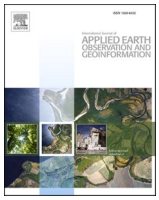
International Journal of Applied Earth Observations and Geoinformation 115 (2022) 103103

Contents lists available at ScienceDirect

International Journal of Applied Earth Observations and Geoinformation

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



WaterHRNet: A multibranch hierarchical attentive network for water body extraction with remote sensing images

Yongtao Yu [a,*], Long Huang [a], Weibin Lu [a], Haiyan Guan [b], Lingfei Ma [c], Shenghua Jin [a], Changhui Yu [a], Yongjun Zhang [a], Peng Tang [a], Zuojun Liu [a], Wenhao Wang [a], Jonathan Li [d]

[a] Faculty of Computer and Software Engineering, Huaiyin Institute of Technology, Huaian, JS 223003, China
[b] School of Remote Sensing and Geomatics Engineering, Nanjing University of Information Science and Technology, Nanjing, JS 210044, China
[c] School of Statistics and Mathematics, Central University of Finance and Economics, Beijing 102206, China
[d] Department of Geography and Environmental Management, University of Waterloo, Waterloo, ON N2L3G1, Canada

ARTICLE INFO

Keywords:
Water resource mapping
High-resolution network
Feature attention
Semantic enhancement
Hierarchical segmentation

ABSTRACT

Water is a kind of vital natural resource, which acts as the lifeblood of the ecosystem and the energy source for the living and production activities of humans. Regularly mapping the conditions of water resources and taking effective measures to prevent them from pollutions and shortages are very important and necessary to maintain the sustainability of the ecosystem. As a preliminary step for image-based water resource analysis, the complete recognition and accurate extraction of water bodies are important prerequisites in many applications. Nevertheless, due to the issues of topology diversities, appearance variabilities, and land cover interferences, there is still a large gap to achieve the human-level water bodies interpretation quality. This paper presents a hierarchical attentive high-resolution network, abbreviated as WaterHRNet, for extracting water bodies from remote sensing imagery. First, by building a multibranch high-resolution feature extractor integrated with global feature semantics aggregation, the WaterHRNet behaves laudably to supply high-quality, strong-semantic feature representations. Furthermore, by inlaying an effective feature attention scheme with the comprehensive exploitation of both the spatial and channel feature significances, the WaterHRNet is forced to strengthen the semantic-determinate, task-aware feature encodings. In addition, by designing a hierarchical processing principle with the progressive enhancement of category-attentive feature semantics, the WaterHRNet performs effectively to export semantic-discriminative, target-oriented feature representations for precise water body segmentation. The WaterHRNet is elaborately verified both quantitatively and qualitatively on three remote sensing datasets. Evaluation results show that the WaterHRNet achieves an average precision of 98.44%, average recall of 97.84%, average IoU of 96.35%, and average F_1-score of 98.14%. Comparative analyses also demonstrate the superior performance and excellent feasibility of the WaterHRNet in segmenting water bodies.

1. Introduction

Water is an irreplaceably crucial part of the natural resources. Statistically, about three fourths of the earth surface is covered by water. The diversities of waters include inland, coastal, and oceanic water bodies, which supply energy and material to industrial productions and daily lives, establish pathways for transportation activities, regulate the climate, and act as the lifespring of the ecosystem. Caused by the change of the global climate, the increase of industrial pollutions, and the impacts of human activities, the surface water bodies are undergoing severe issues with respect to quality and availability, especially the freshwater resources, thereby leading to serious destructions on the stability and sustainability of the ecosystem. Reasonable exploitation and utilization of the water resources are significantly necessary to maintain the earth's water cycles. As well, reducing the consumptions and pollutions of the water resources is quite crucial to ensure the sustainable improvement of human society and the balance of the natural ecosystem. Therefore, regularly monitoring the changes and conditions of the surface water bodies can provide timely evidence to analyze the influences of environmental factors and human activities, and direct the relevant departments to take effective measures to alleviate the further deterioration of the issues.

* Corresponding author.
E-mail address: allennessy@hyit.edu.cn (Y. Yu).

https://doi.org/10.1016/j.jag.2022.103103
Received 26 July 2022; Received in revised form 18 October 2022; Accepted 9 November 2022
Available online 11 November 2022

Compared with the traditional manual survey solutions, which are inefficient and laborious, remote sensing techniques provide a cost-effective and highly-efficient solution to map water bodies in a large area. The remote sensing images collected by satellite or airborne sensors constitute the main data source used for water body mapping applications. The topological structures and spectral properties of water bodies can be well recorded and reflected in the remote sensing images. As a preliminary step in the process of water body mapping, recognizing and extracting water body contents in the remote sensing images play a crucial part in a variety of tasks, such as water body change detection (Sarp and Ozcelik, 2017), water quality analysis (Chu et al., 2018), and water body type categorization (Abid et al., 2021). Specifically, water body extraction aims at marking the image contents belonging to the water body regions in a per-pixel manner. In the literature, massive efforts have been made to investigate advanced techniques for water body extraction on the purpose of either promoting the processing efficiency or upgrading the extraction accuracy (Li et al., 2022a). At the early stages, water bodies are segmented mainly based on spectral thresholding (Nones, 2021), mathematical morphology techniques (Rishikeshan and Ramesh, 2018), water indices (Jin et al., 2021), and machine learning approaches (Chen et al., 2020). However, the performance of these approaches is easily affected by the variations of data sources and the changes of test scenarios. As a result, their robustness and universality are limited to specific cases. Recent advancement of deep learning techniques in vision tasks has brought hopeful light to the intelligent interpretation of remote sensing images. Accordingly, intensive attention has been drawn to water body extraction by exploiting effective deep network architectures. Nevertheless, due to the challenging issues of water bodies including non-rigid and arbitrary topologies, color diversities, indeterminate boundaries, size variations (especially the small-size water bodies), and complexities of the surface conditions and environmental scenarios, there is still a large gap to reach the human-level processing quality towards water body extraction. As a matter of fact, there is great space for enhancing the extraction accuracy and imperative demand for developing advanced solutions to approach the objective of high-quality water body extractions.

