGB-NIR four-band raster images with the binary ground-truth images. The segmentation window size is set to 512 × 512 pixels, allowing the next segmentation window to overlap the current one by half of its area, ensuring an adequate number of samples to prevent overfitting. Considering channel alignment, the registered SWIR single-band raster images differ from the above segmentation by having patches of size 256 × 256. For patches of nonstandard sizes generated when the sliding window reaches the rightmost or bottommost parts of the raster images, we retain them and perform edge padding with constant values.
Fig. 2 shows a randomly selected image and pixel value distribution map for each channel before normalization. The pixel value distributions for this image and the dataset as a whole approximately follow a Gaussian distribution [3]. Therefore, we apply Z-score standardization to each channel across the entire dataset (as in (1)) and then perform min–max normalization (as in (2)).

$$x' = \frac{x - \mu(x)}{\sigma(x)} \tag{1}$$

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}. \tag{2}$$

Here, the mean and standard deviation values $\mu(x)$ and $\sigma(x)$ are determined for this dataset, not ImageNet. $\min(x)$ and $\max(x)$ represent the minimum and maximum values for the channels of the pixel point x.

2) *Channel Separation and Color Normalization:* Our model processes the visible light spectrum and nonvisible spectral bands through separate pathways, employing distinct processing strategies for each. Therefore, we divide a

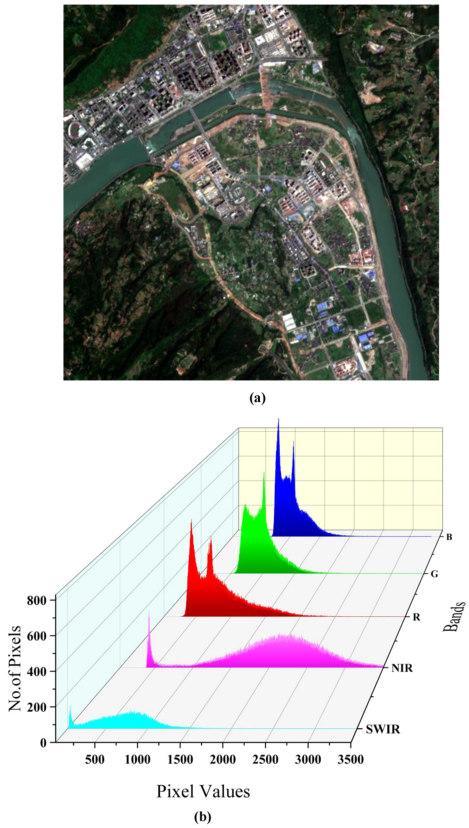


Fig. 2. Pixel distribution map of the five bands corresponding to the randomly chosen patch. The bands are arranged near and far as SWIR, NIR, R, G, and B. Except for NIR, which has a long and narrow tail beyond a pixel value of 3500, pixel values in the other bands are sparsely scattered across a sizeable numerical range (theoretically, 16-b depth can reach 65 535, and the maximum value in the original patch's statistics does not exceed 16 000). Therefore, we choose 3500 as the visualization threshold. In practical operations, to some extent, we applied a 2% linear stretch to reduce the impact of these sparse values on normalization. (a) Patch. (b) Pixel values distribution.

patch of the original RGB-NIR four-band dataset into two equal-sized patches: One containing RGB three bands and the other containing the NIR single band. This separation helps avoid potential complications in the subsequent steps. Even after segmentation and channel separation, the patches maintained a 16-b depth. However, to better visualize the dataset and the convenience of inputting RGB images separately into the comparative network, we compress the bit depth to 8 b (making the images viewable on a personal computer without needing specialized remote sensing software) and perform color normalization.

3) *Sample Rough Screening and Channel Alignment:* Considering the challenges in annotating water bodies, we remove samples with annotation errors resulting from subjective misjudgments during labeling, as illustrated in Fig. 3, where the red-marked areas exhibit clear labeling errors. Unlike RGB-Only datasets, multispectral datasets require that different bands reflecting the same geographical area be pixel-wise aligned. Hence, we conducted a one-time check on the channel alignment of the dataset.

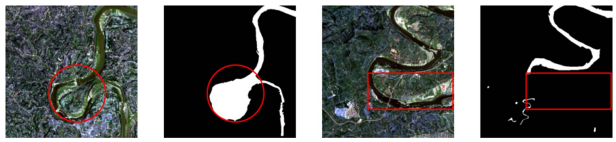


Fig. 3. Samples with erroneous labels.

4) *Image Augmentation:* Before the data was loaded into the network model, during the training phase, we applied four independent data augmentation operations with a probability of 0.5 to the images: Random horizontal flipping, random vertical flipping, random scaling, and random HSV color space transformations.

IV. Multispectral Band-Differentiated Encoding Network

A. Model Structure

Fig. 4 illustrates the architecture of our proposed Multispectral Band-Differential Encoding Network (MBDEN). The network structure primarily follows an encoder-decoder framework. The RGB, NIR, and SWIR channels are simultaneously fed into the network but encoded in separate groups, according to Fig. 4. Before entering the encoder, RGB channels undergo the NLF preprocessing. In contrast, the NIR channel generates a weak label map for fine-tuning the prediction's weight during the WLWA operation. RGB encoder is based on the pretrained ResNet34 [23] backbone from ImageNet, whereas the encoding of the grayscale channel for NIR-SWIR adopts an improved VGG13 [24] structure. Unlike the RGB encoder, the grayscale encoder does not undergo pretraining, but it remains consistent in spatial dimensions and depth. Both pathways perform pixel-wise feature fusion in the later decoding stage, followed by multiscale feature extraction through the Atrous Spatial Pyramid Pooling (ASPP) module. In the fusion-decoding phase, we utilize the D2S operation to restore feature map dimensions while reducing channel dimensions. Furthermore, we employ CBAM to integrate low-level feature maps seamlessly with decoded images. We elaborate on some of these crucial details as follows.

