ISPRS Journal of Photogrammetry and Remote Sensing 206 (2023) 222–241

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

WaterFormer: A coupled transformer and CNN network for waterbody detection in optical remotely-sensed imagery

Jian Kang [a], Haiyan Guan [a,*], Lingfei Ma [b,*], Lanying Wang [c], Zhengsen Xu [a], Jonathan Li [c]

[a] School of Remote Sensing and Geomatics Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China
[b] School of Statistics and Mathematics, Central University of Finance and Economics, Beijing 102206, China
[c] Department of Geography and Environment Management, University of Waterloo, Waterloo, ON N2L 3G1, Canada

ARTICLE INFO

Keywords:
Optical remotely-sensed imagery
Convolutional neural networks (CNNs)
Visual Transformer
Waterbody detection (WD)
Multi-scale feature
Long-range dependency

ABSTRACT

As one of the most significant components of the ecosystem, waterbody needs to be highly monitored at different spatial and temporal scales. Nevertheless, waterbody variations in shape, size, and reflectivity, complicated and varied types of land covers, and environmental scene diversity, present colossal challenges in achieving accurate waterbody detection (WD). In this paper, we propose a novel network coupled with the Transformer and convolutional neural network (CNN), termed WaterFormer, to automatically, efficiently, and accurately delineate waterbodies from optical high-resolution remotely sensed (HR-RS) images. This network mainly includes a dual-stream CNN, a cross-level Vision Transformer, a light-weight attention module, and a sub-pixel up-sampling module. First, the dual-stream network abstracts waterbody features at multi-views and different levels. Then, to exploit the long-range dependencies between low-level spatial information and high-order semantic features, the cross-level Vision Transformer is embedded into the dual-stream, aiming at improving WD accuracy. Afterwards, the light-weight attention module is adopted to provide semantically strong feature abstractions by enhancing discrimination neurons, and the sub-pixel up-sampling module is employed to further generate high-resolution and high-quality class-specific representations. Quantitative and qualitative evaluations demonstrated that the WaterFormer provided a promising means for detecting waterbody areas in satellite images under complex scene conditions. Moreover, comparative analyses with the state-of-the-art (SOTA) alternatives, e.g., MSFENet, MSAFNet, and BiSeNet, also verified the generalization and superiority of the WaterFormer in WD tasks. The assessment results exhibited that the WaterFormer gained an average accuracy of 97.24%, average precision of 94.59%, average recall of 91.95%, average F_1-score of 93.24%, and average Kappa index of 0.9133, respectively. Additionally, we presented an open-access HR satellite imagery waterbody dataset, a mesoscale dataset with high-quality and high-precision waterbody annotation to facilitate future research in this field. The dataset has been released at https://github.com/NJdeuK/WD_Dataset.

1. Introduction

Water, as one of the most essential elements in the Earth's ecosystem, is critical for energy cycles, ecological coordination, and human society development (Vorosmarty et al., 2000; Yang et al., 2018; Liu et al., 2020a). In particular, current global climate changes have intensified the conflict between human society and the natural environment, which makes how to scientifically manage and efficiently utilize water resources for balancing environmental protection and socio-economic development more important. (Pekel et al., 2016; Chen et al., 2020). Consequently, it is of great implication to analyze waterbody characteristics and its temporal-spatial distribution patterns.

Currently, various Earth observation systems have been rapidly built to provide massive remote sensing data at different spatial–temporal scales (Li et al., 2022b). Due to the high-temporal and spatial resolution, large coverage area, rich spectrum information, clear geometric structures, and texture features, the optical high-resolution remotely sensed (HR-RS) images have been widely utilized in different tasks, such as object identification (Zhang et al., 2022) and change detection (Li et al., 2022a). However, waterbody detection or identification from HR-RS images still faces some challenges, as shown in Fig. 1. The difficulties are recapitulated as follows:

* Corresponding authors.
E-mail addresses: 202211610001@nuist.edu.cn (J. Kang), guanhy.nj@nuist.edu.cn (H. Guan), l53ma@cufe.edu.cn (L. Ma), lanying.wang@uwaterloo.ca (L. Wang), junli@uwaterloo.ca (J. Li).

https://doi.org/10.1016/j.isprsjprs.2023.11.006
Received 29 March 2023; Received in revised form 6 November 2023; Accepted 7 November 2023
Available online 17 November 2023

1) Waterbody variations in shape, size, and reflectivity. Natural environments (e.g., land cover, topography, and climate) and human economic activities carve different appearances of waterbody areas, see Fig. 1(a). Ditches and artificial rivers are normally narrow and relatively short, while lakes and natural rivers are wide and long. In some areas, paddy fields and ponds are densely distributed and spotted. Moreover, the reflectivity of waterbodies greatly depends on their clarities and different contents (e.g., algae, sand, and sediment).

2) Environmental scene diversity. Different natural environments shape waterbodies with different styles in RS images. As shown in Fig. 1 (b) low-resolution satellite images blurred the edges or boundaries of objects, i.e., waterbody. On the contrary, HR-RS images provide clear texture and obvious edge contours of the objects. As shown in Fig. 1(c) the appearance of waterbody areas varies with environmental scenes. This is a challenge for deep learning (DL) algorithms to equalize the proportion of negative and positive pixels in the same batch samples.

3) Complicated and varied types of land covers. The "same spectrum foreign matter" is always a huge challenge in optical HR-RS image interpretation tasks. As shown in Fig. 1(d) the spectral information of some waterbodies resembles with those of the shadows caused by high-rise objects, which could easily generate inaccurate alarms. In addition, some building roofs with high reflectivity, also affect the WD accuracy.

4) Deficiency of specialized datasets. Whether supervised or weakly supervised learning requires a certain number of waterbody samples for learning waterbody semantic features from the satellite images. So far, excluding the Gao-fen challenge dataset presented by (Sun et al., 2021), many WD studies assessed their performance on land cover datasets, such as DeepGlobe, GID, and LoveDA (Demir et al., 2018; Tong et al., 2020; Wang et al., 2021a).

To tackle these challenging issues, an ever-increasing number of WD methods have been presented. Most traditional methods such as Normalized Difference Water Index (NDWI) and clustering or classification methods (e.g. Support Vector Machine (SVM) and K-means clustering) detected the waterbody from satellite images that required to adapt the thresholds or manually designed lower-level features, e.g., textural, geometrical edge features, and spectral (Mcfeeters, 2007; Wang et al., 2019; Liu et al., 2020b). These methods have acquired improved accuracies and efficiency in certain circumstances, but robustly, automatically, and accurately detect waterbodies from HR-RS images in large-scale areas is still a big challenge.

Since the powerful abilities of semantic representation, high-level feature characterization, and robustness, the variety of deep learning networks (e.g., CNNs and the Vision Transformer (ViT) network (Dosovitskiy et al., 2021)) have drawn increasing attention to accurately and effectively detect, extract, and classify multi-scale waterbodies in complex scenes. It is noticed that most CNNs usually fail to establish long-range dependencies and represent heterogeneous object regions, resulting in the predicted results containing a large amount of "salt-and-pepper" noise. On the contrary, ViT can represent global relations, but it is poor in preserving local spatial detail information. Moreover, Transformer-based networks suffer from the issues of computational efficiency and input data size.

Therefore, we proposed a novel CNN and Transformer fusion network (WaterFormer) to robustly and accurately detect waterbodies from HR-RS images. By integrating local spatial features with global contextual features, the WaterFormer describes waterbodies using fine-grained semantic feature representation. The WaterFormer architecture includes: 1) a dual-stream CNN-based baseline with parallel-in-branches, which generates multi-scale information-rich feature maps at different levels, 2) a Cross-Level Vision Transformer (CL-ViT), which establishes long-range dependencies between local and global features, and 3) a Light-Weight Attention (LWA) module and a Sub-pixel Up-Sampling (SUS) module are embedded into the WaterFormer, which inhibits and discriminates the irrelevant features capability of the network. The WaterFormer provides a promising detection result of variable waterbodies in shape, size, and reflectivity, environmental scene diversity, and complicated types of land cover in HR-RS images. The contributions of this paper are threefold as follows.

1) We propose a coupled Transformer and CNN network for WD tasks, termed WaterFormer, where a dual-stream encoder-decoder CNN-based network is embedded with a CL-ViT module to integrate fine-grained spatial features and global contextual semantic information in a hierarchical and collaborative manner, which can accurately and completely detect waterbodies with varied shapes and different spatial

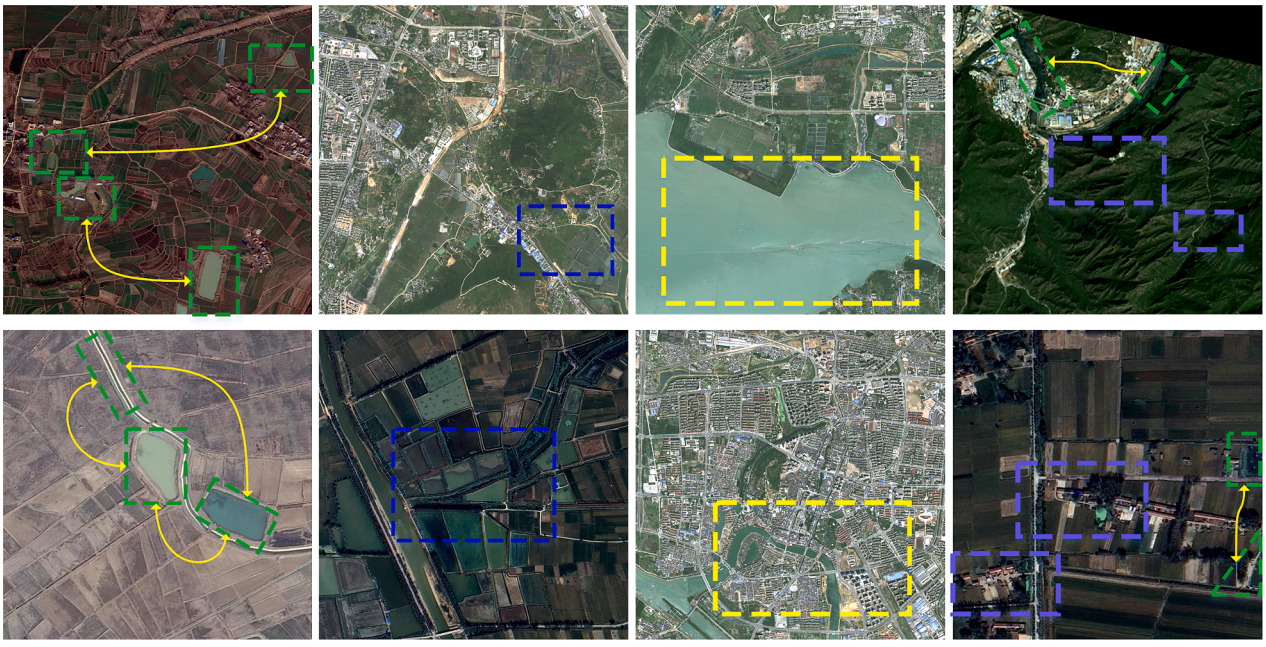

Fig. 1. Challenges of WD from HR-RS images: (a) variable waterbody in shape, size, and reflectivity; (b) different sensor resolutions; (c) diversified scenes; (d) the shadows caused by high-rise objects and high reflectivity buildings.

resolutions in diverse scenes.

2) We design a light-weight attention (LWA) module and a sub-pixel up-sampling (SUS) module to strengthen the abstraction and representation capability of the output features. Specifically, the LWA module aims to enhance the semantic feature representation of waterbodies, preserve feature spatial details, and improve the utilization ratio of model parameters. The SUS module generates high-resolution and high-quality discriminative waterbody feature maps in the decoding contextual branch.

3) We create a waterbody dataset from multi-source HR satellite images, named Multi-Sensor-Resolution Waterbody Dataset (MSRWD), which could be promoted to WD studies using DL techniques in the future. The dataset can be downloaded at https://github.com/NJdeuK/WD_Dataset.

The remaining parts of this paper is organized as follows. Section 2 comprehensively reviews the literatures about WD studies. Section 3 illustrates the architecture of the WaterFormer in detail. Section 4 presents the experimental datasets, implementation details, as well as experimental results and analyses. Finally, Section 5 gives the concluding remarks.

2. Related work

Recently, the detection/identification of waterbodies has drawn increasingly attention in RS image interpretation tasks. In this section, we briefly reviewed the existing image-based works, including tradition-based and CNN-based WD methods, followed by Transformer-CNN fusion methods.

2.1. Traditional waterbody detection

In the past decades, a large number of WD methods have been developed. In terms of fundamental principles, traditional WD methods can be broadly divided into three major categories: threshold-based, machine-learning methods, and hybrid-based methods.

Firstly, the threshold-based methods detected/extracted waterbodies from HR-RS images based on the spectral reflectance characteristics. Regarding the number of employed bands, the threshold-based methods are further grouped into single-band (Shih, 1985) and multi-band threshold methods(Koponena et al., 2002). The former detected waterbodies by using a single spectral difference value, while the latter mainly performed WD by a set of multi-band mathematical logic operations, e.g., NDWI (Mcfeeters, 2007), HR-WI (Yao et al., 2015), Background Difference WI (BDWI) (Li et al., 2021), and Multi-Band WI (MBWI) (Wang et al., 2018). Moreover, using Sentinel-2 data, a Triangle WI (TWI) was proposed to accurately delineate WD results in ice- and snow-covered areas, urban areas with cast shadows, and mountainous regions with highly rugged terrain (Niu et al., 2022). Despite achieving high accuracy in various waterbody types and environmental conditions, the TWI did not take into account the influence of varying solar angles, changes in the physical and chemical characteristics of waterbodies, bathymetry, and the daily or seasonal variations in solar angles. Aroma et al. (2023) designed a deep-blue-NDWI (DBNDWI) for Landsat-8 coastal/aerosol band and demonstrated that the DBNDWI achieved great accuracy improvement in medium spatial resolution images, compared to the Wavelet-based-NDWI (WAWI) and Weighted NDWI (WNDWI).

Secondly, the machine learning based methods for WD from HR-RS images mainly include the following steps: training sample generation, feature engineering, and classifier selection. In terms of training data, the machine-learning methods are also further categorized into two groups: supervised and unsupervised methods. The unsupervised algorithms, such as Fuzzy C-means clustering and K-means clustering (Wang et al., 2019; Zhang et al., 2019), performed the waterbody detection efficiently, but obtained low reliability and poor results. The supervised methods such as SVM and Markov Random Fields (Liu et al., 2020b; Elmi et al., 2016) detected waterbodies from HR-RS images by exploiting their low-order spatial and spectral features such as edges, textures, and shapes. Nevertheless, these methods relied on domain expert knowledge and had limited feature expression capabilities, difficult to capture the deep-level semantic information and the spatial relationship between pixels.

Thirdly, the hybrid-based methods precisely and accurately accomplish WD tasks by integrating waterbody features (e.g., generated by some threshold-based methods) and machine-learning classifiers. Concretely, Gašparović and Singh (2022) proposed an automatic algorithm for urban waterbody mapping (AUWM) using Sentinel-2 data by combining a modified-NDWI (MNDWI), a K-means clustering algorithm, and pan-sharpening techniques. While the algorithm achieved high-quality waterbody mapping with the accuracy of 99.7 %, it exhibited considerable uncertainty applied to very tiny urban waterbodies. Rajendiran and Kumar (2023) proposed a surface water body extraction (SWBE) method, which explored Gabor filters for generating pixel level features (PLF), a variety of spectral indices for describing water features, and an eXtreme Gradient Boosting (XGB) algorithm for classifying water and no-water pixels from Resoucesat-2 images. The image processing processes of the hybrid methods for the WD tasks were intricate, tedious, and high uncertainty with multiple influencing factors.

To summarize, the traditional methods heavily rely on expertise domain knowledge to conduct the feature engineering, suffer from the poor interpretation of spatial relationships, and limit the generality during the WD from the HR-RS imagery.