In this paper, we construct an advanced hierarchical segmentation network architecture for serving water body extraction from remote sensing imagery. The architecture involves a multi-resolution backbone for semantic-strong feature abstraction, a dual-path feature attention module for semantic-significant feature promotion, and a hierarchical segmentation head for task-oriented feature emphasis and high-quality water body prediction. The proposed model demonstrates excellent performance on different water body cases, such as spatial extension diversities, texture appearance variations, and self-condition and surrounding scenario changes. The contributions mainly include the following. (1) A novel lightweight and powerful feature attention principle paralleling a channel-specific attention branch and a spatial-specific attention branch is built for promoting the contributions of the information-relevant and semantic-important features from both the channel and spatial perspectives. Formulated with a different architecture from the existing feature attention mechanisms, the feature semantic encoding quality is significantly improved by suppressing the impacts of the task-irrelevant features, enhancing the differences between the foreground and background features, and attending to the features in the foreground regions with a global perspective in both the channel and spatial domains. (2) A novel hierarchical segmentation scheme cooperated with a semantic-level enhancement module is designed for emphasizing the saliencies of the category-aware semantics and exporting a high-quality, task-oriented feature representation to satisfy the accurate water body extraction. Integrated with the semantic-level enhancement module to exploit category-attentive feature semantics, the feature contrast between the background and foreground contents is significantly promoted to improve the feature semantic encoding quality. Employed by a hierarchical segmentation strategy, the segmentation results from the low-resolution feature maps can provide important region localization cues to the high-resolution feature maps and supervise the segmentation on the high-resolution feature maps to obtain accurate and detailed segmentation results.

2. Related works

A common characteristic of the deep network architectures is that hierarchical feature representations with different scales and semantic levels can be directly and effortlessly produced end-to-end scarcely requiring too much manual intervention. The output feature encodings serve robustly and precisely to portray the inherent properties of the semantic targets. As a result, due to the eye-catching success of deep network architectures in vision tasks, recent studies have devoted positively to conducting water body extraction with deep learning solutions. Generally, most of the existing water body extraction frameworks were formulated into pixel-level segmentation architectures, which can well handle the issues of size variations, non-rigid and arbitrary topologies of water bodies. Wang et al. (2020c) constructed a convolutional neural network (CNN) with a densely connected pattern for water body recognition. The dense connection architecture well supported the feature reuse and the network stability with the export of higher-quality semantics. Wang et al. (2020a) designed a fully convolutional network (FCN) by means of depthwise separable and residual convolutions for lake area identification. Similarly, multiscale dense connections were also employed to access large-range feature contents. The quality-improved feature semantics generated by using dense connections promoted the localization accuracy of the large-area water bodies. However, by depending on the high-level small-size feature maps to recover the high-resolution prediction maps, they suffered from some accuracy degradations in extracting the small-size water bodies and determining the tight water body boundaries. Li et al. (2021c) took advantage of the spectral and spatial attributes to separate water bodies with a DenseNet formulation. In this architecture, feature compression and skip connection techniques were applied to boost the feature abstraction quality. With the inclusion of spatial cues for boosting semantic differences, the water bodies showing varying colors and different contrasts were nicely extracted. However, the correct recognition of small-size water bodies was still unsolved. Wang et al. (2020b) trained a multiscale CNN model, which combined the feature semantics from multiple scales to form a robust representation for water body prediction. The multiscale semantics encoded rich information regarding the water body at different resolutions. Zhang et al. (2021b) developed a cascaded FCN structure to alleviate the resolution loss issue. Worth mentioning, the conditional random field model was connected to refine the initial segmentation results. By promoting the feature semantics and details in the high-resolution prediction maps, these models performed promisingly in recognizing the small-size water bodies and locating the water body boundaries. However, they still behaved unstably in the cases of complex environmental scenarios and surface conditions of water bodies. To sufficiently investigate the multilevel semantics, Duan and Hu (2020) stacked a multiscale refinement CNN architecture. In the refinement process, an erasing-attention component functioned to augment the feature semantics for scale-wise water body prediction. These predictions were eventually integrated to produce the refined segmentation output. Differently, Kang et al. (2021) suggested a context extractor to exploit multiscale feature contexts aiming at well depicting water bodies exhibiting varying forms. To be specific, the context information was delineated through multibranch atrous convolutions. By comprehensively considering multiscale feature contexts for semantic enhancement, the water bodies exhibiting low contrasts and heterogeneous surface properties were well segmented. However, the segmentation details were slightly coarse-grained, especially in the marginal regions. As for Zhang et al. (2021a), the feature attention mechanism was cooperated with the atrous convolutions to retrieve contextual and semantic-level details. As an alternative for multilevel feature fusion, Miao et al. (2018) formulated an encoder-decoder

architecture, which directly concatenated the feature semantics of the same scale from the encoder and decoder pathways. A high-resolution feature map was finally recovered for water body inference. In contrast, in the architecture of Yuan et al. (2021), multiscale semantics from the encoder pathway were integrated together as the auxiliary input to the decoder pathway for the purpose of feature enhancement. Xue et al. (2021) introduced a coordinate attention scheme into the encoder-decoder architecture, which functioned to highlight the important feature semantics from the spatial perspective. The integration of feature attention mechanisms and the formulation of encoder-decoder architectures enforced the models to characterize more task-oriented and semantic-stronger features, which enhanced the discriminations between the water bodies and their surrounding environments, thereby improving the extraction accuracy of the water bodies under different challenging conditions. However, some very small-size water bodies were still failed to be completely extracted.