1) *Pretrained or not:* In image segmentation, traditional and well-established networks, even if their original development might not have been intended for water body segmentation or image segmentation, predominantly rely on RGB-only information. Take, for instance, ResNet34, a classic deep-learning network model known for effectively extracting everyday object features like color and texture. Consequently, the RGB channel coding with a pretrained backbone network accelerates model convergence significantly and provides relatively reliable feature maps from various depths. Conversely, nonpretrained backbone architectures aim to extract features from the NIR-SWIR channels, which are challenging to capture using conventional encoding methods, thus complementing ResNet34's encoding results with fine-grained details.
2) *Weak Label Weight Adjustment:* Common water bodies exhibit light solid absorption characteristics in the NIR

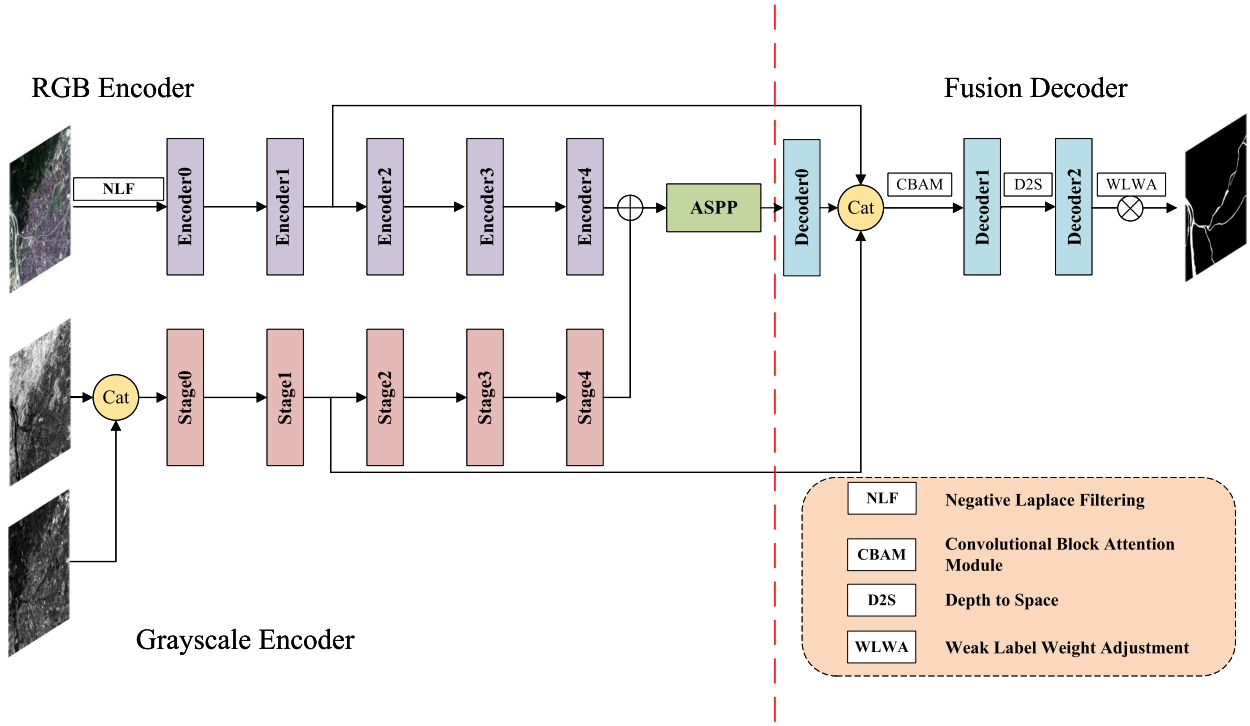


Fig. 4. Proposed MBDEN network architecture.

and SWIR channels, resulting in lower pixel values, which is crucial prior knowledge. As mentioned in Section II-A, various statistical index-based detection methods attempt to maximize the utilization of such prior knowledge. However, none of these statistical index methods can achieve remarkable results when combined with deep learning as we do. Fig. 5 displays a NIR grayscale image and its pixel value visualization heatmap, where the heatmap on the right has undergone pixel inversion. Rough segmentation can differentiate between water and nonwater areas by applying an appropriate pixel threshold (e.g., 220 in the case of Fig. 5). We refer to the thresholded NIR image as a weak label map, where water areas are marked as 1 and nonwater areas as 0. The fusion decoding module, Decoder2, restores the image size to align with a single-channel image of equivalent dimensions to the input RGB image. We refer to this image as the pre-softmax image. Traditionally, the pre-softmax image is followed by a softmax operation. In contrast, we perform pixel-wise weighting based on the weak label map before softmax. Positions marked as 1 in the weak label map are assigned a weight greater than 1, while those marked as 0 are assigned a weight less than 1. These weights are then used to multiply the pre-softmax image pixel-wise. In practice, we employ binary weighting with values of 1.3 and 0.8. We describe the process of WLWA using the following formula:

$$s(x,y) = s(x,y)_{\text{pre}} w(x,y) \tag{3}$$

where, $s(x,y)$ represents the value at row x and column y in the prediction matrix, $s(x,y)_{\text{pre}}$ represents the value at row x and column y in the pre-softmax matrix. $w(x,y)$ represents the weight of the weak label at row and column in the matrix. The weight adjustment equation is as follows:

$$w(x,y) = \begin{cases} 1.3, & N(x,y) \geq THR \\ 0.8, & N(x,y) < THR \end{cases} \tag{4}$$

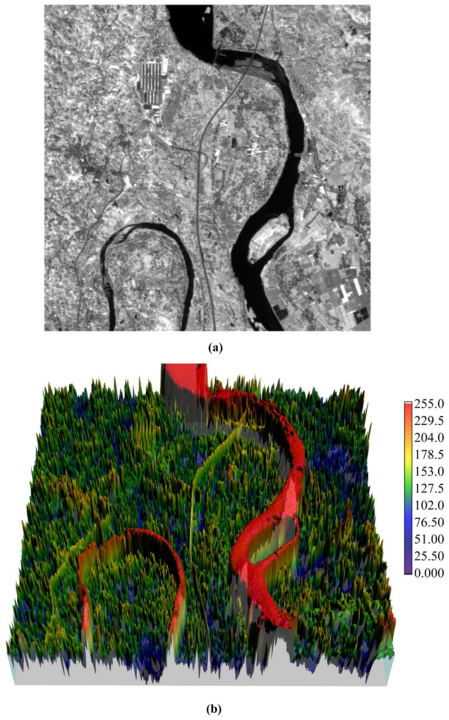


Fig. 5. Pixel value visualization heatmap. (a) NIR in gray. (b) NIR in heatmap.

where $N(x, y)$ represents the value at row x and column y of the pixel matrix in the NIR band. $THR \in (0, 255)$ represents the threshold within the NIR band that roughly distinguishes between water and nonwater bodies.