2.2. Cnn-based waterbody detection

With excellent feature abstracting capabilities and end-to-end automation features, CNNs have been widely used in semantic segmentation. On one hand, HR-RS images characterize rich spectral information, fine texture features, as well as clear pixel geometry and topology, contributing to the WD. On the other hand, such rich information also increases the difficulties of RS image interpretation. WD tasks can be simply regarded as a binary semantic segmentation. Comparatively, the CNN-based WD methods usually achieved better detection accuracies and correctness because these methods usually learned high-order features and semantic information rather than simple and low-order manually designed features, which had proven that CNNs have great potential in WD tasks. The fully convolutional network (FCN) and its variants, such as FCN8s or UNet, have achieved huge success in WD tasks (Li et al., 2019; Feng et al., 2019; Li et al., 2018). To facilitate deep training, dense blocks were further embedded to improve the WD performances (Wang et al., 2020a). Nevertheless, those methods poorly explored deep global contextual information, nonconductive to the expression of high-level semantic features of objects. To address these issues, a multi-scale lake waterbody extraction network was proposed by Wang et al., (2020b), where a multi-scale densely-connected module was embedded into the encoder-decoder structure. To effectively detect waterbodies from aerial and satellite RS dictates, Zhang et al. (2021b) proposed a Multi-feature Extraction and Combination Network (MECNet) to enhance semantic information and increase the diversity of waterbody features. To detect waterbody with size and shape variations, a Multi-Scale Context Extractor Network, named MSCENet, was devised by Kang et al. (2021), where the Res2Net and strip pooling were utilized to integrate high- and low-level feature maps.

Contrary to the traditional image-pixel annotations, a point annotation strategy was adopted in the Neighbor Feature Aggregation Network (NFANet) for WD tasks (Lu et al., 2022). NFANet could minimize its dependence on pixel-level labeled waterbody samples, but it will lose the context space information by utilizing neighboring features. Furthermore, the value of NFANet's parameter was relatively large, reaching 278.7 M. A Multi-Scale Features Extraction Network (MSFENet), proposed by Liu et al., (2023a), faced the challenges posed by large spectral-spatial variations of waterbodies. At the same time, contrastive

learning (CL) was employed to weaken the dependency of WD accuracies on the number of training samples. However, the CL strategy had obvious improvement effects only when the labeled waterbody samples were insufficient. The WD accuracies of weak supervision methods (e.g., NFANet, and MSFENet) depended on the suitability of learning strategies and the performances of segmentation models. In particular, a segmentation model negatively affects the features accurately learned from labeled samples and the full use of unlabeled samples. Additionally, model convergence requires a significant time for model training. To improve the efficiency of model training and further mine waterbody semantic information, a Multiscale Successive Attention Fusion Network (MSAFNet) was proposed to detect waterbodies in complicated situations (Lyu et al., 2023). To trade-off the efficiency and accuracy, Nie et al. (2023) presented a Squeeze-and-Excitation Bilateral Segmentation Network (SE-BiSeNet) for urban WD tasks. Compared with the other SOTA methods, the SE-BiSeNet demonstrated excellent performance in both accuracy and efficiency, with fewer parameters and computational requirements. Although these methods achieved a significant improvement on WD, they still were deficient in capturing fine-grained spatial features and local–global contextual semantic information modeling.

2.3. Transformer-CNN fusion methods in RS tasks

Transformer, a pure attention-based architecture, was first presented in natural language processing (NLP). ViT originally applied Transformers to visual tasks by using patch embedding. Specifically, feature maps or image inputs were divided into a set of patches, each of which was reshaped into a feature vector with positional information. Finally, the generated token embeddings were forwarded to the Transformers. Segmentation Transformer (SETR) achieved the SOTA performance by employing a CNN decoder into the ViT (Zheng et al., 2021). Compared to CNN-based networks, the ViT has the capability to capture long-range dependencies for powerfully representing global information. Thus, to obtain salient and high-order feature representation, many studies presented Transformer-CNN fusion methods. Subsequently, an UNet-like Transformer Network (UNetFormer) was designed for efficient segmenting urban scenes (Wang et al., 2022b), where the lightweight ResNet-18 and an efficient global–local attention mechanism were utilized as the encoder and decoder, respectively. Regarding the influence of image cropping processing on model contextual-aware, Ding et al. (2022) presented a Wide-Context Network (WiCoNet) to address the limitation of input image sizes. Inspired by the Transformer and UNet, Li et al., (2022a) proposed a hybrid Transformer and UNet Network (TransUNetCD) for change detection. Additionally, Transformers have been widely used for detecting other objects, such as buildings, roads, shadows, and waterbodies. For example, Wang et al. (2022) developed a BuildFormer to detect fine-grained buildings; Liu et al. (2023b) proposed a RoadFormer to accurately extract roads with highly structured and long-distance distributed, from RS images; Hu et al. (2023) designed a Multiscale Deformable Transformer Network (MDTNet) for alleviating residual errors in road extraction results; Transformer had been innovatively used in a dual-stream network (DTHNet) for discriminating shadows from dark waters, trees, roads, and other targets (Zhang et al., 2023). Furthermore, Zhong et al. (2022) handled the issues of inaccurate lake boundary detection and lake over-detection via a Noise-canceling Transformer Network (NTNet). Although the NTNet had achieved accurate WD, its parameter of model size was larger than other comparison methods, which was 127.06 M.

In summary, Transformer-CNN fusion methods have achieved significant results in RS tasks. Nonetheless, these methods struggle to achieve an optimal balance between detection efficiency and accuracy in the WD tasks.

3. The proposed method

3.1. WaterFormer architecture overview

The architecture of our WaterFormer, which is composed of a dual-stream network, a CL-ViT module, a LWA module, and a SUS module, is presented in Fig. 2. The dual-stream network consists of two asymmetrical branches, including a spatial branch and a contextual branch, each of which employs an encoder-decoder architecture. The CL-ViT module (see Section 3.2), embedding at the top of the encoder, is utilized to map the spatial scene information and aggregate the contextual semantic knowledge between the dual branches. The LWA module (see Section 3.3) and the SUS module (see Section 3.4) are adopted to enhance semantic features at different scales and levels and to hierarchically restore image resolutions and semantic information representation in the contextual branch, respectively.

It is challenging to trade-off spatial detail information and semantic features, which are critical for achieving high accuracy results. However, current mainstream networks obtained semantic features at a single path via a down-sampling, to enlarge the receptive fields. The down-sampling inevitably loses spatial detail information. Thus, we adopt a dual-stream architecture, which is equipped with a spatial branch to obtain low-order features and spatial details information. Specifically, in this spatial branch, the encoder contains four stages of convolution layers, as shown in Fig. 2. The first convolution stage is stacked by three convolution layers, whose filter kernel size and padding uniformly are set to 3 and 1, respectively. The first convolution stride is set to 2 and the rest to 1. The number of the output channels is successively set to [32, 32, 64]. The filter kernel size, the number of output channels, stride, and padding for the remaining three convolution stages are [3, 64, 2, 1], [3, 64, 2, 1], and [1, 512, 1, 0], respectively. Therefore, in the spatial branch, the original input images are down-sampled at the scales of [1/2, 1/4, 1/8, 1/8], the final output feature maps are 1/8 of the original input image size. The feature map size is straightway scaled to the original image size via a bilinear interpolation up-sampling strategy. Note that the pooling operation is not applied in this branch.

As shown in Fig. 2, the contextual branch is also designed with an encoder-decoder architecture to obtain global contextual information and represent high-level semantic feature. The ResNet-34 (He et al., 2016) is employed in the encoder. Concretely, the original input images are down-sampled at the scales of [1/2, 1/4, 1/8, 1/16, 1/32]. Correspondingly, the number of the feature maps is gradually increased to [128, 64, 128, 256, 512]. Similarly, the decoder consists of five SUS modules to gradually generate high-quality and high-resolution waterbody semantic feature maps. Especially, the LWA module is utilized to enhance the probability weight of waterbody pixels while restraining irrelevant background noise in feature maps at different level-scales. Skip connections are used to connect the waterbody semantic feature maps of the encoder with the corresponding up-sampled feature maps for feature detail recovery. Additionally, the CL-ViT module, which is designed at the top of each branch with 512 channels in the encoder, aims to construct long-range dependencies between the spatial and contextual branches.

3.2. Cross-Level vision Transformer module

We design a CL-ViT module to construct the long-range global feature interactions of the dual-branch features. The CL-ViT is stacked by multiple Transformer blocks, and it exports the overall feature representation of the spatial and contextual features. Note that, before patch embedding, the number of the feature channels is consistent, but the resolution-size is different.

Our CL-ViT contains two input branches and one output branch, as shown in Fig. 3. Assuming the spatial encoding features $\mathbf{S} = [\mathbf{S_1}, \mathbf{S_2}, \mathbf{S_3}.....\mathbf{S_M}] \in R^{M\times H\times W}$, where H, W, and M denote the height, width, and channel number of the feature maps, respectively. Let denote the

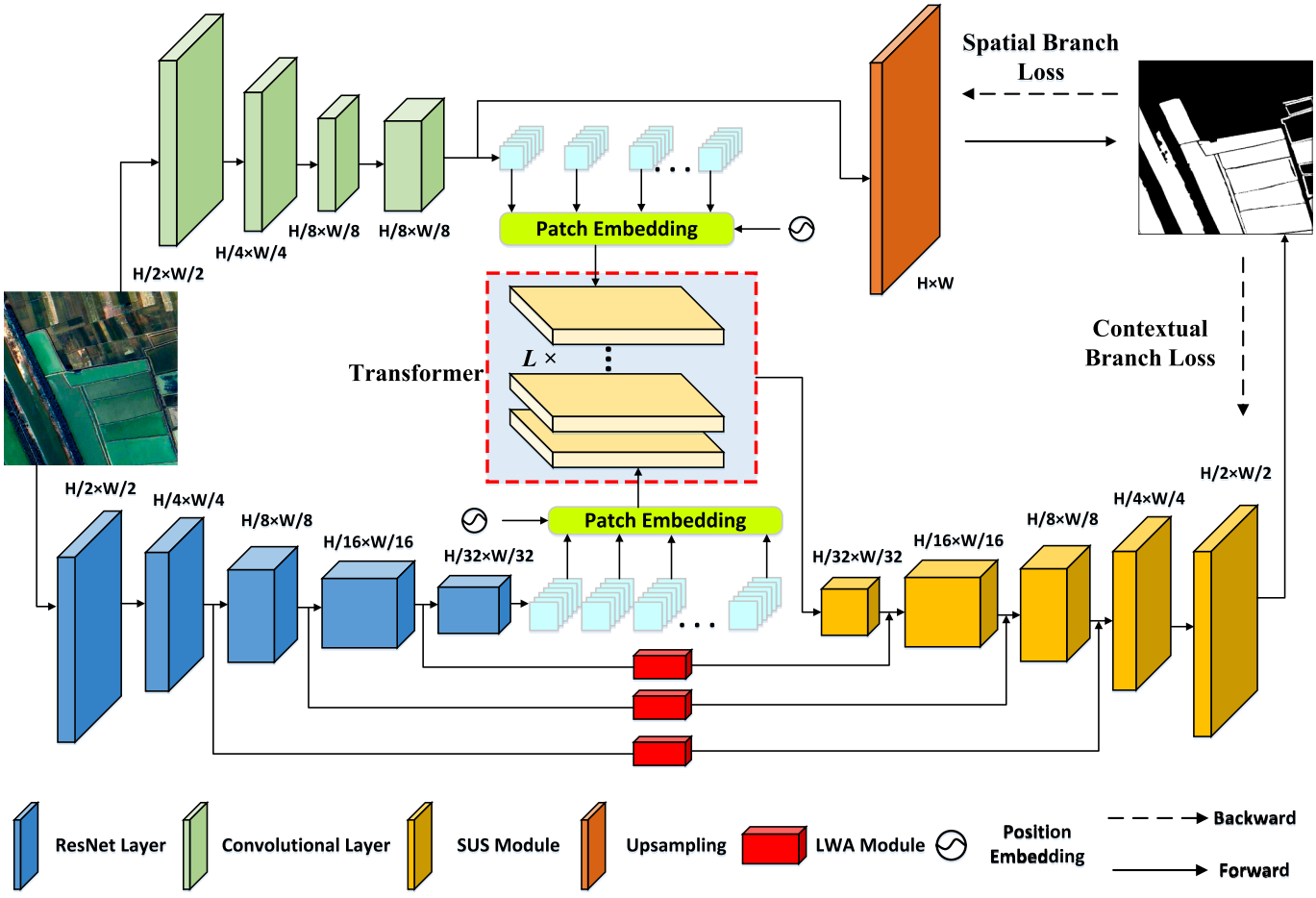

Fig. 2. Overview of the WaterFormer.

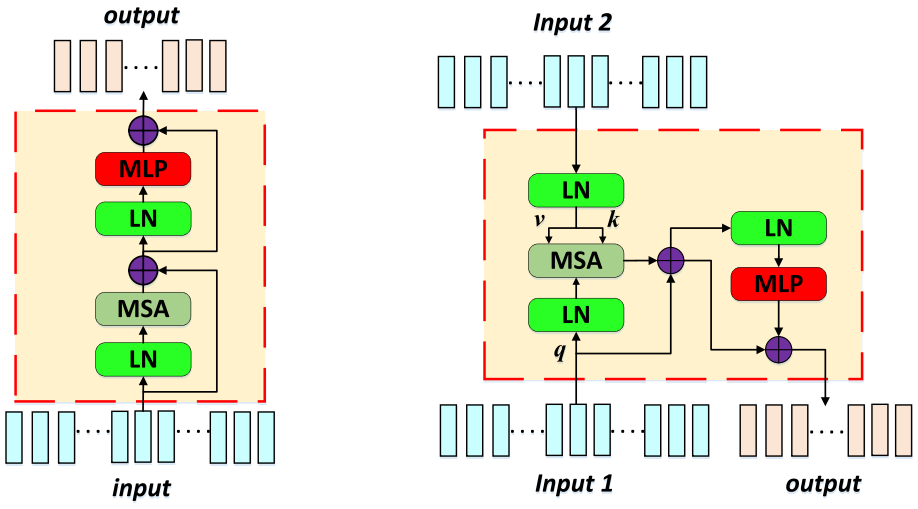

Fig. 3. Vision Transformer (Left) and Cross-Level Vision Transformer (Right).

contextual encoding features as $\mathbf{C} = [\mathbf{C_1}, \mathbf{C_2}, \mathbf{C_3}.....\mathbf{C_N}] \in R^{N\times X\times Y}$, where X, Y, and N denote the height, width, and channel number of the feature maps. Before inputting to the Transformer block, the Patch Embedding is applied to feature maps **S** and **C**. We take the feature map **S** as an example. **S** is first reshaped to a flattened 2D patch set, $\boldsymbol{P_S} = [\boldsymbol{P_1}, \boldsymbol{P_2}, \boldsymbol{P_3}.....\boldsymbol{P_M}] \in R^{G\times(M\times P\times P)}$, for performing tokenization, where the number of the patches is $G = HW/P^2$ and the patch size is $P \times P$. Afterward, the vectorized patch set $\boldsymbol{P_S}$ is mapped into the latent D-dimensional embedding space $\boldsymbol{T_S} \in R^{G\times D}$ via a trainable linear transformation, where D refers to the constant latent vector size of all the layers in the Transformer. Finally, the specific position embedding is learned and added to the patch embedding, to encode the patch spatial positional information. These steps can be represented as follows:

$$Z_S = [T_1E_1; T_2E_2; T_3E_3; .....; T_GE_G] + Epos \quad (1)$$

where $\mathbf{E} = [\mathbf{E}_1, \mathbf{E}_2, \mathbf{E}_3 \ldots\ldots \mathbf{E}_G] \in R^{(P\times P\times M)\times D}$ refers to the patch embedding projection, $\mathbf{Z_S} \in R^{G\times D}$ and $\mathbf{Epos} \in R^{G\times D}$ refer to the deep features and the position embedding, respectively. After obtaining the contextual and spatial embedded tokens, $\boldsymbol{T_C}$ and $\boldsymbol{T_S}$, the Transformer block is employed to conduct the feature projection relationship, which transfers the spatial detail information to the encoding contextual branch to generate context-rich feature representation. As shown in Fig. 3, each single Transformer total-block contains a Multi-Layer Perceptron (MLP) sub-block and a Multi-head Self-Attention (MSA) sub-block. The residual connection and normalization are employed in each inside Transformer block. The MSA aims to conduct the waterbody feature projection relationship between the two branches, obtaining high-order waterbody semantic with rich spatial-contextual information. The MLP effectively enhances the nonlinear transformation capability of the Transformer. In the inside of the Transformer block, the steps can be denoted as follows:

$$\overline{\mathbf{Z}} = \mathrm{MSA}(\mathrm{LN}(\mathbf{Z})) + \mathbf{Z} \tag{2}$$

$$\widetilde{\mathbf{Z}} = \mathrm{MLP}(\mathrm{LN}(\overline{\mathbf{Z}})) + \overline{\mathbf{Z}} \tag{3}$$

where LN denotes a Layer Norm function. $\widetilde{\mathbf{Z}}$ and $\mathbf{Z}$ represent the output and input patch embedding feature maps, respectively. The calculations detailed in the MSA sub-block are:

$$\overline{\mathbf{Z}} = \left[\mathrm{Softmax}\left(\frac{\mathbf{QK}^{\mathrm{T}}}{\sqrt{\mathrm{D/n}}}\right)\right]\mathbf{V} \tag{4}$$

where $\mathbf{V}, \mathbf{Q}, \mathbf{K} \in R^{G\times D/n}$ are the value, query, and key matrices of LN($\mathbf{Z}$). n is the number of the MSA's heads. Assuming $\mathbf{Z_S} \in R^{G\times D}$ and $\mathbf{Z_C} \in R^{O\times D}$ (G and O are the number of the flattened features) represent the spatial and contextual embedded tokens of the spatial branch and contextual branch, respectively. In the Transformer block, spatial query $\mathbf{Q_C}$ is mapped with $\mathbf{Z_C}$, while the contextual key $\mathbf{K_S}$ and value $\mathbf{V_S}$ are mapped with $\mathbf{Z_S}$:

$$\begin{aligned} \boldsymbol{Q}_C &= \mathbf{Z}_C\boldsymbol{W}_Q \in R^{O\times D/n}, \\ \boldsymbol{K}_S &= \mathbf{Z}_S\boldsymbol{W}_K \in R^{G\times D/n}, \\ \boldsymbol{V}_S &= \mathbf{Z}_S\boldsymbol{W}_V \in R^{G\times D/n}, \end{aligned} \tag{5}$$

where $\mathbf{W_V}, \mathbf{W_K}, \mathbf{W_Q} \in R^{D\times D/n}$ are the matching weights of the projection function. The self-attention mechanism is utilized to update $\overline{Z}$, which is calculated as follows:

$$\overline{\mathbf{Z}} = \left[\mathrm{Softmax}\left(\frac{\mathbf{Q}_C\mathbf{K}_S^{\mathrm{T}}}{\sqrt{\mathrm{D/n}}}\right)\right]\mathbf{V}_S \tag{6}$$