Since the widespread use and excellent performance of the U-Net architecture in various segmentation issues, it has also been paid intensive attention to serve the water body extraction task. Due to the recovery of a high-resolution feature map with improved feature semantics and details, the U-Net architecture performed excellently when handling the small-size water bodies and the water bodies under different scenarios. Qin et al. (2022) designed a deep U-Net formulation for extracting small-area water bodies. The contracting route focused on abstracting different-level feature semantics and the expanding route combined these semantics and recovered a high-quality representation for water body segmentation. However, the extraction performance of this model differed greatly on the water bodies of different spectral properties and surface conditions. Aiming at accelerating the processing efficiency and saving the computational overhead, Wang et al. (2021b) trained a lightweight architecture based on the MobileNetV2 model. For postprocessing, morphological approaches were further applied to delete the noise in the exported water body map. As another option, Tambe et al. (2021) introduced an inception layer at each stage of the U-Net for lightweighting the model size, thereby achieving an alleviation on the computation burden. In these ways, both the time and memory consumptions were significantly reduced to well meet the real-time processing requirements, but at the cost of the extraction accuracy degradations due to the quality lowering of the encoded feature semantics. Li et al. (2021a) proposed an improved U-Net structure by adding more skip connections for the purpose of boosting the feature encoding quality. Specifically, an S-shaped circular connection scheme was employed for cross-stage feature fusion. Moreover, deep features and handcrafted features were combined to emphasize the water body semantics. In He et al. (2021), a self-attention module was mounted on the skip connection pathway of the U-Net, which functioned for forcing the network to extract task-aware features. This was achieved by promoting the contributions of the foreground features. Likewise, feature attention and pyramid modules were also taken into account for feature semantic boosting in Li et al. (2021b). By attending to the water body regions or enhancing the feature contrasts between the foreground and background contents to improve the feature representation quality and saliency, the water bodies showing low contrasts, blur boundaries, and heterogeneous spectral properties were well recognized and segmented. However, the increase of the model complexity led to the decrease of the processing efficiency and the increase of the model size. Feng et al. (2019) presented a cascaded processing pipeline for water body detection by combining the U-Net model and the conditional random field model. The U-Net pre-labelled the image contents into water bodies and the background. Then, the conditional random field model and superpixel-based region constraints were cooperated to attain finer prediction results. With the postprocessing procedure for region-wise semantic grouping, the background contents were effectively suppressed and the water body boundaries were tightly adhered. However, due to the superpixel segmentation performance variations on different image scenarios, the segmentation details might be affected in the cases of complex environmental and surface circumstances. In addition, feature pyramid architectures (Li et al., 2019b), capsule networks (Yu et al., 2021), DeepLabV3 + models (Li et al., 2019a), self-supervised learning, weakly supervised learning, unsupervised learning, and transfer learning techniques (Dang and Li, 2021; Li et al., 2022b; Lu et al., 2022; Wang et al., 2021a), and multisource geospatial data fusion strategies (Kim et al., 2021) were also investigated for tackling water body extraction issues.

Nevertheless, despite the expressive achievements made by the advanced deep learning models, there are still some challenges that impede the accurate extractions of water bodies with the remote sensing images. Typical challenges include limitations in spectral properties, variabilities in water body topologies, extensions, and distributions, complexities in water body scenarios and surface conditions, and deficiencies in large-scale annotated datasets of water bodies with rich and diverse patterns. Specifically, the precise identification of the small-size, low-contrast, and blur-boundary water bodies is still a pending issue that cannot be well solved by the existing techniques. To cope with the above challenges, we propose a hierarchical attentive high-resolution network architecture for extracting water bodies based on remote sensing images. First, by formulating a pixel-level segmentation network architecture, the proposed model can deal with water bodies of varying topologies, extensions, and distributions. Second, by employing a multibranch multi-resolution network architecture as the feature extractor with the maintenance of a high-resolution stream, different-size water bodies can be well recognized and extracted, especially the small-size water bodies. Third, by integrating a dual-path feature attention module for attending to the foreground feature semantics and suppressing the impacts of the background feature semantics, the water bodies under different environmental and surface conditions can be well differentiated, especially the low-contrast, spectral-heterogeneous, and small-size water bodies. Finally, by designing a hierarchical segmentation scheme for leveraging the segmentation results from low-resolution feature maps to supervise the segmentation on high-resolution feature maps, the water body segmentation accuracy and detail are greatly improved, especially for the blur boundaries and small-area background contents. Nevertheless, adopted by a supervised learning strategy, the proposed model still requires large amounts of annotated data for high-quality model training. In addition, the extraction performance will be degraded in some extremely challenging conditions, such as occlusions.

3. Methodology

The overall structure of the designed hierarchical attentive high-resolution network for water body extraction (WaterHRNet) is presented in Fig. 1. To be specific, the WaterHRNet comprises three main functional components: the feature extractor, the feature attention module, and the segmentation head. The feature extractor is established as a multibranch HRNet formulation (Wang et al., 2021c) and serves for abstracting target semantics in different feature subspaces. The feature attention module inlaid in all branches of the HRNet is used for promoting the produced feature semantics by attending to more task-oriented encodings. The segmentation head follows a hierarchical processing pattern and functions to successively enhance the feature representations with semantic-level constraints, eventually generating a high-quality feature map for water body prediction.

3.1. HRNet feature extractor

The HRNet architecture is a quite potential and creative design philosophy that is suitable and powerful to serve as the feature extraction backbone. Its novelty and advantage reflect in the exploitation of high-level feature semantics in different-resolution subspaces with a parallel pattern. Worth mentioning, in order to have a global perspective over all the subspaces to strengthen the output feature semantics, information exchange is constantly taken place among different feature