3) *Negative Laplacian Filtering*: The traditional Laplacian filter, while enhancing boundary contrast, eliminates color-consistent regions and retains only partial boundaries, making it challenging for effective application in remote sensing. To address this, we design the NLF kernel, which preserves boundaries while increasing the contrast of adjacent color regions.

The convolution operation involves applying the NLF kernel to the pixel matrix of the input image. The convolution formula is as follows:

$$s(x, y) = f * g = \sum_{h=x-1}^{x+1} \sum_{k=y-1}^{y+1} f(h, k)g(x - h, y - k) \quad (5)$$

where $s(x, y)$ represents the value at the row x and column y of the output matrix s, f represents the NLF kernel, and g represents the pixel matrix of the image being filtered. Equation (5) describes the principle of enhancing image details through the determination of grayscale steps when performing second-order differencing for image sharpening. The second-order difference expression is given by

$$\nabla^2 f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2}$$
$$\nabla^2 f(x, y) = 5f(x, y) - f(x, y - 1) - f(x - 1, y) - f(x + 1, y) - f(x, y + 1). \quad (6)$$

To ensure that the image dimensions remain unchanged before and after filtering the pixel matrix based on (5)–(6) for grayscale step filtering, we first perform edge padding. Suppose the unpadded pixel matrix is as follows:

$$\begin{bmatrix} 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}. \quad (7)$$

The pixel matrix after edge padding is

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \quad (8)$$

After padding, the matrix undergoes the NLF operation, and the color adjustment of the color regions is performed using a color truncation mechanism. The adjustment process is as follows:

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \otimes \begin{bmatrix} 0 & -1 & 0 \\ -1 & 5 & -1 \\ 0 & -1 & 0 \end{bmatrix}$$
$$= \begin{bmatrix} 0 & 0 & -1 & 3 & 2 & 2 & 3 \\ 0 & 0 & -1 & 2 & 1 & 1 & 2 \\ 0 & 0 & -1 & 2 & 1 & 1 & 2 \\ -1 & -1 & -1 & 3 & 2 & 2 & 3 \\ 3 & 2 & 3 & -2 & -1 & -1 & -1 \\ 2 & 1 & 2 & -1 & 0 & 0 & 0 \\ 3 & 2 & 3 & -1 & 0 & 0 & 0 \end{bmatrix} \quad (9)$$
$$\begin{bmatrix} 0 & 0 & -1 & 3 & 2 & 2 & 3 \\ 0 & 0 & -1 & 2 & 1 & 1 & 2 \\ 0 & 0 & -1 & 2 & 1 & 1 & 2 \\ -1 & -1 & -1 & 3 & 2 & 2 & 3 \\ 3 & 2 & 3 & -2 & -1 & -1 & -1 \\ 2 & 1 & 2 & -1 & 0 & 0 & 0 \\ 3 & 2 & 3 & -1 & 0 & 0 & 0 \end{bmatrix}$$
$$\propto \begin{bmatrix} 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 \end{bmatrix} \quad (10)$$

where $\otimes$ denotes the convolution operation, the left and right matrices correspond to the pixel matrix of the filtered image and the NLF kernel, respectively. $\propto$ symbolizes the truncation of color features in the resulting matrix on the left-hand side after the convolution operation, i.e., the postfiltered image.

The values 0 and 1 in the matrix represent two colors, A and B, respectively. A higher count of values smaller than 0 indicates a darker A color, while a higher count of values greater than 1 indicates a deeper B color. The size of the NLF kernel is designed such that the central value in its 3 $\times$ 3 grid must exceed 4. The values on the four sides–top, bottom, left, and right–are sparsely distributed as "0, –1, 0." The central value of the NLF kernel must surpass the sum of values in the sparse positions to maintain color stability within the region. It is important to note that in (9), the difference in values at the boundaries of different-colored regions increases before and after matrix filtering, while the color characteristics within the same colored regions remain stably preserved.

The truncation mechanism activates when pixel values exceed the predefined standard color space, such as when

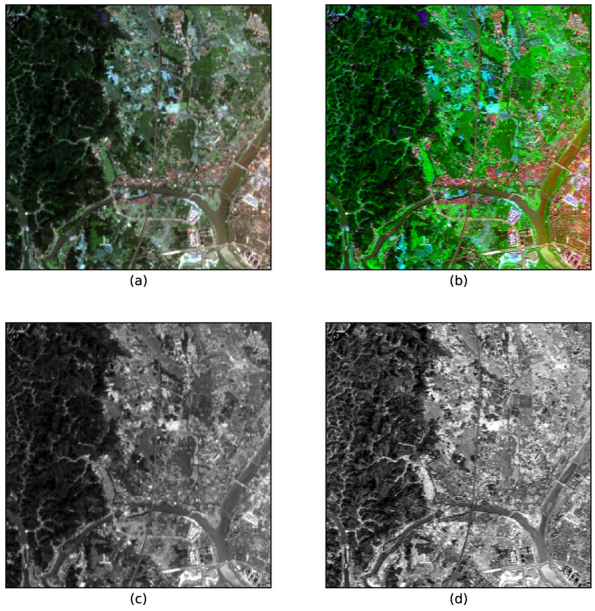


Fig. 6. NLF effect. (a) is the RGB image, (b) is the RGB image after NLF, (c) is the G channel of the RGB image, (d) is the G channel of the RGB image after NLF.

0 signifies standard black and 1 signifies standard white. To elaborate

$$s(x,y) = \begin{cases} 1, & s(x,y) \geq 1 \\ 0, & s(x,y) \leq 0. \end{cases} \tag{11}$$

Fig. 6 shows the effectiveness of NLF operation. It can be seen that the clarity and color contrast of remote sensing images are significantly improved in the visible range after NLF, and the texture and lines are also clearer. This undoubtedly helps the network model to extract small and easily overlooked features.