These steps, together with the MLP calculations, are reduplicated L times, where the contextual dependencies between $\mathbf{Z}_S$ and $\mathbf{Z}_C$ are modeled and imposed. In consequence, the contextually embedded tokens are mapped with the long-range dependencies from the spatially embedded tokens. In the end, the output embedded tokens are reshaped into the 2D feature maps, with waterbody semantic and fine-grained spatial information. We will analyze the different combinations of n and L on the WD performance in Section 4.4.

3.3. Light-weight attention module

Inspired by Yang et al. (2021), we adopted a LWA module to learn more discriminative neurons, enhancing waterbody semantic feature representation while preserving feature map spatial details, and also increasing the utilization ratio of the model parameters. As shown in Fig. 4, the LWA module efficiently highlights the neurons with a high degree of importance while suppressing the surrounding neurons via the following energy function:

$$e_t(w_t, b_t, y, x_i) = (y_t - \widehat{t})^2 + \frac{1}{M-1}\sum_{i=1}^{M-1}(y_0 - \widehat{x}_i)^2 \tag{7}$$

$$\widehat{t} = w_t t + b_t \tag{8}$$

$$\widehat{x}_i = w_t x_i + b_t \tag{9}$$

where $\widehat{x}_i$ and $\widehat{t}$ are the linear transformations of the other neurons feature map x_i, which is the single channel of the input feature $\mathbf{X} \in R^{C\times H\times W}$, and the target neuron t. i denotes the index over spatial dimension. M (M = H × W) denotes the neurons number of in the corresponding channel. $\boldsymbol{b_t}$ and $\boldsymbol{w_t}$ denote the bias and weight of the linear transformations, respectively. If $\boldsymbol{y_t}$ equals to $\widehat{t}$ and $\widehat{x}_i$ equals to y_0, e_t would be minimized.

Note that all variables are scalar in Eq. (7), where $\widehat{x}_i$ and y_0 are given to different values. Following this, the linear separability, between the other neurons and the target neuron t, is found via minimizing the energy function. To understand this, the bilevel labels (i.e., −1 and 1) are assigned to y_0 and y_t. In the final energy function, the regularization is added as follows:

$$w_t = -\frac{2(t-u_t)}{(t-u_t)^2 + 2\sigma_t^2 + 2\lambda} \tag{10}$$

$$b_t = -\frac{1}{2}(t+u_t)w_t \tag{11}$$

where $\sigma_t^2 = \sum_i^{M-1}(x_i - u_t)^2/(M-1)$ and $u_t = \sum_{i=1}^{M-1}x_i/(M-1)$ represent the variance and mean, which calculates over all neurons except t in that channel.

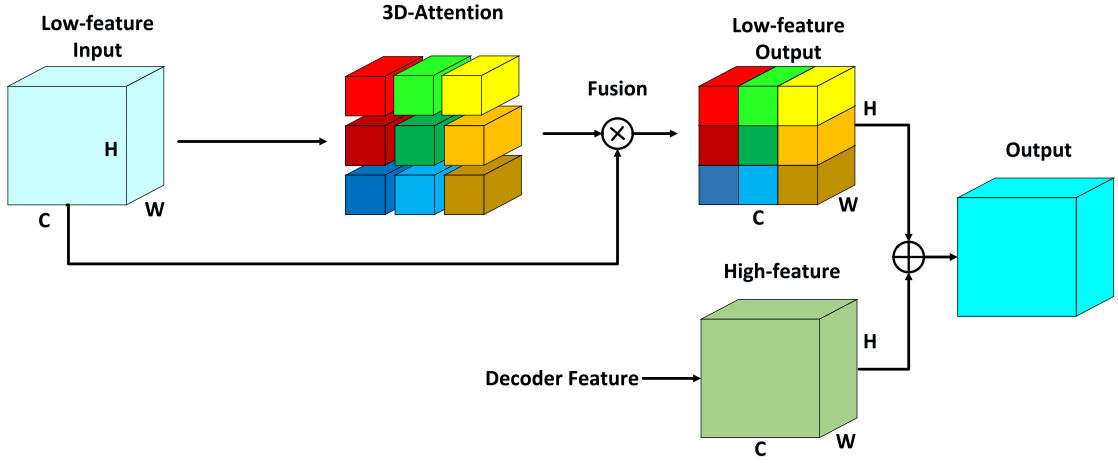

Fig. 4. Illustration of the LWA module.

Following the assumption that the latent representation in a single channel follows the same distribution (Hariharan et al., 2012). The minimal energy can be written as follows:

$$e_t^* = \frac{4(\widehat{\sigma}^2 + \lambda)}{(t - \widehat{u})^2 + 2\widehat{\sigma}^2 + 2\lambda} \tag{12}$$

$$\widehat{u} = \frac{1}{M-1}\sum_{i=1}^{M} x_i \tag{13}$$

$$\widehat{\sigma}^2 = \frac{1}{M-1}\sum_{i=1}^{M-1} (x_i - \widehat{u})^2 \tag{14}$$

It is obvious that the more distinctive the feature of a neuron coming from the others, the higher the weight, the lower the value of energy e_t^*. Thus, the value of $1/e_t^*$ represents the importance of each neuron. Note λ is a coefficient, as a hyper-parameter during model training. To simplify the procedure, the whole refinement step of the LWA module is defined as:

$$\widetilde{X} = \left[\sigma\left(\frac{1}{E}\right)\right] \otimes X \tag{15}$$

where **E** denotes the energy of the corresponding neuron. $\otimes$ and σ refer to the element-wise multiplication and the sigmoid function. $\widetilde{X}$ and X are the output and input feature maps, respectively.

The LWA module is embedded in the contextual branch at the scales of [1/2, 1/4, 1/8]. This module directly generates the full 3-dimensional (3D) weights of the feature map, enhancing the pixel weights of the regions of interest in each feature map while suppressing irrelevant background regions and aggregating the high-level decoder features. The waterbody semantic 3D weights refine the low-level encoding features, being indirectly conducive to the expression of waterbody semantic information in the decoding features. The specific phases can be depicted as follows:

$$F_{opt} = F_{hpt} + F_{lpt} = F_{hpt} + Atten(F_{ipt}) \otimes F_{ipt} \tag{16}$$

where $\mathbf{F_{opt}}$, $\mathbf{F_{hpt}}$, $\mathbf{F_{lpt}}$, and $\mathbf{F_{ipt}}$ denote the output, high-level ouput, low-level output, and input feature maps, respectively. ***Atten*** (·) represents the process of generating 3D-waterbodies feature weights. $\otimes$ is the element-wise multiplication operational procedure.

3.4. Sub-pixel up-sampling module

Inspired by the pixel-shuffle up-sampling algorithms (Shi et al., 2016), a SUS module is designed to generate more discriminative representations in the contextual branch, as illustrated in Fig. 5. The SUS module consists of one pixel-shuffle unit and two convolution layers with a kernel size of 1 × 1. The pixel-shuffle unit scales up the feature map size with a scaling step of two and reduces the channel dimension of the overall feature maps. The two convolution layers are employed to adjust and aggregate feature dimensions. The specific steps can be represented as follows:

$$\boldsymbol{F_{out}} = \boldsymbol{F_d} \oplus \boldsymbol{F_e} = f[SU(f(\boldsymbol{F_{in}}))] \oplus \boldsymbol{F_e} \tag{17}$$

where $\mathbf{F_{out}}$, $\mathbf{F_d}$, and $\mathbf{F_e}$ denote the output, input encoder, and input decoder feature maps, respectively. f (·) and SU (·) represent the 1 × 1 convolution and pixel-shuffle operations. Additionally, assuming the low-resolution (LR) feature map is denoted as $\mathbf{I_{LR}} \in R^{r2C\times W\times H}$ (r is the upscaling ratio), the high-resolution (HR) feature map as $\mathbf{I_{HR}} \in R^{C\times rW\times rH}$, and l as the number of the layers. The projection function can be described as:

$$\boldsymbol{I_{HR}} = SU^l(\boldsymbol{I_{LR}}) = \boldsymbol{W_l} \otimes SU^{l-1}(\boldsymbol{I_{LR}}) + \boldsymbol{b_l} \tag{18}$$

where $\mathbf{b_l}$ and $\mathbf{W_l}$ are the biases and weights of the l^{th} layer. As shown in Fig. 5(b), the upscaling ratio is set to 2, and the feature map size changes from (C × r^2, H, W) to (C, H × r, W × r).

The SUS module is prone to damage the correlation among pixels in the process of periodic arrangement. However, for a LR image feature and its corresponding HR image feature, the SUS module can enhance their correlations due to their structural similarity. Therefore, it can gradually recover a high-quality and high-resolution waterbody

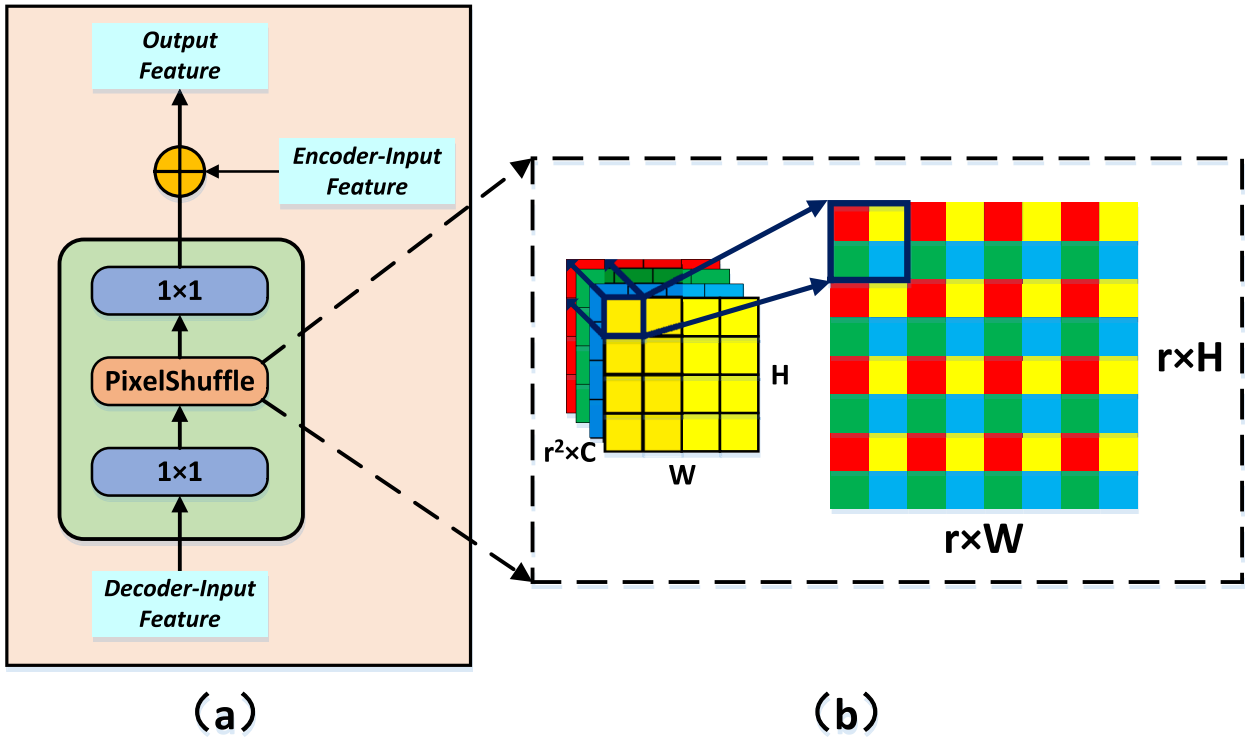

Fig. 5. Illustration of (a) the SUS module and (b) the upscale process of SUS (r = 2, C = 1).

semantic expression, improving the WD accuracy.

3.5. The total loss function

In pixel-level classification tasks, loss functions are employed to estimate the discrepancy between the predicted map and the ground truth (GT) during sample training, and further optimize and adjust network parameters. As shown in Fig. 2, the two predicted maps are generated from the spatial and the contextual branches, respectively, in the proposed WaterFormer. Therefore, a dynamic weighted binary-cross-entropy (BCE) loss, named DCE loss, which consists of the spatial loss and the contextual branch loss, was performed to drive the training process, which is described as follows:

$$L_{DCE} = L_{BCE}(P_C, T) + \alpha L_{BCE}(P_S, T) \tag{19}$$

$$\alpha = \left(1 - \frac{iter}{all\_iter}\right)^2 \tag{20}$$

where P_C and P_S, respectively, represent the segmentation maps generated by the contextual and spatial branches. T represents the GT. α is a weighting parameter, whose value is decreasing with an increase of the iterations.

4. Results and discussion

In this section, the experimental datasets and implementation details are first reported. Next, comparative methods and evaluation indexes are introduced. Finally, the experimental results are discussed and analyzed in detail.

4.1. Descriptions of datasets

To verify the generalization and robustness of the WaterFormer, we tested it on the three HR-RS datasets, i.e., GID (Tong et al., 2020), LoveDA (Wang et al., 2021a), and MSRWD. Table 1 shows the fundamental information of the experimental datasets. The details of these three datasets were explained in the following.

1) **GID**: The dataset consists of GF-2 images, originally annotated with six land covers, i.e., building, farmland, forest, meadow, waterbody, and background. To better facilitate WD tasks, an annotation mask was used to adjust these six land covers to two categories: waterbody and non-waterbody. We cropped the images into a set of non-overlapping image patches with the size of 512 × 512 pixels and removed both the mislabeled images and the images containing no waterbodies. Finally, the patches with their corresponding labels were randomly divided into three parts, i.e., model training set (containing 2291 image pairs, accounting for 60 % of the dataset), model validation set (containing 764 image pairs, 20 %), and model testing set (containing 764 image pairs, 20 %). To reduce over-fitting, we augmented the training subsets via an offline data augmentation, including color dithering, salt-pepper noise interference, and image rotation at three directions of 90°, 180°, and 270°.