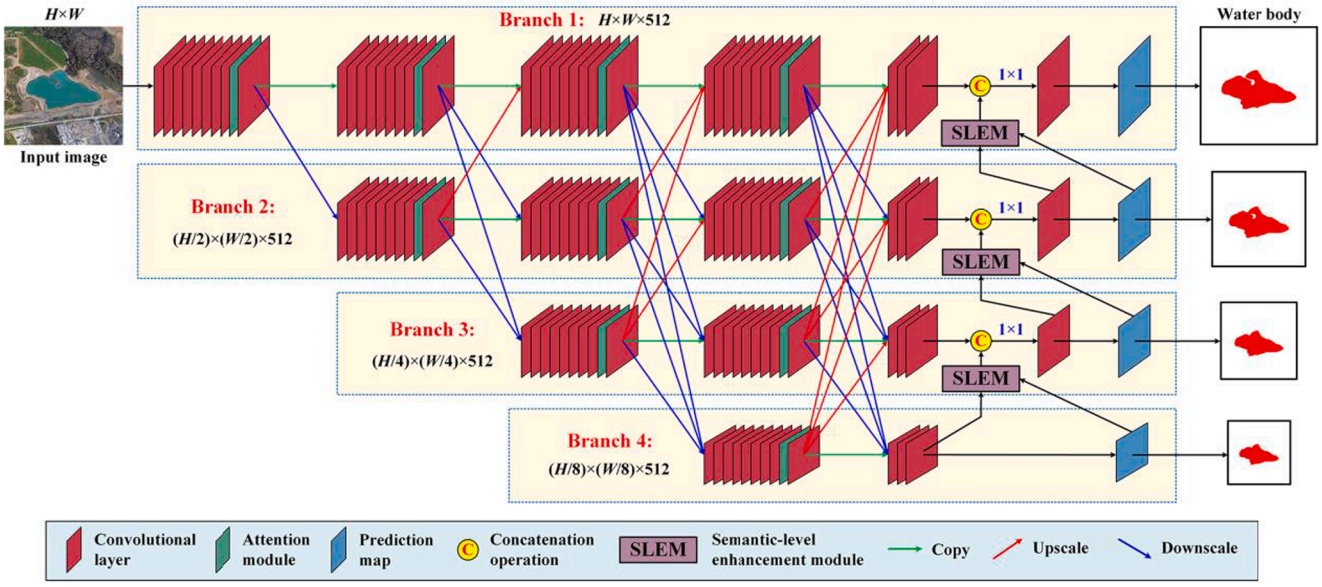


Fig. 1. Overview of the developed water body extraction network.

subspaces, eventually providing a set of strong semantics used for varying reasoning tasks. Therefore, considering the distinct attributes and excellent performance of the HRNet architecture in a broad range of vision issues, we also stack the feature extractor with the HRNet formulation aiming at supplying high-quality, strong, and task-aware feature encodings for more accurately identifying and pixel-wisely delineating the water bodies in the remote sensing images.

As depicted by Fig. 1, the HRNet feature extractor is stacked by four convolution branches, each of which hammers at mining target feature encodings in a specific subspace with a designed spatial resolution. Particularly, from Branch 1 to Branch 4, the spatial resolutions of these subspaces are gradually lowered, resulting in a lower and lower resolution feature map top-down. Note that, the sizes of the feature maps in each branch are totally the selfsame throughout, which benefits significantly for reducing the loss of the localization accuracy under each spatial resolution. Theoretically speaking, the higher-resolution feature semantics are beneficial to the determination of the small-area or elongated water bodies, while the lower-resolution feature semantics are helpful to the recognition of the large-range water bodies. As a result, the advanced parallel structure of the HRNet feature extractor perfectly satisfies the characterization and extraction of the water bodies with different spatial extensions. In the proposed WaterHRNet, Branch 1 of the HRNet feature extractor preserves the highest resolution and the identical size to the input image. The other branches are gradually shrunk in spatial resolutions and sizes by a constant scaling coefficient of 0.5 to exploit semantic attributes in a lower subspace. At the stage of cross-subspace information exchange, the higher-resolution semantics in an upper branch are downsampled into the desired resolution in the target branch based on strided convolution operations, whereas the lower-resolution semantics in a lower branch are upsampled into the desired resolution in the target branch based on deconvolution operations. Afterwards, these resolution-recalibrated semantics are concatenated to the directly copied feature semantics in the target branch and properly aggregated through a 1×1 convolution operation, eventually forming a global semantic augmented feature representation in the target branch. Precisely benefitted by the cross-subspace information exchange, the feature encodings in each branch are constantly refined by taking into account the global perspective of the feature semantics under all subspaces. At the inference stage, the set of multi-resolution feature semantics produced by the four branches will be treated as the semantic evidences to the segmentation head for water body prediction.

3.2. Feature attention module

To achieve position independence and local property interpretation, convolution operations employ a sliding window philosophy to portray the feature semantics within a receptive field that is restrained by the kernel size. The contextual semantics are gradually enriched as the layers go deeper to access larger receptive fields, thereby leading to the abstraction of higher-level feature representations. Nevertheless, there is still a fly in the ointment by merely relying on the convolutions. To be specific, the otherness and relevance of different feature channels, especially those reflecting the foreground, are not expressly taken into consideration, which behaves imperfectly to exploit strong, discriminative feature encodings. Moreover, the significance and saliency of the spatial positions, particularly those occupying the foreground, are not intently paid close attention, which performs unsatisfactorily to investigate robust, task-oriented feature encodings. As effective solutions to feature semantic quality promotion, intensive efforts have been recently made to cooperate feature attention principles with the convolutions on the purpose of explicitly emphasizing the contributions of the useful and relevant feature semantics. Thus, aiming at further enhancing the semantic quality of the exported features by the HRNet extractor, we construct a novel feature attention module and inlay it into each branch of the HRNet extractor at the position before carrying out cross-subspace information exchange for supervising feature abstraction. As Fig. 2 illustrates, the attention module comprises dual parts: a channel-specific attention branch and a spatial-specific attention branch. These two branches function to accentuate the contributions of the important feature semantics from the channel and spatial perspectives, respectively. Eventually, the attention module provides a semantic-enhanced feature representation by combining the recalibrated feature semantics from these two branches.