4) We have included the main parameters for each layer in Table II. It is worth noting that before concatenating with SWIR, the NIR channel undergoes a 2× downsampling operation. The Decoder1 module performed a 1×1 convolution operation to reduce the 384-channel concatenation of lower-level features to 128 channels. The output of the Decoder2 module, an eight-channel image, is further processed with a 1×1 convolution operation to reduce the number of channels to 1, corresponding to the pre-softmax image required for the WLWA operation.

B. Loss Function

We adopt a combination of binary cross-entropy (BCE) and Dice coefficient as the loss function. The specific calculations for these two components are given by (12)–(13).

$$loss_{\text{BCE}} = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)\right] \tag{12}$$

$$loss_{\text{Dice}} = 1 - \frac{2\left|Y \cap \hat{Y}\right| + 1}{|Y| + \left|\hat{Y}\right| + 1} \tag{13}$$

TABLE II
LAYER CONFIGURATION OF EMPLOYED NETWORK ARCHITECTURE

Layer		Output size	
RGB	NIR	(512,512,3)	(512,512,1)
	SWIR		(256,256,1)
NLF	Concatenate	(512,512,3)	(256,256,2)
Encoder0	Stage0	(128,128,64)	(256,256,32)
Encoder1	Stage1	(128,128,64)	(128,128,64)
Encoder2	Stage2	(64,64,128)	(64,64,128)
Encoder3	Stage3	(32,32,256)	(32,32,256)
Encoder4	Stage4	(16,16,512)	(16,16,512)
Add		(16,16,512)	
ASPP		(16,16,1280)	
Decoder0		(128,128,256)	
Concatenate		(128,128,384)	
Decoder1		(128,128,128)	
Decoder2		(512,512,8)	

where n represents the number of pixels in the predicted image, y_i is the label value of the ith pixel (0 or 1), and $\hat{y}_i$ is the predicted value of the ith pixel; Y represents the regions marked as 1 in ground truth, and $\hat{Y}$ represents the regions marked as 1 in predicted map.

It can be observed that (12) primarily focuses on the correspondence of each pixel value in binary classification problems. In contrast, (13) mainly assesses the overlap between the predicted and labeled images. The linear combination of these two equations forms the desired loss function, calculated as

$$loss = \alpha loss_{\text{BCE}} + (1-\alpha) loss_{\text{Dice}}. \tag{14}$$

It is important to note that we do not have sufficient reasons to determine which of the two components of the loss function contributes more significantly. Therefore, the common practice is to choose a weight factor $\alpha = 0.5$ in (14), such that both the $loss_{\text{BCE}}$ and the $loss_{\text{Dice}}$ have equal importance. Similarly, in our practical implementation, we specify equal weights of 1 for both components.

V. EXPERIMENTAL RESULTS AND ANALYSIS

A. Experimental Setup

Our dataset comprises 5214 training samples and 1306 testing samples, each size 8×8. Each sample comprises three patches for the RGB, NIR, and SWIR channels, along with their corresponding binary label maps. All experiments involving classic neural network models for method comparison were conducted on two NVIDIA GTX 3090 GPUs with a batch size set to 8. We implement an early stopping mechanism, which triggers when the loss value has not decreased continuously for seven consecutive times. Otherwise, training continues for a maximum of 300 epochs. The initial learning rate is set to 2e-4, with a decay rate of 0.2 applied after four consecutive nondecreasing loss values, down to a minimum of 5e-7.

TABLE III
IoU Results on Test Set for All Models

Models	ImageNet pre-trained	precision/ [%]	recall/ [%]	IoU/ [%]
U-Net[25]	no	68.93	50.99	41.46
LinkNet[26]	yes	78.16	71.96	59.91
D-LinkNet[27]	yes	81.62	75.34	64.41
DeepLabV3+[28]	yes	83.61	76.90	66.83
MACU-Net[29]	no	77.85	70.87	58.98
MC-WBDN[3]	yes	84.57	77.62	68.14
Ours	yes&no	**87.47**	**82.71**	**73.95**

B. Experimental Results

Table III presents the performance comparison results of various networks on the test set. Since it's a binary classification problem, we use IoU instead of mIoU as the primary evaluation metric for better consistency with human perception.

It should be noted that MC-WBDN is the only network model in the table, apart from ours, that employs a multispectral strategy. We annotate "yes & no" in the "ImageNet pretrained" column to highlight the differences in the two encoding pathways.

As can be seen from Table III, there is a significant performance difference among different network models, and pretrained models generally outperform those that are not pretrained. Except for the U-Net model, which performs noticeably worse as a classic reference model, LinkNet slightly outperforms MACUNet. At the same time, D-LinkNet, DeepLabV3+, and MC-WBDN show a stepwise improvement in performance. Our model exceeds the IoU of the above network models by at least 5.81%. The accuracy rate is 2.9%, and the accuracy rate is 5.09%. The significant lead of D-LinkNet over LinkNet and DeepLabV3+ suggests that multilevel dilated convolution strategies, represented by the ASPP module, effectively extract features of the same object at different scales, closely linking spatial information and context, which is crucial for the demanding task of water body detection. The high performance of MC-WBDN also supports the idea that multispectral strategies can effectively supplement context with fine details.

Fig. 7 displays the comparative prediction performance of different models on patches with varying background intensities. These examples show that water shapes or colors are easily distinguishable when dealing with a strong background, and the segmentation results among different network models are relatively similar. This implies that even with the original design of models like U-Net, good performance can still be achieved. However, the significant performance gap between the U-Net model and other models cannot be solely attributed to weak background scenarios. For example, in the first row of patches where the main river channel's edge is located, the segmentation result of the U-Net model is coarser compared to other models.

The patches in rows 1 and 3 of Fig. 7 represent water distribution in densely populated areas and between farmlands. These two rows showcase our model's exceptional capability in handling small water bodies. Whether it is the tributaries hidden amidst green vegetation and modern buildings in row 1 or the stream seamlessly blending with the green background in the third row, our model outperforms other models by a significant margin. Of course, the MC-WBDN model, which also utilizes a multispectral strategy, performs second best after ours, indicating the positive effect of additional NIR and SWIR spectral information. It is worth noting that there are many small water puddles in the third-row patch, and other models exhibit noticeable issues, such as false negatives and significant shape distortions, which are less prominent in our model.