2) **LoveDA**: The LoveDA dataset has 4191 annotated images, with seven types of land covers, i.e., road, barren, agriculture, waterbody, forest, and background. Similarly, these seven land covers were also adjusted to two categories, i.e., waterbody and non-waterbody. Each image was also cropped into a set of non-overlapping image patches with a size of 512 × 512 pixels. The training, validation, and testing subsets included 4061, 2436, and 1624 images, respectively, and the corresponding proportions of 50 %, 30 %, and 20 % of the whole dataset. We enlarged the training data via an online data augmentation, i.e., randomly image rotating, color dithering, and image flipping in horizontal and vertical directions.

3) **MSRWD**: As there have been few publicly available datasets for WD, we constructed a fine-grained HR satellite image dataset, especially for WD tasks in this paper. This dataset consists of 660 images, including 73 GF-2 images, 74 ZY-3 images, and 513 images acquired from the Google Earth services, thereby the multi-sensor-resolution waterbody dataset was termed as MSRWD. All images were annotated as waterbody and non-waterbody pixel-by-pixel. All images in the MSRWD dataset were uniformly clipped to a set of non-overlapping image patches with the size of 512 × 512 pixels. After the removal of the mislabeling images and the images with no waterbodies, 2419 image patches were remained. Finally, according to the ratios of 5:2:3, 1209, 485, and 725 image patches were used as training, validation, and testing subsets, respectively. We also employed an offline data augmentation, including random rotation, color dithering, salt-and-pepper noise addition, translation transformation, and scale transformation. Note that, compared with the GID and LoveDA datasets, the MSRWD dataset is more challenging due to: (1) varied scales, reflectivity, sizes, and shapes of waterbodies, e.g., artificial, or natural rivers, ponds, ditches, lakes, and paddy-fields, and (2) complex scenes, e.g., urban–rural scene, hills landform, loess plateau, and variability in image spatial resolution.

Table 1
Experimental datasets.

Dataset	GID	LoveDA	MSRWD
Data sources	GF-2	Google Earth	Google Earth, GF-2, ZY-3
Spatial resolution (m)	4.0	0.3	1.0, 4.0, 5.8
Image size (pixels)	7200 × 6800	1024 × 1024	1024 × 1024
Annotation category	6	7	2
Total image number	150	4191	660
Tasks	Object classification	Object classification	Waterbody detection

4.2. Network implementation details

The WaterFormer and the comparison methods were implemented on Pytorch 1.10.0 and Python 3.8.12. The hardware environment of all experiments is a workstation with two 24 GB NVIDIA GeForce RTX 3090, and an Intel(R) Xeon(R) Silver 4210R CPU (2.40 GHz, 10 cores, and 128 GB RAM). During the training stage, the Adam optimization algorithm was employed. The hyper-parameters of the GID, LoveDA, and MSRWD datasets were presented in Table 2. To better train the networks, an "ExponentialLR" scheme was employed for updating the learning rate, where the base of exponential, γ, was set to 0.98.

4.3. Comparative methods and evaluation indexes

To confirm the effectiveness of the WaterFormer on WD task, we compared it with four groups of the eight SOTA DL segmentation methods including CNN-based networks (e.g., LinkNet (Chaurasia and Culurciello, 2017), DeepLabV3+ (Chen et al., 2018)), attention-based networks (e.g., DANet (Fu et al., 2019), CCNet (Huang et al., 2020)), Prue Transformer-based networks (e.g., SwinUNet (Cao et al., 2022), CSwin (Dong et al., 2022)), and Transformer-CNN fusion networks (e.g., BANet (Wang et al., 2021b), TransFuse (Zhang et al., 2021a)), and six SOTA WD methods including MECNet (Zhang et al., 2021b), MSCENet (Kang et al., 2021), MSNANet (Lyu et al., 2022), BiSeNet (Nie et al.,

Table 2
Experimental hyper-parameters.

Hyper-parameter	GID	LoveDA	MSRWD
Epoch	50	120	60
Batch size	2	4	4
Initial learning rate	2.5e-5	1e-4	1e-4

2023), MSAFNet (Lyu et al., 2023), MSFENet (Liu et al., 2023a).

Furthermore, to quantitatively analyze the WD performance of different methods, five commonly used metrics, i.e., F_1-score (F_1), recall (R), precision (P), Kappa coefficient (KC), and overall accuracy (OA), were adopted by comparing the GT with the prediction pixel maps.

$$OA = \frac{T_P + T_N}{T_P + F_P + T_N + F_N} \tag{21}$$

$$P = \frac{T_P}{T_P + F_P} \tag{22}$$

$$R = \frac{T_P}{T_P + F_N} \tag{23}$$

$$F_1 = \frac{2 \times P \times R}{P + R} \tag{24}$$

$$P_e = \frac{(T_P + F_N) \times (T_P + F_P)}{(T_P + T_N + F_P + F_N)^2} + \frac{(T_P + F_P) \times (T_N + F_N)}{(T_P + T_N + F_P + F_N)^2} \tag{25}$$

$$KC = \frac{OA - P_e}{1 - P_e} \tag{26}$$

where T_P, T_N, F_P, and F_N are the number of correct waterbody detection pixels, correct no-waterbody detection pixels, incorrect waterbody detection pixels, and undetected waterbody pixels, respectively.

4.4. Parameters setting analysis

In the WaterFormer, there are three adjustable parameters, i.e., L and n in the CL-ViT module and λ in the LWA module, which significantly affect the WD performance. In this section, we conceived and designed two groups of experiments on the GID, LoveDA, and MSRWD datasets, to investigate the sensitivity of the WaterFormer to the selection of the aforementioned three parameters.

The n and L were the number of the MSA heads and the number of Transformer blocks, respectively. In this group of experiments, we assigned the original values of n and L to 2 and 4, respectively. We varied n from 2 to 8, and L from 4 to 8. Fig. 6 shows the experimental results of the three datasets on OA and F_1 when using different combinations of L and n. Obviously, the best OA values on the MSRWD, GID, and LoveDA datasets were obtained at n = 8 and L = 4, respectively. Also, the highest F_1 values were acquired on the MSRWD, GID and LoveDA datasets at n = 8 and L = 4. As shown in Fig. 6, as the number of heads in the MSA unit increases, the OA and F_1 values gradually increase with the same Transformer blocks on the LoveDA. There was no observed pattern of variation for the MSRWD and GID. The reason might be that the multi-head attention mechanism facilitates modeling the spatial information between the two CNN branches for the dataset with higher spatial resolution. However, the WaterFormer achieved a relatively stable WD performance, especially, on the GID and MSRWD datasets when n invariant and L increased. The reason might be that extensive Transformer blocks cause overfitting, and thus degrading the overall WD performance.

The λ is the parameter in the LWA module, which affects the 3D weights by determining the value of the energy function. We used n = 8 and L = 4, and varied λ from 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3} to 10^{-2} on the GID, LoveDA, and MSRWD datasets. The OA and F_1 values obtained by the WaterFormer, were presented in Fig. 7. It can be observed that the greatest OA and F_1 on the LoveDA and MSRWD were acquired when λ = 10^{-4}. The best values of OA and F_1 were obtained when λ was 10^{-4} or 10^{-3}. The reason behind this discrepancy may be that the spatial resolution of the GID images is lower than those of the LoveDA and MSRWD images. When processing low-resolution images, it is hard to distinguish the targets or peripheral neurons for the corresponding energy generated by λ = 10^{-4} and λ = 10^{-3}.

4.5. Comparison with SOTA DL methods

Figs. 8-10 present a visualization comparison of the WD results on the GID, LoveDA, and MSRWD datasets. The comparative methods and the GT were marked by red and blue boxes, respectively. The yellow boxes represent the WaterFormer. Additionally, Table 3 reports the quantitative verification results of the WD results acquired by those comparative methods.

*a) **GID:*** Fig. 8 assumes a subset of some typical WD results acquired by the comparative methods on the GID dataset. On the whole, visual inspection showed that the WaterFormer was superior to other methods under complicated scenes. To be more specific, as shown in the first rows of Fig. 8, the images contained many bright artificial buildings and shadows spreading around the tiny straight river. The WaterFormer achieved the best visual performance, and the WD map was greatly consistent with the GT. In the second rows of Fig. 8, most of the methods achieved better WD results. In particular, the WaterFormer and Transfuse clearly and accurately delineated the fine inlets, benchlands, and river islands. As shown in the third rows of Fig. 8, there was a large lake with dark spectrum brightness. Note that the WaterFormer and BANet accurately and completely delineated the lake, whereas the rest of the

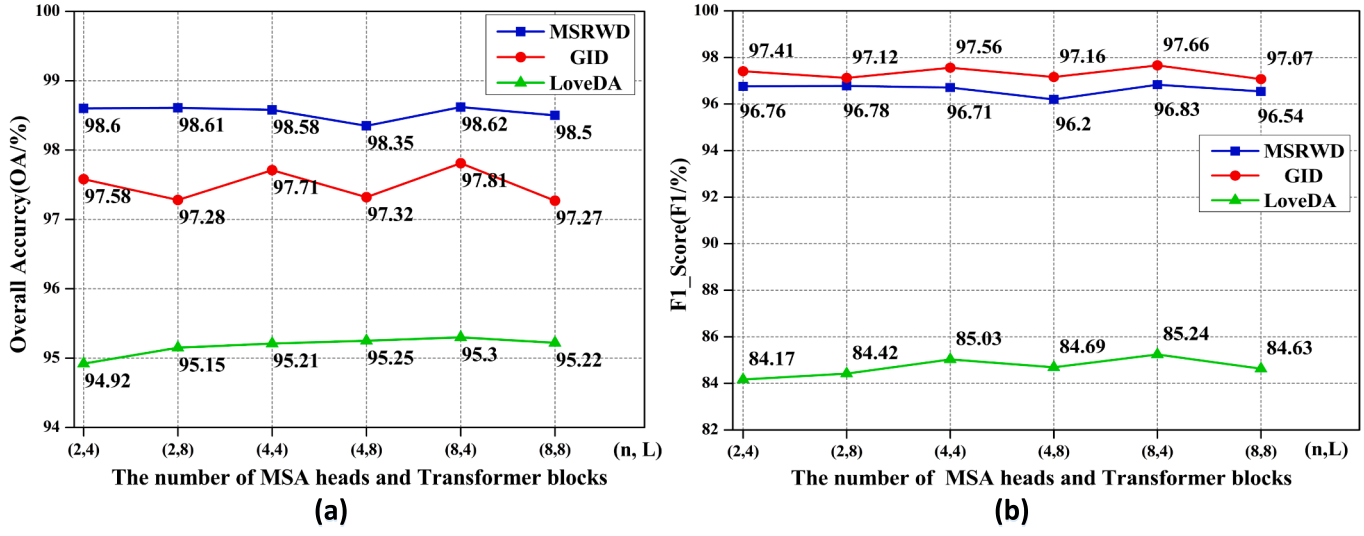

Fig. 6. Effects of the hyper-parameters in Transformer on the accuracies of the WaterFormer. (a) OA and (b) F_1.

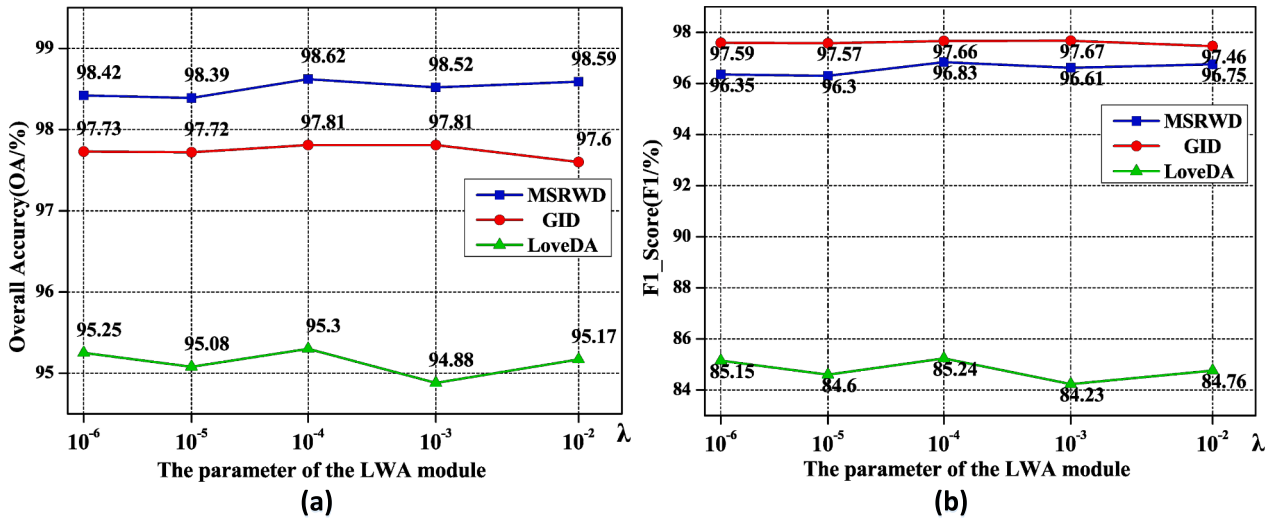

Fig. 7. Effects of the hyper-parameters in LWA module on the accuracies of the WaterFormer. (a) OA and (b) F_1.

methods obtained poor WD results with false alarms and uncertain areas. In addition, the WaterFormer generated WD maps with more accurate and smoother boundaries than others. In the bottom row of Fig. 8, there exists a remarkable "synonyms spectrum" phenomenon in the image with densely distributed buildings and a tiny cross-river bridge. The WD result obtained by the WaterFormer was also consistent with the GT. Thus, benefiting from the advantages of the CL-ViT module, the WaterFormer accomplished the promising WD performance when handling the waterbodies with complicated scenarios and multi-scale characteristics (e.g., geometric and spectral variations).

As shown in TABLE 3, our WaterFormer achieved the best WD accuracies against the reference methods. More specially, compared with the LinkNet, the WaterFormer achieved a better WD performance, with an improvement of 2.01 %, 1.46 %, 2.86 %, 2.18 %, and 4.04 % of OA, P, R, F_1, and KC, respectively. These values are 4.08 %, 3.91 %, 4.85 %, 4.39 %, and 8.19 %, respectively, comparable to DeepLabV3 +. The reason is that the dual-stream CNN encoder has a stronger feature abstraction ability than the light-weight Xception and the single-branch residual architecture. In particular, compared with the CNN-based attentional networks, our WaterFormer obtained a mean increase of 1.33 %, 1.75 %, 1.06 %, 1.41 %, and 2.67 % of OA, P, R, F_1, and KC, respectively. This is because the CL-ViT module mapped the spatial detail information into high-level semantic feature maps, improving WD integrality and accuracy. Additionally, our method was superior to the pure Transformer-based networks by a gain of 0.34 %, 0.33 %, 0.39 %, 0.36 %, and 0.66 % of OA, P, R, F_1, and KC, respectively. The WaterFormer, integrating the CNN with the Transformer, had a stronger ability to capture local features of waterbodies than the pure Transformer-based methods. Unlike the BANet, the better WD performance of the WaterFormer benefited from the LWA module to enhance waterbody semantic features at multiple levels while preserving the raw feature map spatial details. Compared with TrusFuse, the WaterFormer obtained a slight improvement of 0.03 %, 1.33 %, and 0.04 % of OA, P, and KC, respectively. This is because the SUS module gradually recovers a high-quality and high-resolution waterbody semantic expression.

*b) **LoveDA:*** Fig. 9 shows the WD results obtained by the comparative approaches. It can be seen that our WaterFormer could refine the waterbody boundaries and reduce the false alarms, thus generating finer detection maps, as presented in the first row of Fig. 9. In the second and third row of Fig. 9, the waterbody surrounded by artificial buildings and bare soil appear strong spectral reflectivity. The WaterFormer generated the WD result maps, extremely consistent with the GT. When dealing with the waterbody isolated, staggered with buildings and vegetation, or covered with shadows, as observed in the bottom row of Fig. 9, the WaterFormer completely and correctly delineated the waterbody, while the other methods failed because little spatial information was fully mapped into the high-level semantic feature maps during forward propagation. Thus, this can be drawn a conclusion that the WaterFormer has a superior robustness and noise suppression capability.