The channel-specific attention branch serves to increase the weights of the task-relevant feature channels. In this regard, the input feature map with the dimension of $H \times W \times C$ (H, W, and C mean the height, width, and channel number) is first transformed into feature representations $F_1 \in R^{H\times W\times C}$ and $Q_1 \in R^{H\times W\times 1}$ via two separate 1×1 convolutions. F_1 acts as a feature response map, each of whose position reflects the task-sensitive feature response corresponding to the same point on the input feature map. In other words, a larger value at a position encodes a stronger feature response. Q_1 is a spatial weighting map, which measures the significance of the feature response at a position. That is, the feature response associated with a larger weight will be considered

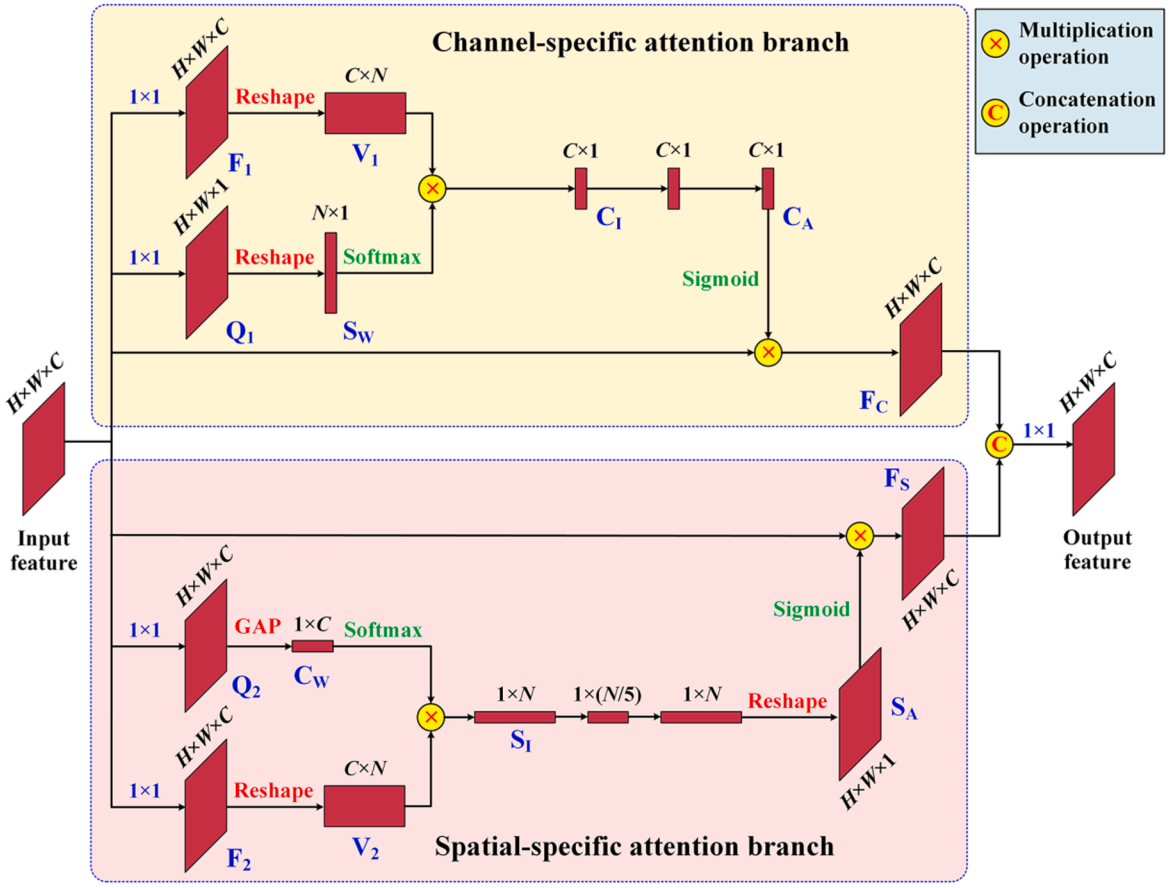


Fig. 2. Structure of the designed feature attention module.

more. To simplify channel feature relevance investigation, F_1 and Q_1 are, respectively, reshaped into a feature matrix $V_1 \in R^{C\times N}$ and a feature vector $S_W \in R^{N\times 1}$, where $N{=}H \times W$. Afterwards, V_1 is multiplied with S_W via matrix multiplication operations to weighted aggregate the feature responses in each channel of V_1, resulting in a channel informativeness vector $C_I \in R^{C\times 1}$. Particularly, in order to normalize the contributions of the feature responses, the softmax function is operated on S_W before conducting matrix multiplication. Here, each entry of C_I reflects the feature informativeness relating to a channel in the input feature map. Next, to exploit the interdependencies among different channels, two fully-connected layers are further appended to comprehensively interpret the channel-wise informativeness. The last layer outputs a channel attention vector $C_A \in R^{C\times 1}$, where each element encodes the relevance and importance relating to a channel in the input feature map. To well characterize the channel feature relevance on the same baseline, C_A is normalized by the sigmoid function. Eventually, by multiplying each attentive factor in C_A to all the positions in the corresponding channel of the input feature map, we finalize a semantic-enhanced feature representation $F_C \in R^{H\times W\times C}$ that expressly upgrades the contributions of the informative and relevant channel feature semantics.