What is particularly remarkable is that, compared to other models, ours demonstrates less susceptibility to overinterpretation. In row 2 of Fig. 7, two artificial water storage facilities are built along the river on the left side of the river's main channel. These facilities are typically used for freshwater aquaculture, resulting in a narrow gap between them and the main river channel. However, examining all models closely, only ours accurately captures the gap between these two water bodies. In other words, all other models, despite their beneficial feature extraction efforts, tend to overinterpret during the decoding phase. We believe that a more refined upsampling helps prevent this "overinterpretation." However, dense operations like transpose convolution for upsampling introduce higher computational costs. Therefore, we achieve a compromise by utilizing the D2S operation. Meanwhile, weak label weight adjustment also positively affects similar scenarios.

The patch in row 4 of Fig. 7 was artificially intervened to test the performance stability of various models in dealing with weather conditions such as clouds and fog. We notice that the performance differences among the models in this situation are generally consistent with their performance under normal conditions. Our model demonstrate the same level of stability in maintaining connectivity as usual. In introducing Fig. 1 of Section II, we illustrate the problem that methods based on exponential statistics do not work in weak backgrounds. In contrast, deep learning methods remain effective even in highly challenging weak background conditions, such as rows 5 and 6 in Fig. 7. In these cases, our model exhibits superior performance. In row 5, the scene is so challenging that the human eye can hardly distinguish the stream from the background. In the sixth row, the scene involves a mountain road intertwined with a stream, which shares similar colors and morphological features.

C. Ablation Study

To better illustrate the contributions of the introduced components, we study three main modules, namely, dual-encoder, NLF, and WLWA, using DeepLabV3+ as the baseline. The results are presented in Table IV.

From Table IV, it can be observed that the dual-encoder module made the most significant contribution to our ablation study, indicating a beneficial role of the grayscale encoding pathway in enhancing feature extraction for RGB color images, especially in detail augmentation. Our analysis suggests that, although color images and grayscale images are fundamentally pixel value matrices, different channels exhibit relatively stable

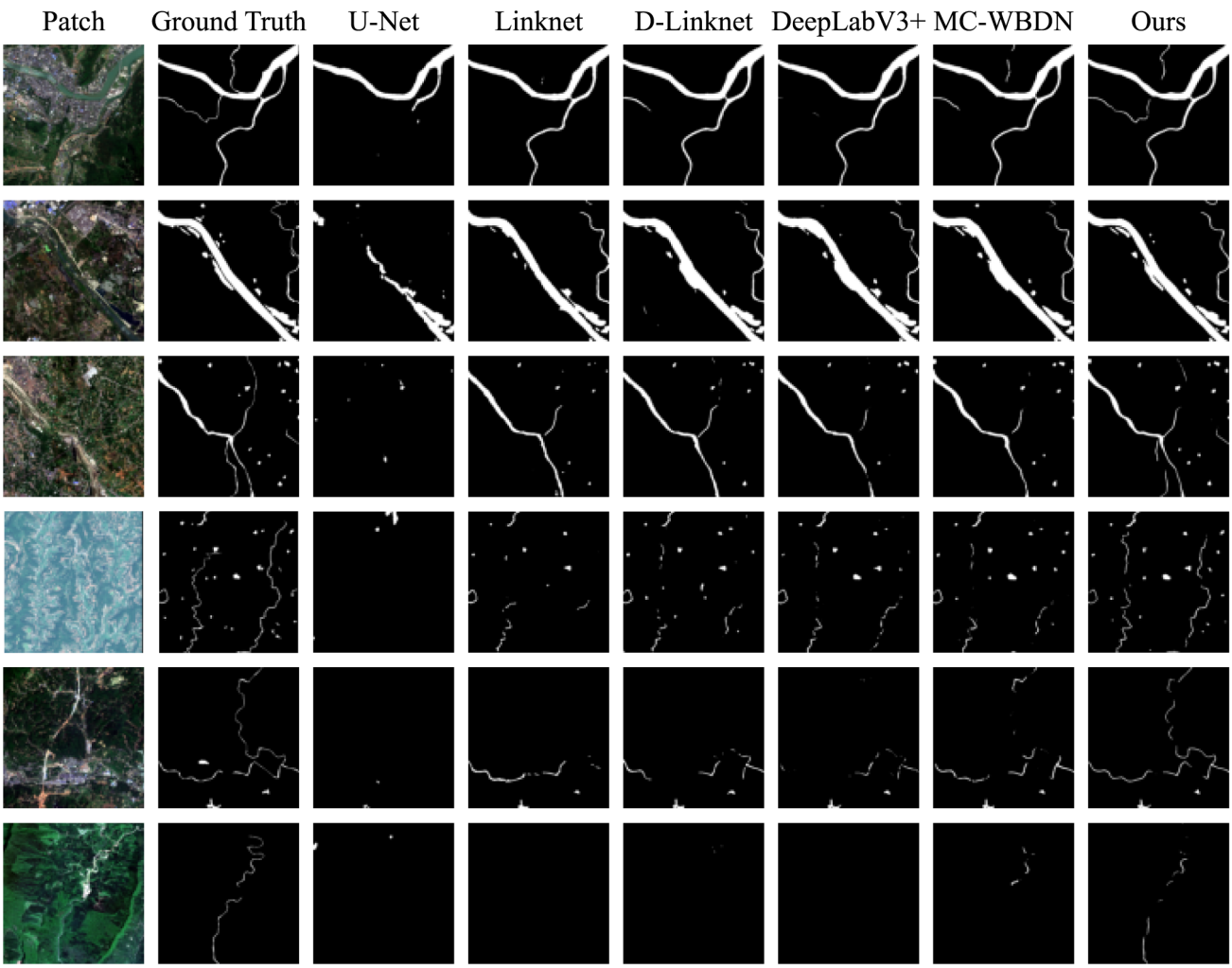


Fig. 7. Test results for different models on example patches.