As reported in Table 3, the WaterFormer presented a considerable performance boost over other methods with OA, F_1, R, and KC values reaching 95.30 %, 85.24 %, 82.91 %, and 0.8245, respectively. Compared with LinkNet and DeepLabV3+, our WaterFormer acquired a mean increase of 0.74 %, 0.83 %, 4.21 %, 2.66 %, and 3.08 % for OA, P, R, F_1, and KC, respectively. However, these improvements are 0.55 %, 0.59 %, 3.18 %, 1.98 %, and 2.29 %, respectively, comparable to the DANet and CCNet. This is because the attention mechanism is more beneficial for the constructed long-distance correlation of each pixel than other optimization units in the clear-texture and fine-detail images. The WaterFormer outperformed the BANet, with a great increase of 1.01 %, 4.03 %, 7.47 %, and 4.58 % of OA, F_1, R, and KC, respectively. Furthermore, these values are 0.36 %, 2.06 %, 6.53 %, and 2.21 %, respectively, comparable to the TransFuse.

*c) **MSRWD:*** As shown in Fig. 10, the ZY-3 and GF-2 satellite images were presented in the first and second rows, and the Google images were in the last three rows. Accordingly, we found that the WaterFormer achieved superior results to the other methods. Take the example of the ZY-3 image, the river presents thin, long, and meander. Furthermore, it can be carefully observed that the LR-RS images presented typical surface objects (e.g., buildings, roads, and farmlands) with rough texture, blurred edges, and monotonic spectrums. In fact, detecting waterbodies accurately and completely from such images is still an extraordinarily challenging task. However, the WaterFormer still achieved the SOTA WD results with a clear and complete delineation of waterbodies. In contrast, the GF-2 images contain rich spectrum information. The WaterFormer missed detections, that is, waterbody with bright reflectivity or containing eutrophic matters were largely missed. The last three rows of Fig. 10 present the Google images, containing regular farmland water nets, densely distributed large waterbodies, and thin-long artificial irrigation canals. The narrow waterbody areas were detected completely and precisely by the WaterFormer, comparable to the other methods. As shown in the fourth row of Fig. 10, the WaterFormer delineated accurate waterbody boundaries, whereas the DeepLabV3+, SwinUNet, BANet, and TransFuse achieved poor WD results without fine and smooth boundaries. The reason might be that the SUS module is capable of enhancing the correlation between pixels in similar topological structure images. Therefore, it improved the detection results of waterbody boundaries. In addition, for the thin-long artificial irrigation

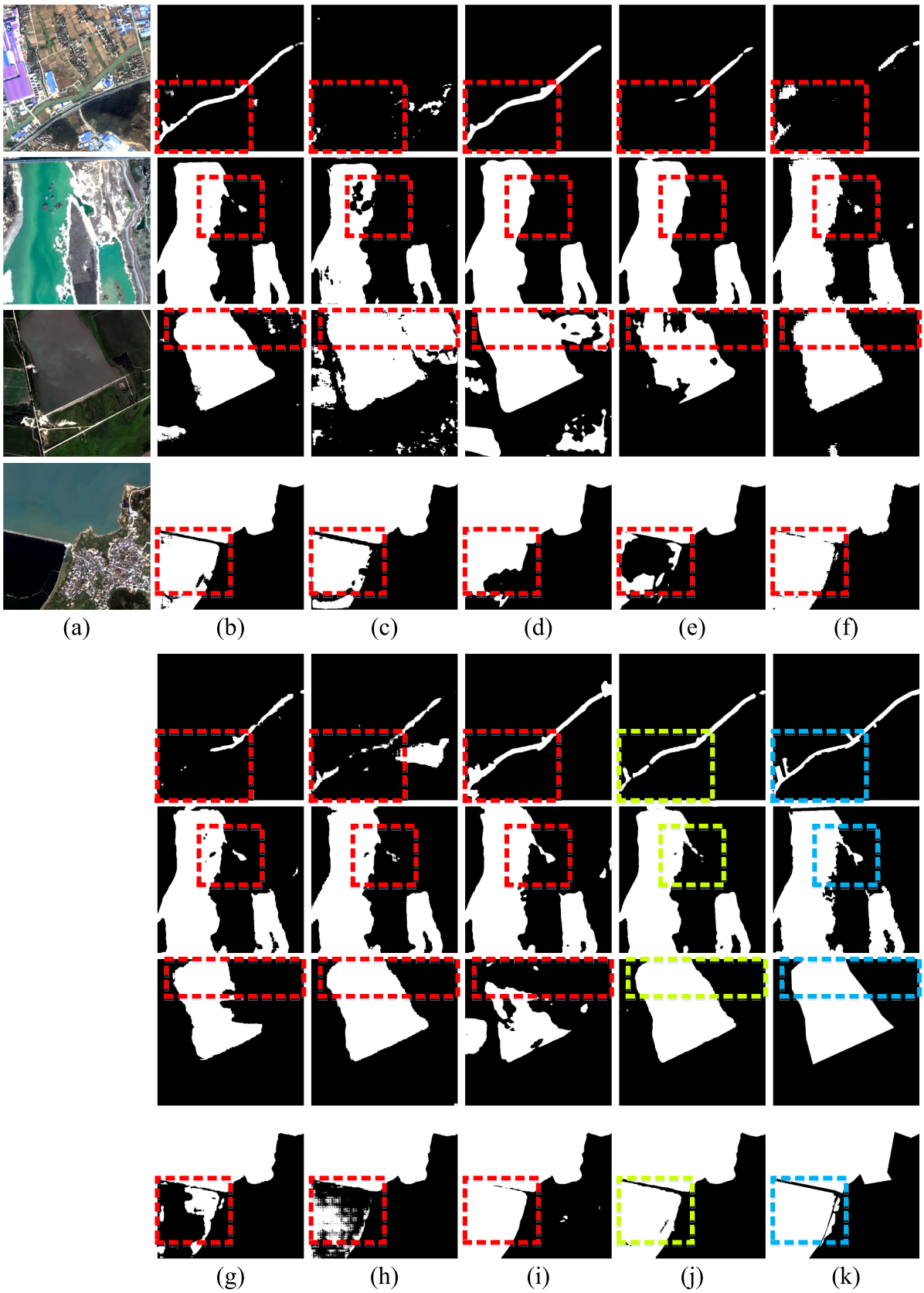

Fig. 8. Visual analysis of WD results using the GID: (a) Raw image (b) LinkNet; (c) DeepLabV3+; (d) DANet;(e) CCNet; (f) SwinUNet; (g) CSwin; (h) BANet; (i) TransFuse; (j) WaterFormer; and (k) GT.

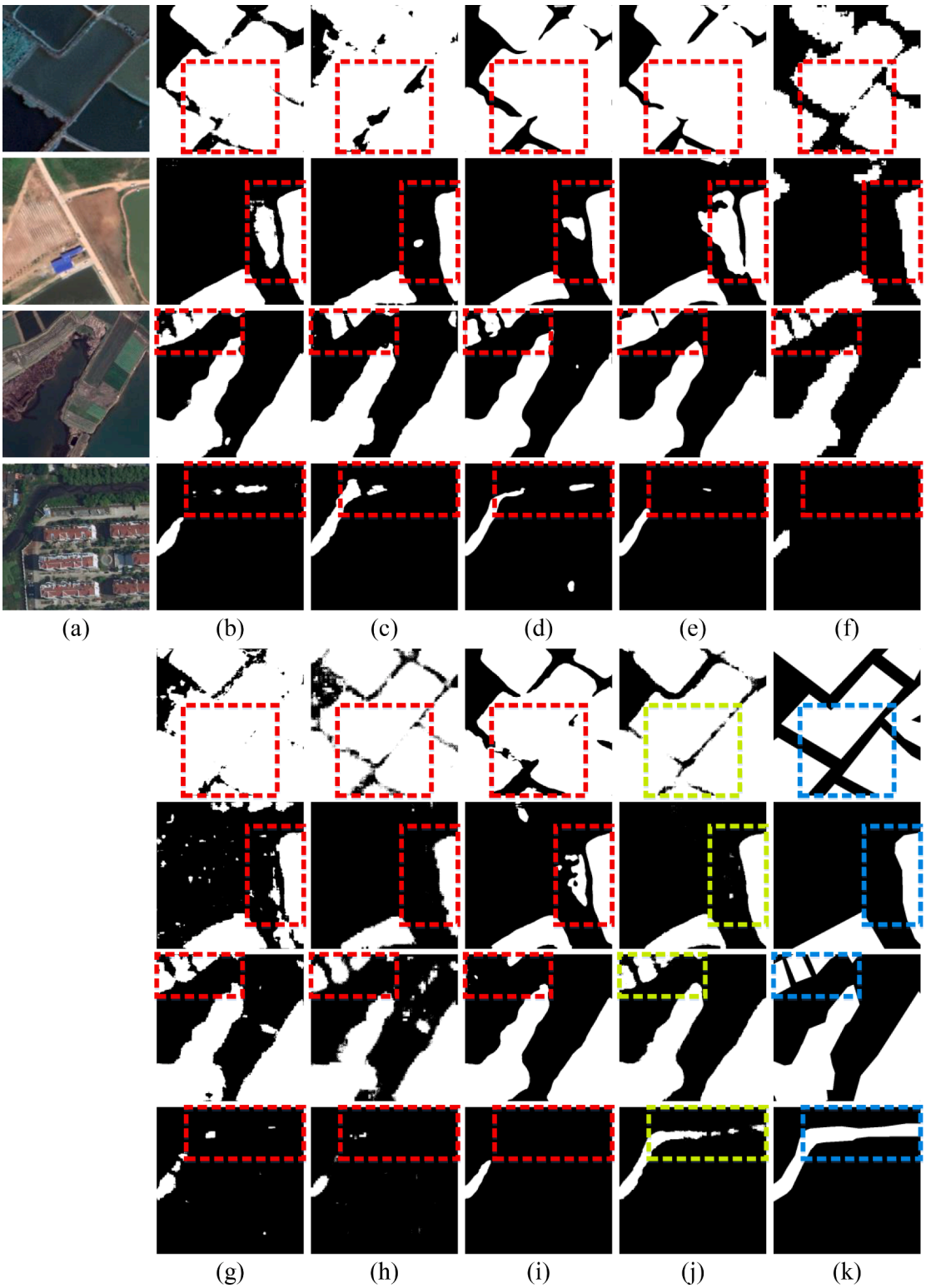

Fig. 9. Visual analysis of WD results using the LoveDA: (a) Raw image (b) LinkNet; (c) DeepLabV3+; (d) DANet; (e) CCNet;(f) SwinUNet; (g) CSwin; (h) BANet; (i) TransFuse; (j) WaterFormer; and (k) GT.

canals, our WaterFormer obtained the WD maps more consistent with the GT.

Our WaterFormer obtained the best WD accuracies with OA, P, R, F1, and KC values reaching 98.62 %, 97.62 %, 96.04 %, 96.83 %, and 0.9595, respectively, as shown in Table 3. The WaterFormer outperformed all CNN-based methods, obtaining an increase of 0.84 %, 1.58 %, 2.30 %, 1.97 %, and 2.50 % of OA, P, R, F_1, and KC, respectively. These values are 0.54 %, 0.79 %, 1.73 %, 1.28 %, and 1.62 %, respectively, compared to the SwinUNet and CSwin. This might be because the pure Transformer-based methods advanced the establishment of long-range dependencies in the pixel-level segmentation tasks. Compared with the BANet, the WaterFormer obtained an increase of 0.94 %, 1.29 %, 3.12 %, 2.23 %, and 2.83 % of OA, P, R, F_1, and KC. The reasons might be that the feature encoding ability of the light-weight ResT-Lite,

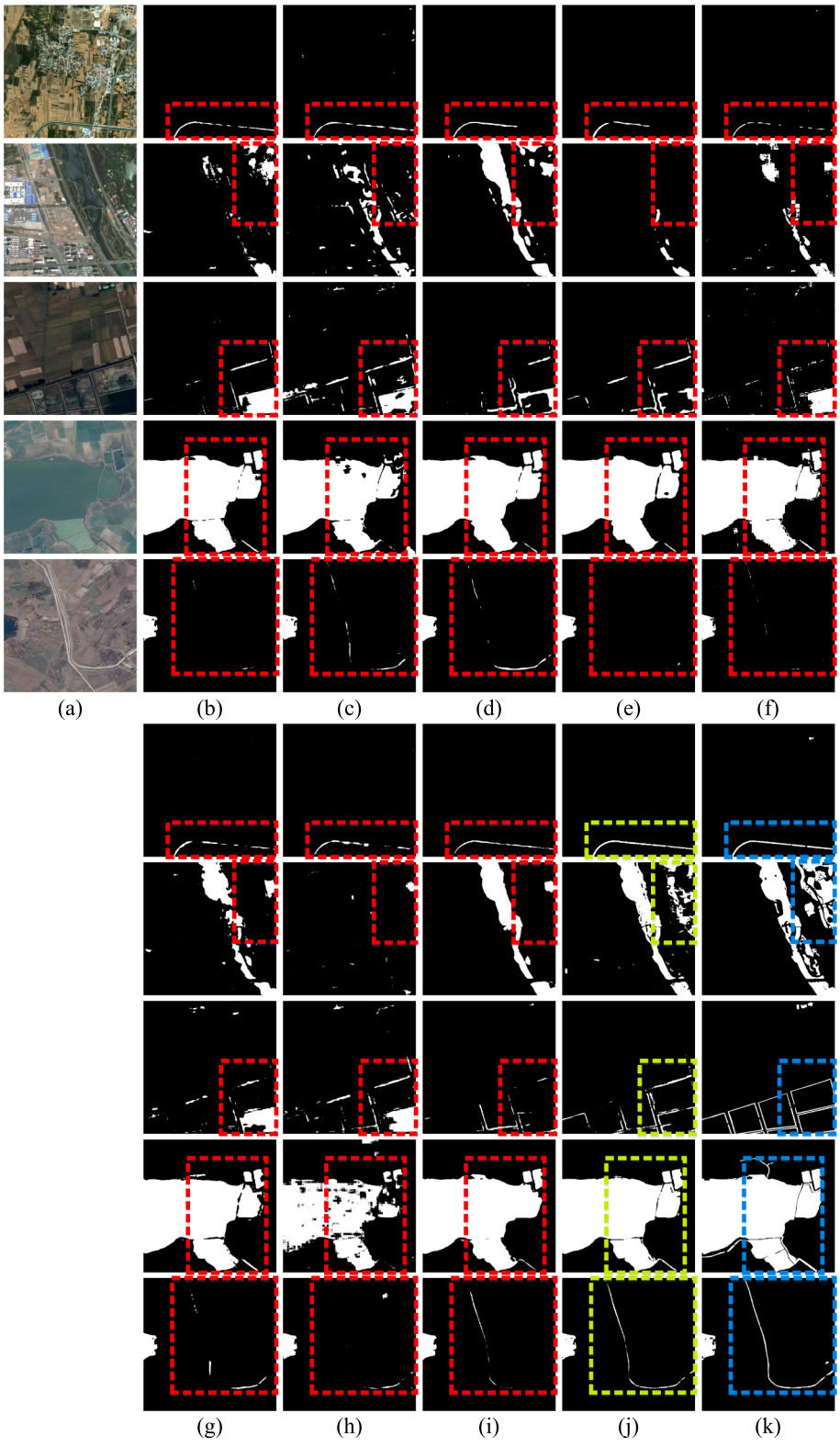

Fig. 10. Visual analysis of WD results using the MSRWD: (a) Raw image (b) LinkNet; (c) DeepLabV3+; (d) DANet; (e) CCNet; (f) SwinUNet; (g) CSwin; (h) BANet; (i) TransFuse; (j) WaterFormer; and (k) GT.

embedded in the Transformer-branch of the BANet, is seriously poor, while the utilization of multi-scale strengthened features is more effective than that of the aggregated features to achieve semantic representation for these great challenging datasets. Finally, the WaterFormer achieved a gain of 0.20 %, 1.10 %, 0.48 %, and 0.60 % for OA, R, F_1, and KC, compared to the TransFuse. The statistics also verified that the CL-ViT module contributed to long-range dependency. Through the comparative analyses, we could draw a conclusion that, owing to the superiority of mapping spatial local features into global context information, our WaterFormer was capable of abstracting waterbody semantic features from HR images.

Table 3
Quantiaive comparison results of GID, LOVEDA, and MSRWD (BOLD: BEST; BOLD-ITALIC: SECOND BEST).