The spatial-specific attention branch serves to concentrate more on the semantics related to the foreground positions. In this regard, likewise, given the input feature map, two separate 1×1 convolutions are first operated to transform it into feature representations $F_2 \in R^{H\times W\times C}$ and $Q_2 \in R^{H\times W\times C}$. Specifically, Q_2 is further processed via global average pooling (GAP) to compress each channel into a single element, resulting in a feature vector $C_W \in R^{1\times C}$. Similarly, F_2 is also regarded as a feature response map that reflects the task-sensitive feature responses at the positions of the input feature map. Whereas, C_W behaves as a channel weighting map, which measures the strength of the feature responses in each channel. It means that the feature channel associated with a larger weight will cast more contributions. To simplify spatial feature correlation exploitation, F_2 is reshaped into a feature matrix $V_2 \in R^{C\times N}$. Then, V_2 is multiplied by C_W via matrix multiplication operations to weighted aggregate the feature responses at each position of V_2, resulting in a spatial informativeness vector $S_I \in R^{1\times N}$. Specifically, in order to normalize the contributions of the feature responses, C_W is operated by the softmax function before conducting matrix multiplication. Here, an element in S_I mirrors the feature informativeness relating to a position in the input feature map. Next, to mine the correlations among different positions, two fully-connected layers are further connected to intently analyze the spatial informativeness with a bottle-neck pattern. After reshaping the output of the last layer into a feature map $S_A \in R^{H\times W\times 1}$ along the row direction, we obtain a spatial attention map, each of whose elements encodes the relevance and saliency relating to a position in the input feature map. To well characterize the spatial feature relevance on the same baseline, S_A is also normalized by the sigmoid function. Finally, by multiplying each attentive factor in S_A to the corresponding position in each channel of the input feature map, we attain a semantic-enhanced feature representation $F_S \in R^{H\times W\times C}$ that expressly highlights the significance of the salient and task-relevant spatial feature semantics.

As Fig. 2 illustrates, the semantic-recalibrated feature representations F_C and F_S from the channel-specific and spatial-specific attention branches are ultimately concatenated and rationally integrated through a 1×1 convolution, finalizing a desired semantic-upgraded feature representation that coinstantaneously emphasizes the channel-spatial relevant and important feature semantics. Noteworthily, by designing a shallow network architecture without too many complex and computationally intensive matrix multiplication and convolution operations, the proposed feature attention module is quite lightweight and

efficient.

3.3. Segmentation head

As shown in Fig. 1, the HRNet extractor exports a set of multi-resolution feature semantics for supervising the determination of water bodies. Theoretically, the lower-resolution feature semantics behave excellently to smooth the noise influences, thereby beneficial to suppress the interior texture heterogeneities of water bodies. In contrast, the higher-resolution feature semantics behave promisingly to depict target details, thereby favorable to the delineation of water body boundaries. Thus, by organically fusing the multi-resolution feature semantics, we formulate the segmentation head as a hierarchical processing structure, which progressively optimizes the feature semantics in the higher-resolution branches based on the predictions in the lower-resolution branches in a bottom-up manner. Concretely, first, a primary prediction map is inferred using the feature semantics in Branch 4. For each branch, the prediction map involves two category-specific channels with a softmax output form: one for the water body and the other for the background. Then, the prediction map alongside with the feature semantics in the current branch are input to a semantic-level enhancement module (SLEM) for the generation of category-attentive feature semantics. These category-attentive feature semantics are further added to the feature semantics in the upper branch to supervise feature semantic enhancement. Next, the semantic-level-enhanced features are applied to infer a higher-resolution prediction map in the upper branch. As detailed in Fig. 1, the above procedure is hierarchically repeated bottom-up to progressively enhance the higher-resolution feature semantics. Eventually, the semantic-level-enhanced features produced in Branch 1 with the highest quality are leveraged to predict the final water body segmentation map.

The SLEM functions to aggregate the contextual semantics of each individual category from a global perspective to constitute a unified category-oriented semantic encoding. As Fig. 3 depicts, the SLEM combines the feature map F and the prediction map P from the current branch as the input and exports a contextual category-attentive semantic feature map F_E. To be specific, the positional semantics in feature map F can be organized into two category sets based on the prediction map P as follows:

$$S_F = \left\{ F_{[i,j,*]} \middle| \operatorname{argmax}\left(P_{[i,j,*]}\right) = 2 \right\} \tag{1}$$

$$S_B = \left\{ F_{[i,j,*]} \middle| \operatorname{argmax}\left(P_{[i,j,*]}\right) = 1 \right\} \tag{2}$$

where 1 and 2 represent the category entry indices of the binary softmax outputs of the prediction map corresponding to the background and the foreground, respectively; $F_{[i,j,*]}$ and $P_{[i,j,*]}$, respectively, refer to the semantic vectors from F and P at position (i, j); S_F and S_B are the semantic vector sets associated with the foreground and the background, respectively. Likewise, the positional predictions in prediction map P can be also organized into two category sets as follows:

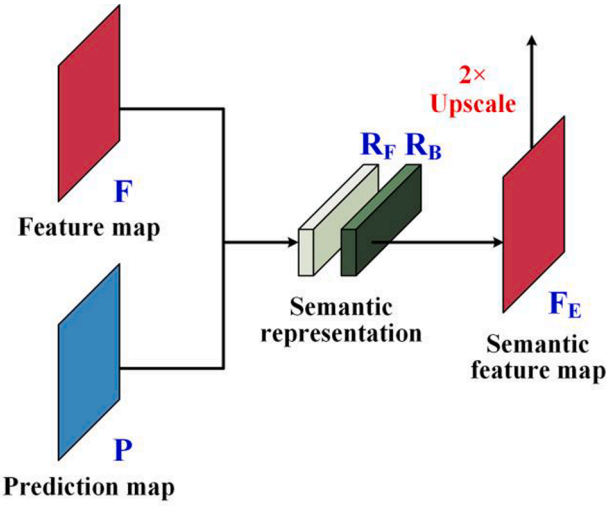


Fig. 3. Overview of the semantic-level enhancement module (SLEM).