TABLE IV
RESULTS OF ABLATION STUDY

Dual-encoder	NLF	WLWA	IoU/[%]
			66.83
✓			70.19
	✓		67.72
		✓	68.63
✓	✓		71.34
✓		✓	71.47
	✓	✓	68.82
✓	✓	✓	72.41

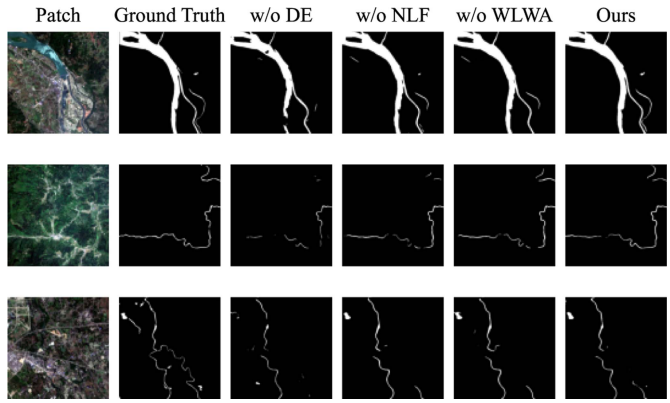


Fig. 8. Ablation results of main innovative points.

correlations when subjected to high-dimensional mapping within neural networks. This correlation implies specific requirements on the number and sequence of channels in traditional networks, particularly those with pretrained encoders.

Fig. 8 illustrates the ablation studies of the three main innovations from Table IV in three different scenarios. Here, "w/o" stands for "without." It is evident that the contributions of each module to the detection of small water bodies are positive, and the contributions of each module are highly consistent with the quantitative results.

In Section IV-A, we explain how NLF works, and indeed, NLF does lead to a performance improvement of nearly 1% in our model. However, this improvement is lower than we expect,

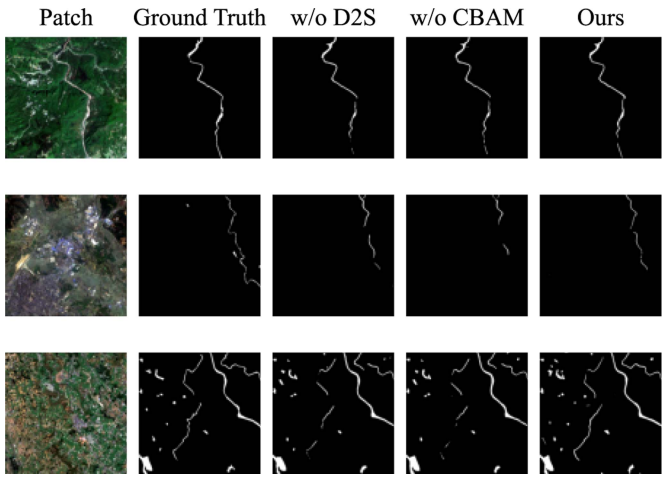


Fig. 9. Ablation results of main auxiliary module.

TABLE V
RESULTS OF NLF STUDY

Methods	IoU/[%]
U-Net + NLF	↑ 6.77
LinkNet + NLF	↑ 1.77
D-LinkNet + NLF	↓ 1.30
DeepLabV3+ + NLF	↓ 0.83
MACUNet + NLF	↑ 3.13

considering the noticeable effect of NLF in Fig. 6. Table V shows that we further apply NLF to several other models to investigate why.

Combining the information from Tables III and V, it is evident that the NLF yields significant improvements for U-Net and MACUNet, which have not undergone pretraining. However, its effectiveness varies when applied to models like D-LinkNet and DeepLabV3+ that already employ multiscale fusion strategies. Since we plan to choose the encoding scheme of DeeplabV3+ (backbone network using ResNet34) as one of our encoding pathways, the results in Table V are unfavorable. Therefore, we conduct multiple experiments and ultimately apply the NLF preprocessing operation to our network with a probability of 0.3. This decision accounts for the nearly 1% performance improvement in Table IV.

When observing the results in Fig. 5, we believe many would be tempted to assume that effective water body segmentation can be achieved using only NIR or SWIR data. This initial motivation is incorporating the WLWA module to adjust the network's output. However, the reality is that there is not a single optimal threshold that can effectively partition all positive samples in a dataset containing such complex scenes. First, the difference in pixel value distributions between water and background objects exists within a relatively flexible range, which can often fluctuate due to changes in lighting conditions. Second, colored metals and buildings constructed with these metals, tall landforms, and shadows from buildings can exhibit highly similar absorption patterns to water bodies in the NIR and SWIR bands. As a result, our WLWA module only led to a modest performance improvement of 1.8% in our model.

TABLE VI
RESULTS OF ABLATION STUDY (B)

D2S	CBAM	IoU/[%]
		72.41
✓		73.19
	✓	73.44
✓	✓	73.95

Based on the baseline integration of the above three modules, we test the effectiveness of the D2S module and CBAM module. The quantitative and qualitative results are provided in Table VI and Fig. 9, respectively. Table VI demonstrates that the D2S operation can replace bilinear interpolation in specific scenarios and offer some optimization benefits. In addition, CBAM effectively refines and enhances the low-level feature channels from different encoding pathways and the results of upsampling in the model. In Fig. 9, a visual comparison of the two auxiliary modules is presented, with the example in the second row demonstrating that CBAM makes a greater contribution to the consideration of channel information in terms of line continuity and detection completeness.

D. Discussion

In this article, we introduce a novel DNN model for remote sensing water body detection, dubbed MBDEN. The proposed model outperforms benchmark models in pixel-level classification of water bodies across diverse scenarios with varying intensities. We particularly highlight MBDEN's superior performance in detecting small water bodies (with pixel widths of only a few pixels), a feat closely tied to our targeted design strategy. In crossing disciplinary boundaries between deep learning techniques and remote sensing detection, it is common to validate the efficacy of cutting-edge technologies against existing baselines and then investigate the reasons behind their effectiveness. Contrarily, this study begins with an understanding of the prior knowledge on remote sensing water bodies, acknowledging the issues of low spatial resolution in remote sensing datasets and the underutilization of multispectral information by existing networks.