Methods	GID OA(%)/P(%)/R (%)/F_1(%)/KC	LoveDA OA(%)/P(%)/R (%)/F_1(%)/KC	MSRWD OA(%)/P(%)/R (%)/F_1(%)/KC
LinkNet	95.80/96.98/ 94.03/95.48/ 0.9156	94.62/86.13/ 80.00/82.96/ 0.7977	98.28/97.47/ 94.60/96.01/ 0.9492
DeepLabV3+	93.73/94.53/ 92.04/93.27/ 0.8741	94.51/87.63/ 77.40/82.20/ 0.7897	97.14/95.58/ 91.15/93.31/ 0.9150
DANet	96.18/95.53/ 96.40/95.97/ 0.9234	94.91/87.53/ ***80.37/83.80***/ ***0.8079***	97.47/93.70/ 94.81/94.25/ 0.9264
CCNet	96.78/97.85/ 95.26/96.54/ 0.9353	94.59/86.71/ 79.09/82.72/ 0.7953	98.22/97.40/ 94.39/95.87/ 0.9474
SwinUNet	97.35/97.76/ 96.60/97.18/ 0.9469	94.20/***88.66***/ 74.02/80.68/ 0.7731	98.04/97.14/ 93.79/95.44/ 0.9419
CSwin	97.60/**98.47**/ 96.41/***97.43***/ 0.9519	94.38/87.27/ 76.87/81.74/ 0.7844	98.12/96.52/ 94.83/95.66/ 0.9447
BANet	96.96/96.92/ 96.62/96.77/ 0.9390	92.73/82.06/ 71.14/76.21/ 0.7195	97.68/96.33/ 92.92/94.60/ 0.9312
TransFuse	***97.78***/97.11/ **98.22**/97.66/ ***0.9556***	***94.94***/**91.31**/ 76.38/83.18/ 0.8024	***98.42***/**97.79**/ ***94.94***/***96.35***/ ***0.9535***
WaterFormer	**97.81**/***98.44***/ ***96.89***/**97.66**/ **0.9560**	**95.30**/87.71/ **82.91**/**85.24**/ **0.8245**	**98.62**/***97.62***/ **96.04**/**96.83**/ **0.9595**

4.6. Comparison with SOTA WD methods

To further substantiate the advantages of our WaterFormer, a comparative analysis was conducted between the WaterFormer and the existing SOTA WD methods. Fig. 11 present a visualization comparison of the WD results on the GID, LoveDA, and MSRWD datasets. The comparative methods and the GT were marked by red and blue boxes, respectively. The yellow boxes represent the WaterFormer. Additionally, TABLE 4 reports the quantitative verification results of the WD results acquired by those comparative methods.

a) ***Quantitative analysis*****:** As shown in TABLE 4, our WaterFormer performed obvious superiority over all the SOTA WD methods in terms of OA, R, F_1, and KC. Specifically, the WaterFormer acquired the best OA, R, F_1, and KC of 97.81 %, 96.89 %, 97.66 %, and 0.9560 on the GID, respectively. For the OA, R, F_1, and KC, the values were 95.30 %, 82.91 %, 85.24 %, and 0.8245 on the LoveDA, respectively. While on the MSRWD, the values for OA, R, F_1, and KC were 98.62 %, 96.04 %, 96.83 %, and 0.9595, respectively. More specially compared with the other WD methods, the MSFENet obtained the second-best R, F_1, and KC on the three datasets. The reason behind this might be that the dense atrous convolution module embedded in the network eliminates the semantic gap and enhances waterbody feature expression, greatly improving the WD accurateness. With respect to the GID, the WaterFormer obtained a mean increase of 3.07 %, 1.58 %, 5.06 %, 3.43 %, and 6.19 % of OA, P, R, F_1, and KC, compared with the MECNet, MSCENet, MSNANet and BiSeNet, respectively. The MSAFNet obtained the best P indices of 99.62 % and 92.51 % on the GID and the LoveDA, respectively. However, the P value of the MSAFNet was lower than other methods on the MSRWD. This is primarily because the MSAFNet utilizes the pre-trained ResNet-50 backbone to generate multi-level feature maps, which indicates high P values of the GID and LoveDA, while the MSRWD is more complicated and challenging in comparison to the former datasets. Relative to the MECNet, the proposed WaterFormer had a small decrease in P on the MSRWD. Specifically, its P dropped by 1.35 % on the MSRWD. The reason might be that the deep supervision strategy is adopted to impose loss constraints to raise the P value of the WD results. This loss function employed in the MECNet was demonstrated to be more effective compared with the DCE loss on the MSRWD.

b) ***Qualitative analysis:*** Visual inspection is also performed to analyze the effectiveness of the proposed WD method. Fig. 11 indicates that the proposed WaterFormer outperformed the comparative methods. The WaterFormer obtained the WD results, extremely consistent with the GT than most of the comparative methods. Specifically, as observed in the first row of Fig. 11, the city and the tiny cross-river bridge surrounded by the waterbody, and the artificial buildings with various colors, shapes, and styles, could cause false positives and negatives in the WD task. However, the WaterFormer obtained the best visual performance, and the WD results were closely aligned with the GT. In the second and third rows of Fig. 11, most of the WD methods obtained poor WD results. Only the WaterFormer plainly and accurately delineated the fine waterbody boundaries. In particular, as shown in the fourth row of Fig. 11, there were three distinct waterbody regions distributed, with different spectral brightness. Note that the WaterFormer, MSCENet, and MSAFNet correctly and completely delineated the waterbody regions, marked by the dashed rectangles, whereas the rest of the methods obtained poor WD results with false alarms and uncertain areas. This is primarily because the transfer learning pretrained strategy utilized in the MSCENet and MSAFNet accelerates network training convergence and enhances the waterbody feature extraction ability.

Additionally, as seen in Fig. 11, ZY-3, GF-2, and Google images in the MSRWD were arranged from the fifth to the eighth rows. The first two low spatial resolution images exhibited fine-broken spatial distribution waterbodies with significant global contextual semantic knowledge. While the last high spatial resolution image showed delicate and exquisite waterbodies with rich spatial local information. Compared with the six WD methods, our WaterFormer accurately and completely detected the narrow and tiny waterbody areas in the complicated urban scenes. Moreover, for the widely distributed waterbodies in the suburban scenes, the WaterFormer generated the WD maps with smoother boundaries and more veracious. The reason is that the WaterFormer obtained superior performance lying in the integration of fine-grained spatial features and global contextual semantic information in a hierarchical and collaborative way.

4.7. Computational efficiency analysis

The floating-point operations (Flops), number of parameters (Params), and weight parameters size were summarized to explore the complexity of the experimental methods, as shown in Fig. 12. Notably, our WaterFormer had 45.09G Flops, 32.082 M Params, and 131 MB weight parameters size. Among all the methods, although the BANet required the smallest numbers of the involved Params, Flops, and weight parameters size, its WD performance was modest, which demonstrated the low parameter utilization rate. As shown in Fig. 12(a1) -(c1), DeepLabv3+, DANet, and CCNet had the highest value of Flops, Params, and weight parameters size because of the complicated features encoder part in these networks. Besides, the SwinUNet, BANet, and WaterFormer spent the numbers of Flops smaller than other DL segmentation methods, which also confirmed that our WaterFormer was relatively light-weight and uncomplicated in model complexity. Furthermore, compared with the BiSeNet and MSAFNet, the WaterFormer had the higher value of Flops, Params and weight parameters size, as presented in Fig. 12(a2) -(c2). This is due to the use of the dual-stream network and the CL-ViT module.

Additionally, we analyzed the computational consumption of our WaterFormer by comparing it with the eight DL segmentation and the six WD methods. Fig. 13 and Fig. 14 show the training time and inference time comparison of nine methods on the three datasets. As shown in Fig. 13(a1) -(c1), the LinkNet, SwinUNet, and CSwin methods, which were a single-branch with the encoding–decoding structure, required the lowest mean training time on the three datasets. On the contrary, the DeepLabV3+, DANet, and CCNet methods required a longer training

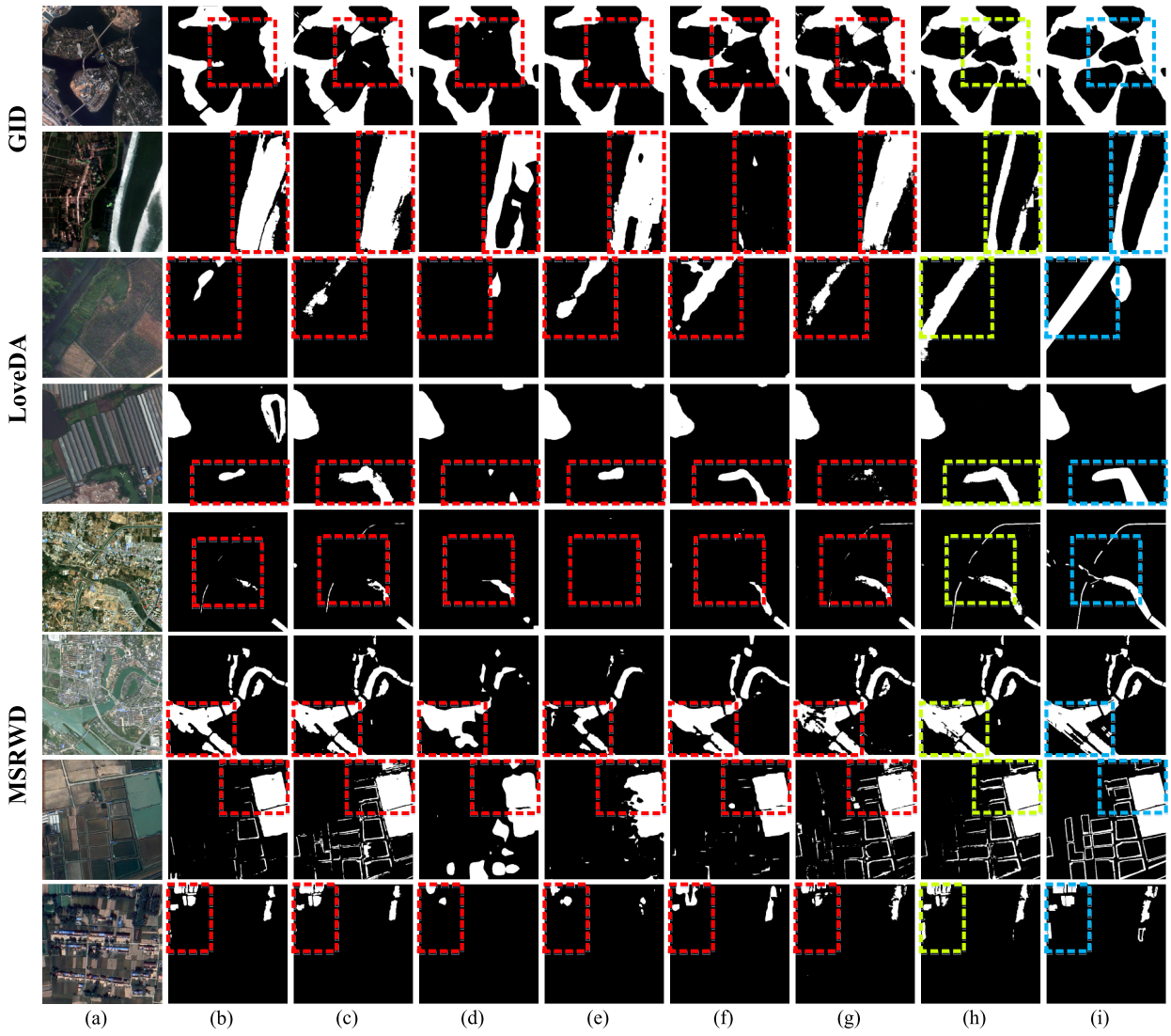

Fig. 11. Visual analysis of WD results using the MSRWD: (a) Raw image (b) MECNet; (c) MSCENet; (d) MSNANet; (e) BiSeNet; (f) MSAFNet; (g) MSFENet; (h) WaterFormer; (i) GT.

time due to the higher complexity of their network architectures by utilizing atrous convolutions. Note that, compared with BANet and TransFuse, the WaterFormer obtained an average increase of 1.32 h and 1.35 h for the training times on the LoveDA and MSRWD datasets, respectively. The reason might be that the dynamic weighted binary-cross-entropy loss, composed of two BCE losses, was employed in the WaterFormer, which required more training time. However, for the GID dataset, the whole training time is only 21.793 h, comparable to the BANet and TransFuse. As presented in Fig. 13(a2) -(c2), except for the BiSeNet and MSAFNet, our WaterFormer took the least time to complete model training on the three datasets. Moreover, the inference time of all methods were shown in Fig. 14(a2) -(c2). The time that our WaterFormer generated the WD result map was 39.042 s, 77.541 s, and 39.106 s on the GID, LoveDA, and MSRWD, respectively. Therefore, the WaterFormer achieved a better balance between WD accuracies and computational efficiencies.

4.8. Ablation analysis

To reveal the effectiveness of the two segmentation branches, the CL-ViT module, the SUS module, and the LWA module, this section presents several ablation studies. Fig. 15 shows a visualization comparison of the ablation studies and Table 5 reports the quantitative WD results. Note that, we named the WaterFormer that removed the contextual branch and the spatial branch as WaterFormer-S-branch and WaterFormer-C-branch, respectively. We named the WaterFormer only with the SUS module as WaterFormer-SUS and termed the WaterFormer without the LWA module as WaterFormer-SUS-CLViT. Finally, the bilinear upsampling algorithm was replaced with the SUS module in the WaterFormer-LWA-CLViT.

1) Effect of the segmentation branch: The WD experimental results of the WaterFormer-S-branch, the WaterFormer-C-branch, and the WaterFormer were analyzed. As shown in the second, third, seventh columns of Fig. 15, our WaterFormer generated the waterbody heat map, extremely consistent with the GT than those of WaterFormer-S-branch and WaterFormer-C-branch.

Compared with the WaterFormer-S-branch and WaterFormer-C-branch, our WaterFormer acquired a mean increase of 5.71 %, 2.94 %, and 1.29 % for OA, 6.22 %, 11.82 %, and 3.07 % for F_1, 11.47 %, 13.38 %, 3.88 % for KC on the GID, LoveDA, and MSRWD, respectively, as seen in TABLE 5. This is because the WaterFormer can accurately and completely detect waterbodies with varied shapes and different spatial resolutions by integrating fine-grained spatial features and global contextual semantic information.

2) Effect of the CL-ViT module: Fig. 15 presents the influence of the CL-ViT module on the performance of the method trained with the GID,

Table 4
Quantiaive comparison results of GID, LOVEDA, and MSRWD (BOLD: BEST; BOL-ITALIC: SECOND BEST).

Methods	GID OA(%)/P(%)/R (%)/F_1(%)/KC	LoveDA OA(%)/P(%)/R (%)/F_1(%)/KC	MSRWD OA(%)/P(%)/R (%)/F_1(%)/KC
MECNet	95.57/97.69/ 92.81/95.19/ 0.9110	94.45/87.82/ 76.72/81.89/ 0.7864	98.25/**98.97**/ 92.98/95.88/ 0.9477
MSCENet	96.38/96.16/ 96.17/96.17/ 0.9275	95.19/***90.80***/ 78.60/84.26/ 0.8144	98.31/97.41/ 94.82/96.09/ 0.9502
MSNANet	91.69/96.64/ 85.35/90.64/ 0.8323	94.29/87.39/ 76.11/81.36/ 0.7801	97.06/96.67/ 89.64/93.02/ 0.9117
BiSeNet	95.32/96.96/ 92.99/94.93/ 0.9059	94.88/87.34/ 80.33/83.69/ 0.8066	97.55/***97.70***/ 90.96/94.21/ 0.9266
MSAFNet	94.65/**99.62**/ 89.00/94.01/ 0.8922	***95.27***/**92.51**/ 77.35/84.26/ 0.8150	97.46/94.05/ 94.36/94.20/ 0.9258
MSFENet	***96.48***/95.91/ ***96.65***/***96.28***/ ***0.9294***	95.15/88.27/ ***81.17***/***84.57***/ ***0.8170***	***98.36***/97.22/ ***95.22***/***96.21***/ ***0.9517***
WaterFormer	**97.81**/***98.44***/ **96.89**/**97.66**/ **0.9560**	**95.30**/87.71/ **82.91**/**85.24**/ **0.8245**	**98.62**/97.62/ **96.04**/**96.83**/ **0.9595**

LoveDA, and MSRWD datasets. The WaterFormer-SUS-CLViT achieved the better performance against the WaterFormer-SUS by reducing the false alarms and improving the integrity of the small-thin waterbodies, thus generating a finer detection map, as shown in the second and sixth rows of Fig. 15. The fourth and seventh rows of Fig. 15 show the waterbodies surrounded by vegetation and buildings with strong spectral reflectivity. Moreover, the WaterFormer-SUS-CLViT generated the promising WD maps, closed to the GT.