$$C_F = \left\{ P_{[i,j,2]} \middle| \operatorname{argmax}\left(P_{[i,j,*]}\right) = 2 \right\} \tag{3}$$

$$C_B = \left\{ P_{[i,j,1]} \middle| \operatorname{argmax}\left(P_{[i,j,*]}\right) = 1 \right\} \tag{4}$$

where C_F and C_B are the category confidence sets associated with the positions being inferred as the foreground and the background, respectively. Next, to constitute a unified category-oriented semantic encoding, we comprehensively and globally aggregate the contextual semantics in each category as follows:

$$R_F = \sum_k \frac{e^{C_{F[k]}}}{\sum_n e^{C_{F[n]}}} S_{F[k]} \tag{5}$$

$$R_B = \sum_k \frac{e^{C_{B[k]}}}{\sum_n e^{C_{B[n]}}} S_{B[k]} \tag{6}$$

where $S_{F[k]}$, $S_{B[k]}$, $C_{F[k]}$, and $C_{B[k]}$ denote the k-th term in the corresponding set; R_F and R_B are, respectively, the constituted category-attentive semantic representations of the foreground and the background (Fig. 3). Finally, these two semantic representations are used to design the semantic feature map F_E as follows:

$$F_{E[i,j,*]} = \begin{cases} R_F, & if \ \operatorname{argmax}\left(P_{[i,j,*]}\right) = 2 \\ R_B, & if \ \operatorname{argmax}\left(P_{[i,j,*]}\right) = 1 \end{cases} \tag{7}$$

Note that, in order for feature semantic enhancement in the upper branch, F_E is further upscaled into the twice size and concatenated to the feature semantics from the upper branch. Afterwards, a 1×1 convolution is operated on the concatenated features to generate a category semantic enhanced feature representation in the upper branch.

3.4. Loss function

As a matter of fact, the quality of the prediction results in the lower branches impacts significantly on the determination of the category-attentive feature semantics used for semantic-level feature enhancement, thereby affecting the feature representation performance and the segmentation accuracy in the upper branches. Thus, to well supervise the optimization of the WaterHRNet towards high-quality water body extraction, each of the four branches is bound by a binary ground-truth segmentation map, in which the positions marked with labels of 1 belong to the water body areas and the positions marked with labels of 0 signify the background areas. The loss function is formulated as the weighted summation of the losses from all the branches as follows:

$$L = \sum_{i=1}^{4} 2^{i-4} \left(L_{FL}^{i} + L_{IoU}^{i} \right) \tag{8}$$

where L_{FL}^{i} is defined by the focal loss (Lin et al., 2017) of the softmax predictions corresponding to the labelled ground truths in the i-th branch; L_{IoU}^{i} is formulated by the intersection over union (IoU) loss (Qin et al., 2019) between the binary segmentation results and the ground-truth segmentation map in the i-th branch. Specifically, considering the size discrepancy between the prediction maps in different branches, a weighting coefficient of 2^{i-4} is introduced to balance the contributions of the branches to the overall loss of the network.

4. Results and discussions

4.1. Datasets

For convincingly verifying the capability of the constructed WaterHRNet in water body extraction issues, confirmatory experiments were intensively conducted on the following three datasets: GE-Water (Yu et al., 2021), CN-Water (Yu et al., 2021), and LandCover.ai (Boguszewski et al., 2021). The GE-Water dataset was specifically built for water body extraction tasks. It consists of 9000 satellite images containing different types and topologies of water bodies collected all around the world using the Google Earth engine. These water bodies vary greatly in patterns, sizes, shapes, appearances, and environmental scenarios. Every image is pixel-wisely annotated by the water body and the background ground truths and formatted into 800 × 800 pixels in size. The CN-Water dataset is also a water body extraction dataset. The images in the CN-Water dataset were captured by the GaoFen-2 satellite in China. It includes a total of 1000 images containing water bodies of mainly rivers, lakes, ponds, and coastal areas. Every image is cropped into 800 × 800 pixels in size and associated with an annotated binary label map for depicting the water body regions and the background. In contrast, the LandCover.ai dataset collected in Poland was initially constructed for land cover classification purposes. In this dataset, the land covers were annotated into four categories including building, woodland, water, and road. It comprises 33 aerial images having the image resolution of about 9000 × 9500 pixels and another 8 aerial images having the image resolution of about 4200 × 4700 pixels. In order to reasonably examine the performance of the WaterHRNet, the land covers in the LandCover.ai dataset were simply categorized into two types: water and non-water. Furthermore, to well facilitate model construction and examination, each image in this dataset was split into image blocks with 800 × 800 pixels based on a 200-pixel overlapping size along both directions. Therefore, the final dataset used for performance evaluation contains 8368 image samples. To be specific, the images in the three datasets are all optical images with red, green, and blue color channels. For each dataset, 60 % of the images were randomly picked out to form the training set, 5 % of the images were selected at random to constitute the validation set, and the other 35 % of the images were treated as the test set.

4.2. Network implementation

At the model optimization stage, the WaterHRNet was trained using the Adam optimizer supervised by the loss function defined in Eq. (8) in a cloud computing environment. This platform is packed with ten 16-GB GPUs, a 128-GB memory, and a 16-core CPU. During training, each training batch is bound with two image samples on each GPU and intently processed for 800 epochs towards model parameter optimization. The learning rate was initialized to be 0.001 and decayed to 0.0001 in the last 400 epochs. Specifically, training sample augmentation was also taken into account aiming at promoting the model quality. To this end, first, a training sample was horizontally flipped to form a duality image. Then, this couple of images were, respectively, clockwise rotated with a step interval of 90 degrees to form another three images for each. As a result, a training sample was eventually augmented into eight variants. The augmented training set was further randomly permutated and finally leveraged for model construction and optimization.

4.3. Water body extraction assessment

To supply quantified and convincing assessments on the water body extraction performance, we employed the following four metrics widely used in semantic segmentations: precision, recall, IoU, and F_1-score. These four metrics appraise the model capability from different aspects. To be specific, precision focuses on the correctly marked water body pixels in the entire predicted water body elements. Recall concerns the successfully recognized water body pixels in the annotated ground truths. IoU and F_1-score assess the overall performance by comprehensively taking into account both the true predictions and the false predictions. For all metrics, the larger the values, the better the model performance.