We recognize that the spatial resolution issue of remote sensing water bodies is more of a practical challenge than a technical one. Hence, the employed datasets strive to maintain clarity while achieving high ground coverage. A clearer input image undoubtedly provides a better starting configuration for the DNN's gradient descent process. Initially, we aimed to employ multiple filters with nonzero elements distributed along the directions $y=0$, $x=0$, $y=x$, and $y=-x$ for preprocessing, to give linear water bodies greater initial weight in these directions. This approach was meant to make slender water bodies, which are typically challenging to detect, more noticeable to the neural network. However, due to the sliding window mechanism, the filtering effects of adjacent receptive fields interfered with each

other, leading to a further degradation of the linear filtering results due to noise accumulation. Furthermore, designing these linearly initialized filtering kernels as weight-learnable initial convolution kernels proved ineffective. In contrast, our designed NLF filtering kernels counteract the filtering interference from the four directions (up, down, left, right) due to the sparsity at the edges, while the central value ensures the color stability of the inner uniform areas, thereby enhancing color contrast and line clarity.

Similar to the development of the NLF, we proposed a dual-channel separate encoding approach and the WLWA module from the perspective of fully leveraging the multispectral information of water bodies, culminating in a remote sensing water body detection framework that encompasses NLF preprocessing, differentiated encoding, and weight-adjusted postprocessing. Recognizing the usefulness of information carried by nonvisible light bands, it was intuitive to also use the more relevant NIR-SWIR as input channels, providing additional valuable information. However, the strict requirements of pretrained feature extractors on the number of convolutional kernels per network layer led to the popular approach of designing fusion heads to handle early-stage channels, discarding the ResNet34's initial convolutional design for three-channel input [3]. This conventional design method overlooks the inherent correspondence between the channels of frozen layers and those of the preceding layer or even the initial input channels. Therefore, separate encoding for NIR-SWIR not only maintains the advantages of pretrained shallow networks in extracting fundamental features such as edges and textures from RGB images, but also offers flexibility in feature fusion across different layers due to the freedom in channel design.

The conception of WLWA was also not coincidental. The introduction of various statistical indices in Section II-A has proven the superior performance of multispectral information consideration, like NDWI, in water body detection over any single spectrum. However, we found adjusting prediction weights based on NDWI's prior knowledge to be unstable and computationally expensive due to division operations. We ultimately opted for WLWA based on the NIR single band, adjusting weights for both water and nonwater parts. Adjusting weights for water bodies alone is viable experimentally, and we speculate this method offers better generalization across different water body scenarios. Yet, for the dataset used, adjusting both water and nonwater parts yielded better performance.

VI. Conclusion

In this article, we have developed a multispectral water body detection network model based on deep learning. There are two main differences between this model and existing deep learning-based methods. First, it enhances image clarity and contrast by autonomously proposing the NLF, thus alleviating the problem of image quality degradation caused by insufficient spatial resolution. Second, to better utilize multispectral information, the model introduces band-differentiated encoding methods and NIR weight adjustment measures. Combining these contributions, we incorporate the D2S operation and the widely used and practical CBAM module [30], [31], [32], [33]. These enhancements significantly elevate the water body detection performance of the proposed MBDEN model to a high level. The experiments have demonstrated that our model excels in detecting small water bodies and performs exceptionally well under weak background conditions. In future work, we will focus on the encoding part of the proposed framework and attempt to modify the encoding architecture using neural architecture search algorithms [34], [35]. We will optimize the encoding scheme by integrating the idea of stacking CNN encoding structures [1] with ViT [36], [37] technology. At the same time, in terms of model training, we draw on SpectralGPT [22], multimodal remote sensing dataset technology [38], and the idea of self supervised learning [39] to improve the quality of remote sensing datasets. Finally, we will propose a remote sensing target extraction network with excellent performance and generalization ability.

Acknowledgment

The opinions expressed here are entirely that of the author. No warranty is expressed or implied. User assumes all risk.

References

[1] E. Protopapadakis, A. Doulamis, N. Doulamis, and E. Maltezos, "Stacked autoencoders driven by semi-supervised learning for building extraction from near infrared remote sensing imagery," *Remote Sens.*, vol. 13, no. 3, 2021, Art. no. 371.

[2] J. Zhang et al., "Water body detection in high-resolution SAR images with cascaded fully-convolutional network and variable focal loss," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 316–332, Jan. 2021.

[3] 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.

[4] 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.

[5] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, "An augmented linear mixing model to address spectral variability for hyperspectral unmixing," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1923–1938, Apr. 2019.

[6] T. Bijeesh and K. Narasimhamurthy, "Surface water detection and delineation using remote sensing images: A review of methods and algorithms," *Sustain. Water Resour. Manage.*, vol. 6, pp. 1–23, 2020.

[7] M. Malviya et al., "Correlates of physiological stress and habitat factors in reintroduction-based recovery of tiger (panthera tigris) populations," *Hystrix*, vol. 29, no. 2, 2018, Art. no. 195.

[8] E. Trochim, A. Prakash, D. Kane, and V. Romanovsky, "Remote sensing of water tracks," *Earth Space Sci.*, vol. 3, no. 3, pp. 106–122, 2016.

[9] M. Boschetti, F. Nutini, G. Manfron, P. A. Brivio, and A. Nelson, "Comparative analysis of normalised difference spectral indices derived from modis for detecting surface water in flooded rice cropping systems," *PLoS One*, vol. 9, no. 2, 2014, Art. no. e88741.

[10] P. S. Frazier et al., "Water body detection and delineation with landsat TM data," *Photogrammetric Eng. Remote Sens.*, vol. 66, no. 12, pp. 1461–1468, 2000.

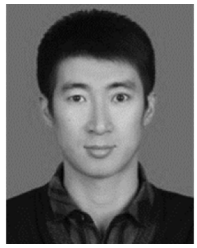
[11] N. Pierdicca, M. Chini, L. Pulvirenti, and F. Macina, "Integrating physical and topographic information into a fuzzy scheme to map flooded area by SAR," *Sensors*, vol. 8, no. 7, pp. 4151–4164, 2008.