As seen in Table 5, the CL-ViT module improved the WD performance, with an OA increase of 1.03 %, 0.46 %, and 0.35 %, and an F_1 increase of 1.16 %, 1.5 %, and 0.83 %, and a KC increase of 2.08 %, 1.78 %, and 1.05 % on the GID, LoveDA, and MSRWD, respectively. That is due to the CL-ViT module constructed the dependencies between spatial and contextual information, thus achieving synergistic complementarity of global semantic knowledge and local rich detail information.

3) *Effect of the LWA module*: To validate the efficacy of the LWA module, an ablation study was performed. As shown in the first, third, and eighth rows of Fig. 15, the WD maps were highly consistent with the GT, and showed a higher confidence of waterbody pixels in each heat map, comparable to the WaterFormer-SUS and WaterFormer-SUS-CLViT.

Table 5 summarizes the quantitative results of the ablation study, in terms of OA, F_1, and KC. One can observe that the WaterFormer achieved the best performance, i.e., the OA and F1 values reaching 97.81 % and 98.62 % for the GID, as well as 97.66 % and 96.83 % for MSRWD, respectively. For the WaterFormer-SUS-CLViT, the OA and F_1 values decreased to 97.65 % and 98.46 % for the GID, and 97.50 % and 96.47 % for the MSRWD, respectively. This powerfully proved that the LWA module conduced to waterbody semantic feature enhancement while preserving feature map spatial details at different levels. However, the module degraded the OA and F_1 values to 95.30 % and 85.24 %, respectively, for the LoveDA dataset. The reason is that the LWA module poorly explores the spatial contextual relationships, especially for the LoveDA dataset with sub-meter HR-RS images.

4) *Effect of the SUS module*: The SUS module generates high-resolution waterbody feature maps and high-quality discriminative waterbody feature maps in the decoding contextual branch. We tested its performance by removing the SUS module from our WaterFormer. It can be seen that WaterFormer-LWA-CLViT reduced the false alarms and refined the waterbody boundaries, thus the WD results were inconsistent with the GT, as presented in the first row to the sixth row of Fig. 15.

Table 5 demonstrates that the proposed WaterFormer without the SUS module obtained an accuracy decrease in OA, F_1, and KC on the GID, LoveDA, and MSRWD, respectively. Specifically, its OA drops by 3.18 %, 1.04 %, and 1.28 % on the three datasets, respectively; its F_1 drops by 3.18 %, 2.93 %, and 2.82 %, respectively; and its KC drops by 6.33 %, 3.57 %, and 3.65 %, respectively. The reason might be that the SUS module can further weaken the semantic gap between the contextual branch encoder and decoder and eliminate dislocation and misalignment of waterbody features. The experimental results indicate that the SUS module is a key component in the WaterFormer for the WD task.

5. Conclusion and future work

This paper presents a coupled Transformer and CNN Network, named WaterFormer, to precisely and accurately detect waterbodies

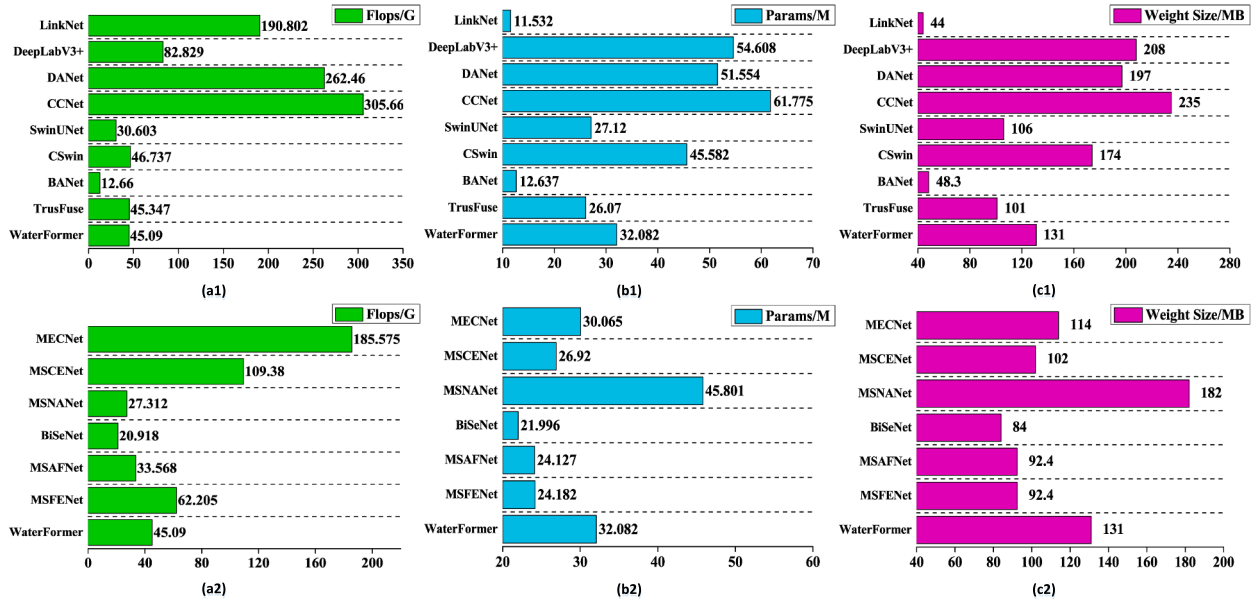

Fig. 12. Comparison of the DL and WD model properties: (a) Flops; (b) Params; (c) Weight Parameters Size.

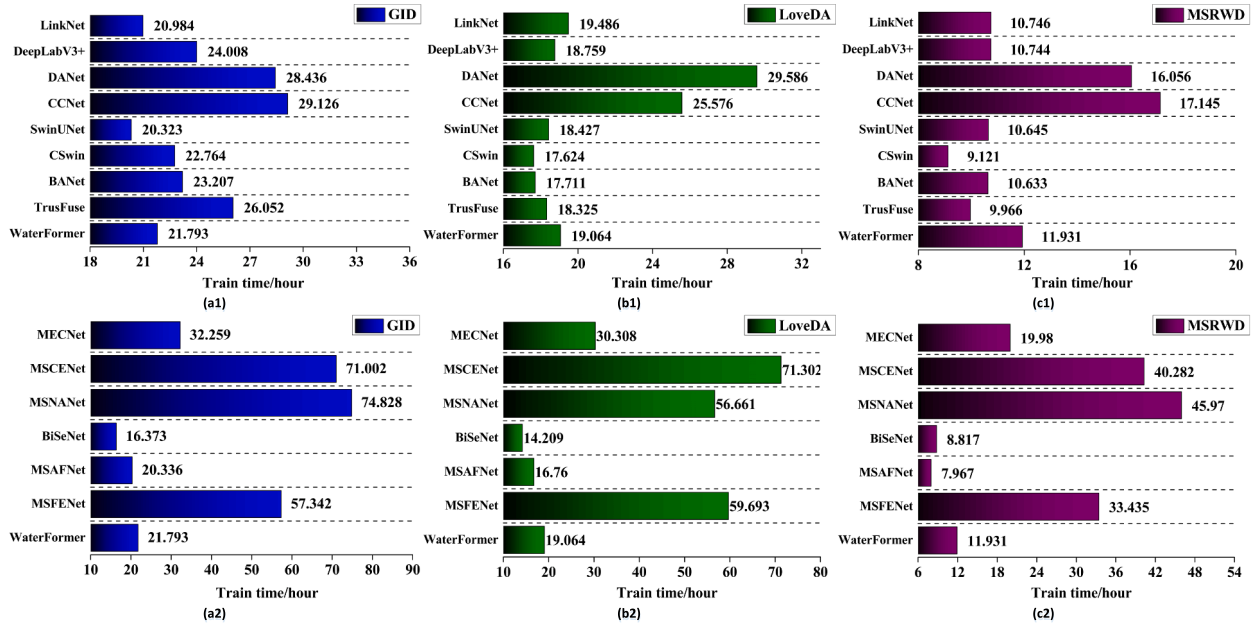

Fig. 13. Training time comparison of the SOTA DL and WD method. (a) GID;(b) LoveDA;(c) MSRWD.

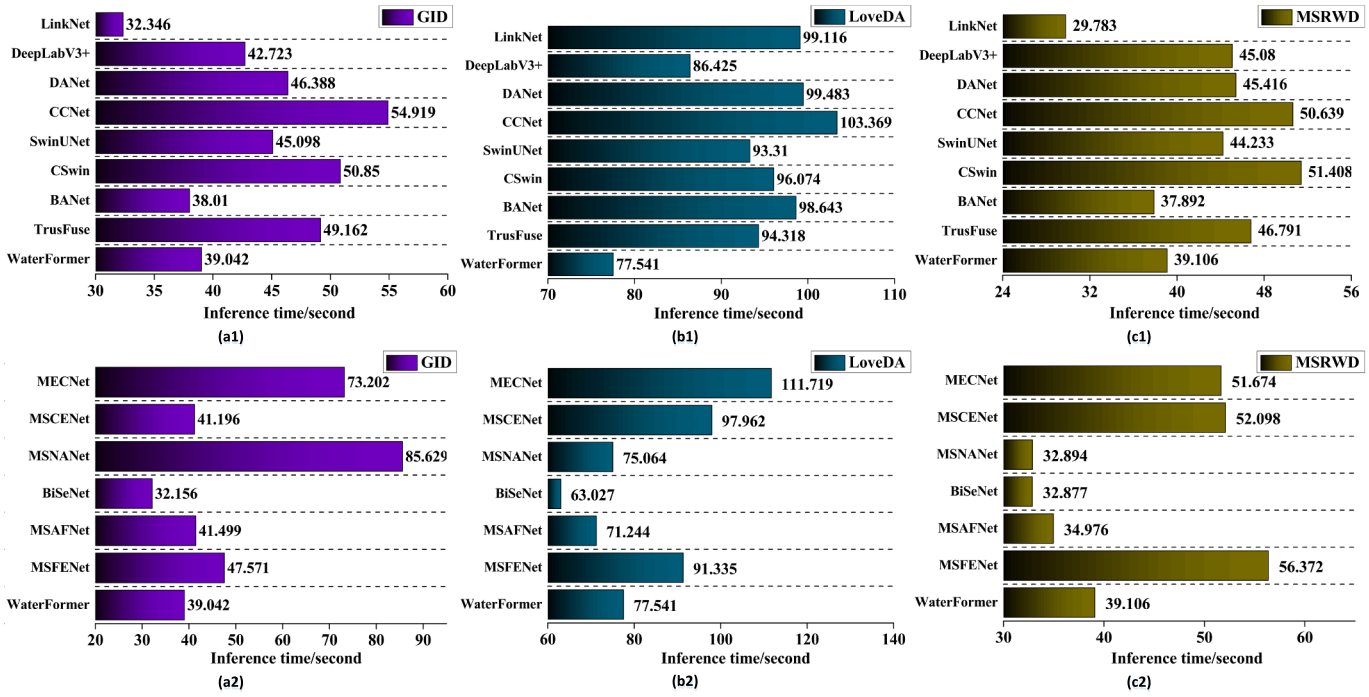

Fig. 14. Inference time comparison of the SOTA DL and WD method. (a) GID;(b) LoveDA;(c) MSRWD.

from HR-RS images. In the WaterFormer, a dual-stream deep CNN-based architecture is employed to acquire high-order and high-quality task-aware feature semantics knowledge for generating highly precise water maps. By embedding the CL-ViT into the network, the WaterFormer learned contextual properties by constructing the long-range dependence between the low-order spatial information and high-level semantic features. In addition, by adopting the LWA module, WaterFormer highlighted the feature channels tightly related to the waterbodies and concentrated on the spatial features of waterbodies, while effectively suppressing the background noise. Furthermore, by devising the SUS module over the feature expression, the WaterFormer generated high-resolution and high-quality discriminative representations of feature maps.

To facilitate the WD under different environments, we released a highly professional and high-quality MSRWD dataset. Through extensive experiments on GID, LoveDA, and MSRWD, we gained a promising performance in detecting waterbodies with large variations in size, shape, and environment. Comparative experiments are performed with the Transformer-based, CNN-based, hybrid-based methods, and the specialized waterbody detection methods, the WaterFormer better models both spatial scene information and global contextual semantic knowledge, thus achieving competitive and advantageous performance in WD tasks.

Notwithstanding the significant advancements achieved through the

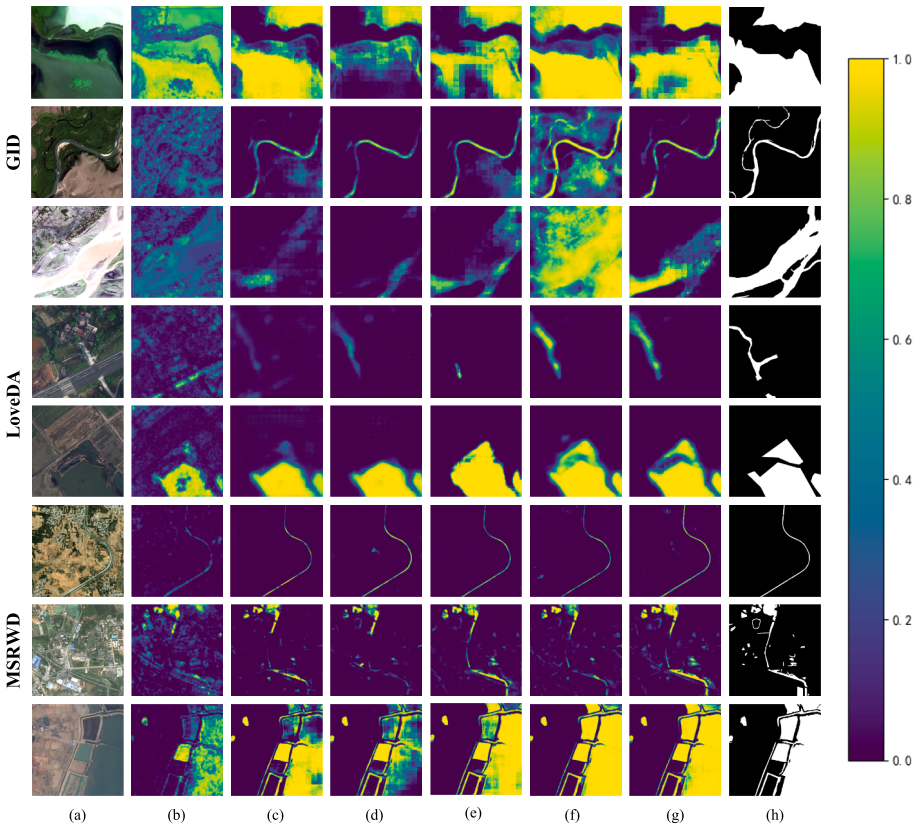

Fig. 15. Visual analysis of the ablation experiments on GID, LoveDA, and MSRWD: (a) Raw image; (b)WaterFormer-S-branch; (c) WaterFormer-C-branch; (d) WaterFormer-SUS; (e) WaterFormer-SUS-CLViT; (f) WaterFormer-LWA-CLViT; (g) WaterFormer;(h) GT.

Table 5
Quantiaive comparison results of the ablation study on GID, LOVEDA, and MSRWD.

Model	GID OA(%)/F_1(%)/ KC	LoveDA OA(%)/F_1(%)/ KC	MSRWD OA(%)/F_1(%)/ KC
WaterFormer-S-branch	89.66/88.90/ 0.7923	89.99/64.54/ 0.5889	96.42/91.61/ 0.8934
WaterFormer-C-branch	94.55/93.98/ 0.8903	94.74/82.30/ 0.7925	98.24/95.92/ 0.9480
WaterFormer-SUS	96.62/96.34/ 0.9321	95.15/84.47/ 0.8160	98.11/95.64/ 0.9444
WaterFormer-SUS-CLViT	97.65/97.50/ 0.9529	**95.61/85.97/ 0.8338**	98.46/96.47/ 0.9549
WaterFormer-LWA-CLViT	94.63/94.48/ 0.8927	94.26/82.31/ 0.7888	97.34/94.01/ 0.9230
WaterFormer	**97.81/97.66/ 0.9560**	95.30/85.24/ 0.8245	**98.62/96.83/ 0.9595**

proposed WaterFormer in this paper, it is important to acknowledge its inherent shortcomings in the field of detecting waterbodies from optical RS images. Firstly, there is a pressing necessity to reinforce the generalization performance of the proposed method to accommodate the diverse range of images characterized by low resolutions, temporal variations, and large geographical regions. Secondly, advanced WD networks will be developed in future research to achieve superior accuracy while minimizing time costs. Finally, amidst the multifaceted and diverse advancements in RS sensors, a large volume of multimodal and heterogeneous data (e.g., SAR and hyperspectral images) can be utilized for WD tasks to compensate for the limitations of optical imagery.

Funding Sources

This research was partially funded by the National Natural Science Foundation of China under Grants No. 41971414 and 42101451, and also partially funded by the Emerging Interdisciplinary Project of Central University of Finance and Economics from China.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Aroma, R.J., Raimond, K., Estrela, V.V., 2023. A coastal band spectral combination for water body extraction using Landsat 8 images. Int. J. Environ. Sci. Technol. https://doi.org/10.1007/s13762-023-05027-z.

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M., 2022. Swin-Unet: Unet-Like Pure Transformer for Medical Image Segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 205–218. https://doi.org/10.1007/978-3-031-25066-8_9.

Chaurasia, A., Culurciello, E., 2017. In: LinkNet: Exploiting Encoder Representations for Efficient Semantic Segmentation, in, pp. 1–4. https://doi.org/10.1109/VCIP.2017.8305148.

Chen, F., Chen, X., Van de Voorde, T., Roberts, D., Jiang, H., Xu, W., 2020. Open water detection in urban environments using high spatial resolution remote sensing imagery. Remote Sens. Environ. 242 https://doi.org/10.1016/j.rse.2020.111706.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 801–818. https://doi.org/10.1007/978-3-030-01234-2_49.

Demir, I., Koperski, K., Lindenbaum, D., Pang, G., 2018. DeepGlobe 2018: A Challenge to Parse the Earth through Satellite Images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 172–181. https://doi.org/10.1109/cvprw.2018.00031.

Ding, L., Lin, D., Lin, S.F., et al., 2022. Looking Outside the Window: Wide-Context Transformer for the Semantic Segmentation of High-Resolution Remote Sensing

Images. IEEE Trans. Geosci. Remote Sens. 60, 1–13. https://doi.org/10.1109/TGRS.2022.3168697.
Dong, X., Bao, J., Chen, D., Zhang, W., 2022. CSWin Transformer: A General Vision Transformer Backbone with Cross-Shaped Windows. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr52688.2022.01181.
Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., 2021. An image is worth 16×16 words: Transformers for image recognition at scale. In: Proceedings of the International Conference on Learning Representations (ICLR), pp. 1–20.
Elmi, O., Tourian, M., Sneeuw, N., 2016. Dynamic River Masks from Multi-Temporal Satellite Imagery: An Automatic Algorithm Using Graph Cuts Optimization. Remote Sens. 8 https://doi.org/10.3390/rs8121005.
Feng, W., Sui, H., Huang, W., Xu, C., An, K., 2019. Water Body Extraction From Very High-Resolution Remote Sensing Imagery Using Deep U-Net and a Superpixel-Based Conditional Random Field Model. IEEE Geosci. Remote Sens. Lett. 16, 618–622. https://doi.org/10.1109/lgrs.2018.2879492.
Fu, J., Liu, J., Bao, Y., Tian, H., Fang, Z., Li, Y., Lu, H., 2019. Dual Attention Network for Scene Segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3141–3149. https://doi.org/10.1109/cvpr.2019.00326.
Gašparović, M., Singh, S.K., 2022. Urban surface water bodies mapping using the automatic k-means based approach and sentinel-2 imagery. Geocarto Int. https://doi.org/10.1080/10106049.2022.2148757.
Hariharan, B., Malik, J., Ramanan, D., 2012. Discriminative Decorrelation for Clustering and Classification. In: Proceedings of the European Conference on Computer Vision (ECCV). https://doi.org/10.1007/978-3-642-33765-9_33.
He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep Residual Learning for Image Recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778. https://doi.org/10.1109/cvpr.2016.90.
Hu, P.C., Chen, S.B., Huang, L.-L., et al., 2023. Road Extraction by Multiscale Deformable Transformer From Remote Sensing Images. IEEE Geosci. Remote Sens. Lett. 20, 1–5. https://doi.org/10.1109/LGRS.2023.3299985.
Huang, Z., Wang, X., Huang, L., 2020. CCNet: Criss-Cross Attention for Semantic Segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 603–612. https://doi.org/10.1109/iccv.2019.00069.
Kang, J., Guan, H., Peng, D., Chen, Z., 2021. Multi-scale context extractor network for water-body extraction from high-resolution optical remotely sensed images. Int. J. Appl. Earth Obs. Geoinf. 103 (102499), 2021. https://doi.org/10.1016/j.jag.2021.102499.
Koponena, S., Pulliainena, J., Kalliob, K., Hallikainena, M., 2002. Lake water quality classification with airborne hyperspectral spectrometer and simulated MERIS data. Remote Sens. Environ. 79, 51–59. https://doi.org/10.1016/s0034-4257(01)00238-3.
Li, Y., Dang, B., Zhang, Y., Du, Z., 2022b. Water body classification from high-resolution optical remote sensing imagery: Achievements and perspectives. ISPRS J. Photogram. Remote Sens. 187, 306–327. https://doi.org/10.1016/j.isprsjprs.2022.03.013.
Li, R., Liu, W., Yang, L., Sun, S., Hu, W., Zhang, F., Li, W., 2018. DeepUNet: A Deep Fully Convolutional Network for Pixel-Level Sea-Land Segmentation. IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens. 11, 3954–3962. https://doi.org/10.1109/jstars.2018.2833382.
Li, L., Yan, Z., Shen, Q., Cheng, G., Gao, L., Zhang, B., 2019. Water Body Extraction from Very High Spatial Resolution Remote Sensing Data Based on Fully Convolutional Networks. Remote Sens. 11 https://doi.org/10.3390/rs11101162.
Li, L., Su, H., Du, Q., Wu, T., 2021. A novel surface water index using local background information for long term and large-scale Landsat images. ISPRS J. Photogram. Remote Sens. 172, 59–78. https://doi.org/10.1016/j.isprsjprs.2020.12.003.
Li, Q., Zhong, R., Du, X., Du, Y., 2022a. TransUNetCD: A Hybrid Transformer Network for Change Detection in Optical Remote-Sensing Images. IEEE Trans. Geosci. Remote Sens. 60, 1–19. https://doi.org/10.1109/tgrs.2022.3169479.
Liu, B., Du, S.H., Bai, L.B., et al., 2023a. Water extraction from optical high-resolution remote sensing imagery: a multi-scale feature extraction network with contrastive learning. Giscience Remote Sens. 60 (1), 2166396. https://doi.org/10.1080/15481603.2023.2166396.
Liu, Q., Huang, C., Shi, Z., Zhang, S.M., 2020b. Probabilistic River Water Mapping from Landsat-8 Using the Support Vector Machine Method. Remote Sens. 12 https://doi.org/10.3390/rs12091374.
Liu, X., Wang, Z., Wan, J., et al., 2023b. RoadFormer: Road Extraction Using a Swin Transformer Combined with a Spatial and Channel Separable Convolution. Remote Sens. 15 (4), 1049. https://doi.org/10.3390/rs15041049.
Liu, H., Zheng, L., Jiang, L., Liao, M., 2020a. Forty-year water body changes in Poyang Lake and the ecological impacts based on Landsat and HJ-1 A/B observations. J. Hydrol. 589 https://doi.org/10.1016/j.jhydrol.2020.125161.
Lu, M., Fang, L.Y., Li, M.X., et al., 2022. NFANet: A Novel Method for Weakly Supervised Water Extraction From High-Resolution Remote-Sensing Imagery. IEEE Trans. Geosci. Remote Sens. 60, 1–14. https://doi.org/10.1109/TGRS.2022.3140323.
Lyu, X., Fang, Y., Tong, B., Li, X., Zeng, T., 2022. Multiscale Normalization Attention Network for Water Body Extraction from Remote Sensing Imagery. Remote Sens. 14, 4983. https://doi.org/10.3390/rs14194983.
Lyu, X., Jiang, W., Li, X., Fang, Y., Xu, Z., Wang, X., 2023. MSAFNet: Multiscale Successive Attention Fusion Network for Water Body Extraction of Remote Sensing Images. Remote Sens. 15, 3121. https://doi.org/10.3390/rs15123121.
McFeeters, S.K., 2007. The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features. Int. J. Remote Sens. 17, 1425–1432. https://doi.org/10.1080/01431169608948714.
Nie, P., Cheng, X., Song, Z.Y., Mao, M.Q., Wang, T.-T., Meng, L.K., 2023. Rethinking BiSeNet: A Lightweight Network for Urban Water Extraction. IEEE Trans. Geosci. Remote Sens. 61, 1–10. https://doi.org/10.1109/TGRS.2023.3266034.
Niu, L.F., Kaufmann, H., Xu, G.C., et al., 2022. Triangle Water Index (TWI): An Advanced Approach for More Accurate Detection and Delineation of Water Surfaces in Sentinel-2 Data. Remote Sens. 14 (21), 5289. https://doi.org/10.3390/rs14215289.
Pekel, J.F., Cottam, A., Gorelick, N., Belward, A.S., 2016. High-resolution mapping of global surface water and its long-term changes. Nature 540, 418–422. https://doi.org/10.1038/nature20584.
Rajendiran, N., Kumar, L.S., 2023. Pixel Level Feature Extraction and Machine Learning Classification for Water Body Extraction. Arab. J. Sci. Eng. 48, 9905–9928. https://doi.org/10.1007/s13369-022-07389-x.
Shih, S.F., 1985. Comparison of ELAS classification and density slicing Landsat data for water-surface area assessment. Hydrol. Appl. Space Technol. 160, 91–97.
Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z., 2016. Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), pp. 1874–1883. https://doi.org/10.1109/cvpr.2016.207.
Sun, X., Wang, P., Yan, Z., Diao, W., Lu, X., Yang, Z., Zhang, Y., Xiang, D., Yan, C., Guo, J., Dang, B., Wei, W., Xu, F., Wang, C., Hansch, R., Weinmann, M., Yokoya, N., Fu, K., 2021. Automated High-Resolution Earth Observation Image Interpretation: Outcome of the 2020 Gaofen Challenge. IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens. 14, 8922–8940. https://doi.org/10.1109/jstars.2021.3106941.
Tong, X.-Y., Xia, G.-S., Lu, Q., Shen, H., Li, S., You, S., Zhang, L., 2020. Land-cover classification with high-resolution remote sensing images using transferable deep models. Remote Sens. Environ. 237 https://doi.org/10.1016/j.rse.2019.111322.
Vorosmarty, C.J., Green, P., Salisbury, J., Lammers, R.B., 2000. Global water resources: vulnerability from climate change and population growth. Science 289 (5477), 284–288. https://doi.org/10.1126/science.289.5477.284.
Wang, L., Fang, S., Meng, X., Li, R., 2022a. Building Extraction With Vision Transformer. IEEE Trans. Geosci. Remote Sens. 60, 1–11. https://doi.org/10.1109/tgrs.2022.3186634.
Wang, Z., Gao, X., Zhang, Y., Zhao, G., 2020b. MSLWENet: A Novel Deep Learning Network for Lake Water Body Extraction of Google Remote Sensing Images. Remote Sens. 12 https://doi.org/10.3390/rs12244140.
Wang, L., Li, R., Wang, D., Duan, C., Wang, T., Meng, X., 2021b. Transformer Meets Convolution: A Bilateral Awareness Network for Semantic Segmentation of Very Fine Resolution Urban Scene Images. Remote Sens. 13 https://doi.org/10.3390/rs13163065.
Wang, L.B., Li, R., Zhang, C., Fang, S.H., Duan, C.X., et al., 2022b. UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. ISPRS J. Photogram. Remote Sens. 190 https://doi.org/10.1016/j.isprsjprs.2022.06.008.
Wang, X., Ling, F., Yao, H., Liu, Y., Xu, S., 2019. Unsupervised Sub-Pixel Water Body Mapping with Sentinel-3 OLCI Image. Remote Sens. 11 https://doi.org/10.3390/rs11030327.
Wang, G., Wu, M., Wei, X., Song, H., 2020a. Water Identification from High-Resolution Remote Sensing Images Based on Multidimensional Densely Connected Convolutional Neural Networks. Remote Sens. 12 https://doi.org/10.3390/rs12050795.
Wang, X., Xie, S., Zhang, X., Chen, C., Guo, H., Du, J., Duan, Z., 2018. A robust Multi-Band Water Index (MBWI) for automated extraction of surface water from Landsat 8 OLI imagery. Int. J. Appl. Earth Obs. Geoinf. 68, 73–91. https://doi.org/10.1016/j.jag.2018.01.018.
Wang, J., Zheng, Z., Ma, A., Lu, X., Zhong, Y., 2021a. LoveDA: A Remote Sensing Land-Cover Dataset for Domain Adaptive Semantic Segmentation. Proceedings of the Conference and Workshop on Neural Information Processing Systems(NeurIPS).
Yang, X., Qin, Q., Grussenmeyer, P., Koehl, M., 2018. Urban surface water body detection with suppressed built-up noise based on water indices from Sentinel-2 MSI imagery. Remote Sens. Environ. 219, 259–270. https://doi.org/10.1016/j.rse.2018.09.016.
Yang, L., Zhang, R.Y., Li, L., Xie, X., 2021. SimAM: A Simple, Parameter-Free Attention Module for Convolutional Neural Networks. In: Proceedings of the 38th International Conference on Machine Learning. https://doi.org/10.1007/s11263-019-01283-0.
Yao, F., Wang, C., Dong, D., Luo, J., Shen, Z., Yang, K., 2015. High-Resolution Mapping of Urban Surface Water Using ZY-3 Multi-Spectral Imagery. Remote Sens. 7, 12336–12355. https://doi.org/10.3390/rs70912336.
Zhang, S., Cao, Y.G., Sui, B.K., 2023. DTHNet: Dual-Stream Network Based on Transformer and High-Resolution Representation for Shadow Extraction from Remote Sensing Imagery. IEEE Geosci. Remote Sens. Lett. 20, 1–5. https://doi.org/10.1109/LGRS.2023.3290176.
Zhang, Y., Liu, X., Zhang, Y., Ling, X., Huang, X., 2019. Automatic and Unsupervised Water Body Extraction Based on Spectral-Spatial Features Using GF-1 Satellite Imagery. IEEE Trans. Geosci. Remote Sens. 16, 927–931. https://doi.org/10.1109/lgrs.2018.2886422.
Zhang, Y., Liu, H., Hu, Q., 2021a. TransFuse: Fusing Transformers and CNNs for Medical Image Segmentation. In: Proceedings of the Medical Image Computing and Computer Assisted Intervention Society (MICCAI), pp. 14–24. https://doi.org/10.1007/978-3-030-87193-2_2.
Zhang, W., Jiao, L., Li, Y., Huang, Z., Wang, H., 2022. Laplacian Feature Pyramid Network for Object Detection in VHR Optical Remote Sensing Images. IEEE Trans. Geosci. Remote Sens. 60, 1–14. https://doi.org/10.1109/tgrs.2021.3072488.

Zhang, Z., Lu, M., Ji, S., Yu, H., Nie, C., 2021b. Rich CNN Features for Water-Body Segmentation from Very High Resolution Aerial and Satellite Imagery. Remote Sens. 13, 1912. https://doi.org/10.3390/rs13101912.

Zheng, S., Lu, J., Zhao, H., Fu, Y., 2021. Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), pp. 6877–6886. https://doi.org/10.1109/cvpr46437.2021.00681.

Zhong, H.-F., Sun, Q., Sun, H.-M., Jia, R.-S., 2022. NT-Net: A Semantic Segmentation Network for Extracting Lake Water Bodies From Optical Remote Sensing Images Based on Transformer. IEEE Trans. Geosci. Remote Sens. 60, 1–13. https://doi.org/10.1109/tgrs.2022.3197402.