[12] M. S. Hassan and S. Mahmud-Ul-Islam, "Detection of waterlogging areas based on passive remote sensing data in jessore district of khulna division, Bangladesh," *Int. J. Sci. Res. Pub.*, vol. 4, 2014, Art. no. 12.

[13] H. Gao, L. Wang, L. Jing, and J. Xu, "An effective modified water extraction method for landsat-8 oli imagery of mountainous plateau regions," *Proc. IOP Conf. Ser.: Earth Environ. Sci.*, vol. 34, no. 1, 2016, Art. no. 012010.

[14] C. Verpoorter, T. Kutser, and L. Tranvik, "Automated mapping of water bodies using landsat multispectral data," *Limnol. Oceanogr.: Methods*, vol. 10, no. 12, pp. 1037–1050, 2012.

[15] S. Abd Manaf, N. Mustapha, M. N. Sulaiman, N. A. Husin, and M. R. A. Hamid, "Comparison of classification techniques on fused optical and sar images for shoreline extraction: A case study at northeast coast of peninsular Malaysia," *J. Comput. Sci.*, vol. 12, no. 8, pp. 399–411, 2016.
[16] C. Huang, B. D. Nguyen, S. Zhang, S. Cao, and W. Wagner, "A comparison of terrain indices toward their ability in assisting surface water mapping from Sentinel-1 data," *ISPRS Int. J. Geo- Inf.*, vol. 6, no. 5, Art. no. 140, 2017.
[17] A. Paul, D. Tripathi, and D. Dutta, "Application and comparison of advanced supervised classifiers in extraction of water bodies from remote sensing images," *Sustain. Water Resour. Manage.*, vol. 4, pp. 905–919, 2018.
[18] S. Solbo, E. Malnes, T. Guneriussen, I. Solheim, and T. Eltoft, "Mapping surface-water with radarsat at arbitrary incidence angles," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. Proc.*, 2003, vol. 4, pp. 2517–2519.
[19] A. Ch, R. Ch, S. Gadamsetty, C. Iwendi, T. R. Gadekallu, and I. B. Dhaou, "ECDSA-based water bodies prediction from satellite images with UNet," *Water*, vol. 14, no. 14, 2022, Art. no. 2234.
[20] N. Yokoya et al., "Breaking limits of remote sensing by deep learning from simulated data for flood and debris-flow mapping," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4400115.
[21] X. Chang, B. Deng, Z. Bao, X. Guo, and F. Yuan, "A modified D-linknet for water extraction from high-resolution remote sensing," in *Proc. IEEE 5th Int. Conf. Adv. Electron. Mater., Comput. Softw. Eng.*, 2022, pp. 151–156.
[22] D. Hong et al., "SpectralGPT: The first remote sensing foundation model customized for spectral data," Jan. 2024, doi: 10.5281/zenodo.10533809.
[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, 2015.
[25] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput.-Assist. Interv.: 18th Int. Conf., Proc., Part III 18*, Springer, 2015, pp. 234–241.
[26] A. Chaurasia and E. Culurciello, "Linknet: Exploiting encoder representations for efficient semantic segmentation," in *Proc. IEEE Vis. Commun. Image Process.*, 2017, pp. 1–4.
[27] L. Zhou, C. Zhang, and M. Wu, "D-linknet: Linknet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 182–186.
[28] 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.
[29] R. Li et al., "Multiattention Network for Semantic Segmentation of Fine-Resolution Remote Sensing Images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2022, doi: 10.1109/TGRS.2021.3093977.
[30] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.
[31] H. Fu, G. Song, and Y. Wang, "Improved YOLOv4 marine target detection combined with CBAM," *Symmetry*, vol. 13, no. 4, 2021, Art. no. 623.
[32] W. Wang, X. Tan, P. Zhang, and X. Wang, "A CBAM based multiscale transformer fusion approach for remote sensing image change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 6817–6825, 2022.
[33] Y. Zhang et al., "A lightweight winter wheat planting area extraction model based on improved deeplabv3 and cbam," *Remote Sens.*, vol. 15, no. 17, 2023, Art. no. 4156.
[34] H. Liu, K. Simonyan, and Y. Yang, "Darts: Differentiable architecture search," 2018, *arXiv:1806.09055*.
[35] Y. Gao, H. Bai, Z. Jie, J. Ma, K. Jia, and W. Liu, "MTL-NAS: Task-agnostic neural architecture search towards general-purpose multi-task learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11543–11552.
[36] Z. Tu et al., "MaxViT: Multi-axis vision transformer," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2022, pp. 459–479.
[37] D. Zhang and F. Zhou, "Self-supervised image denoising for real-world images with context-aware transformer," *IEEE Access*, vol. 11, pp. 14340–14349, 2023.
[38] D. Hong et al., "Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks," *Remote Sens. Environ.*, vol. 299, 2023, Art. no. 113856.
[39] D. Hong, J. Yao, C. Li, D. Meng, N. Yokoya, and J. Chanussot, "Decoupled-and-coupled networks: Self-supervised hyperspectral image super-resolution with subpixel fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5527812.



Debin Wei was born in 1978. He received the Ph.D. degree in discipline of control science and engineering from the Nanjing University of Science and Technology (NUST), Nanjing, China, in 2023.

He is currently an Associate Professor with Dalian University, Dalian, China. His research interests include space-ground integrated network transmission technology, traffic engineering, and image processing.

Hongji Xie received the B.S. degree in food science and engineering from Shandong Agricultural University, Taian, China, in 2016. He is currently working toward the master's degree in computer science and technology with the Communication and Network Laboratory, Dalian University, Dalian, China.

His current research interests include computer vision and deep learning.

Pinru Li received the B.S. degree in digital media technology from Zhengzhou Normal University, Zhengzhou, China, in 2021. He is currently working toward the master's degree in computer science and technology with the Communication and Network Laboratory, Dalian University, Dalian, China.

His current research interests include computer vision and deep learning.

Yongqiang Xu received the B.S. degree in data science and big data technology from the Anhui University of Science and Technology, Huainan, China, in 2022. He is currently working toward the master's degree in computer science and technology with the Communication and Network Laboratory, Dalian University, Dalian, China.

His current research interests include computer vision and deep learning.