For providing more convincing evidence to testify the practical feasibility and advanced superiority of the WaterHRNet, a group of contrastive analyses were also carried out with some state-of-the-art deep network models serving for water body extraction issues. The considered models include the multiscale lake water extraction network (MSLWENet) (Wang et al., 2020a), the dense-local-feature-compression network (DLFC-Net) (Li et al., 2021c), the multiscale refinement network (MSR-Net) (Duan and Hu, 2020), the multiscale context extractor network (MSCENet) (Kang et al., 2021), the multifeature extraction and combination network (MECNet) (Zhang et al., 2021a), the improved U-Net (Qin et al., 2022), the deep multifeature water body segmentation network (W-Net) (Tambe et al., 2021), and the self-attention capsule feature pyramid network (SA-CapsFPN) (Yu et al., 2021). Amongst the above eight models, different strategies are employed to reasonably aggregate the multilevel/multiscale feature semantics with the purpose of obtaining strong feature representations for high-quality water body inference. Specifically, the MSLWENet, DLFC-Net, MSCENet, and MECNet are formulated as the encoder-decoder architecture, the improved U-Net and W-Net follow the architecture of the U-Net, and the MSR-Net and SA-CapsFPN are designed with the FPN architecture. In addition, feature attention schemes are intently considered in the MSR-Net, MECNet, and SA-CapsFPN for task-aware feature semantic promotion purpose. Aiming at conducting comparative assessments on the same baseline, all the eight models were optimized and examined on the same datasets and on the same cloud computing platform used in this paper and coupled with the same training data augmentation principle. Quantified evaluations on these models also took place using the same precision, recall, IoU, and F_1-score metrics.

Tables 1, 2, and 3 record the quantified assessment results obtained by the WaterHRNet and the eight compared models on the three test datasets. As reported by the statistics in these tables, the WaterHRNet demonstrated excellent extraction accuracies on all the three datasets. The high precision metric indicated that the water body regions were correctly recognized and well separated from the non-water targets, thereby leading to a small number of false detections. Moreover, the high recall metric indicated that the majority of the water body areas were successfully located and accurately segmented from the background, thereby resulting in a small number of missing detections. The promising performance can be also reflected by the high values of the IoU and the F_1-score metrics. Concretely speaking, on the GE-Water dataset, the WaterHRNet obtained the water body identification accuracy of 97.72 %, 96.94 %, 94.80 %, and 97.33 % on precision, recall, IoU, and F_1-score, respectively. As for the CN-Water dataset, the precision, recall, IoU, and F_1-score evaluation results were, respectively, 98.97 %, 98.86 %, 97.85 %, and 98.91 %. On the LandCover.ai dataset, the WaterHRNet achieved a performance of 98.62 %, 97.73 %, 96.41 %,

Table 1
Water body extraction performances achieved by different models on the GE-Water dataset.

Model	Precision (%)	Recall (%)	IoU (%)	F_1-score (%)	FPS
WaterHRNet	**97.72**	**96.94**	**94.80**	**97.33**	18
MSLWENet	92.84	92.17	86.05	92.50	17
DLFC-Net	92.55	91.94	85.60	92.24	20
MSR-Net	94.86	93.77	89.24	94.31	**25**
MSCENet	92.72	92.13	85.92	92.42	19
MECNet	95.33	94.62	90.43	94.97	20
Improved U-Net	90.86	90.02	82.55	90.44	12
W-Net	91.17	90.35	83.08	90.76	10
SA-CapsFPN	96.65	95.88	92.80	96.26	7

Table 2
Water body extraction performances achieved by different models on the CN-Water dataset.

Model	Precision (%)	Recall (%)	IoU (%)	F_1-score (%)	FPS
WaterHRNet	**98.97**	**98.86**	**97.85**	**98.91**	18
MSLWENet	94.38	93.87	88.90	94.12	17
DLFC-Net	94.11	93.72	88.53	93.91	20
MSR-Net	96.65	95.38	92.33	96.01	**25**
MSCENet	94.26	93.84	88.77	94.05	19
MECNet	97.41	96.53	94.11	96.97	20
Improved U-Net	91.79	90.94	84.10	91.36	12
W-Net	92.04	91.22	84.55	91.63	10
SA-CapsFPN	98.76	97.79	96.60	98.27	7

Table 3
Water body extraction performances achieved by different models on the LandCover.ai dataset.

Model	Precision (%)	Recall (%)	IoU (%)	F_1-score (%)	FPS
WaterHRNet	**98.62**	**97.73**	**96.41**	**98.17**	18
MSLWENet	94.57	93.64	88.86	94.10	17
DLFC-Net	94.23	93.42	88.37	93.82	20
MSR-Net	95.33	94.45	90.27	94.89	**25**
MSCENet	94.75	93.66	89.04	94.20	19
MECNet	96.14	95.30	91.79	95.72	20
Improved U-Net	91.66	91.32	84.31	91.49	12
W-Net	91.79	91.45	84.54	91.62	10
SA-CapsFPN	97.13	96.21	93.55	96.67	7

and 98.17 %, respectively, on the precision, recall, IoU, and F_1-score in extracting water body regions.

As for the compared models, the SA-CapsFPN, MECNet, and MSR-Net demonstrated apparently more advanced water body extraction capabilities than the rest five models, whereas the W-Net and improved U-Net performed less promisingly amongst all the models. To be specific, the performance differences between the best model SA-CapsFPN and the worst model improved U-Net with respect to the average IoU and F_1-score metrics are about 10.67 % and 5.97 %, respectively. For the SA-CapsFPN, capsule primitives are used to constitute a feature pyramidal structure and feature attention and contextual semantic augmentation philosophies are also cooperated for high-quality, target-sensitive feature representation characterization, thereby resulting in the dramatic performance advantage. The performance gains of the MECNet and MSR-Net also benefit from the consideration of multilevel feature fusion and the integration of task-aware feature attention. In contrast, by using only the pure U-Net architecture, the W-Net and improved U-Net demonstrated relatively weaker capabilities, especially when tackling the small-area water bodies and the water bodies showing severe heterogeneities.

Comparatively, most of the models behaved the best on the CN-Water dataset, while they performed relatively less promisingly on the GE-Water dataset. For instance, for the WaterHRNet, the performance differences with regard to the IoU and F_1-score metrics were about 3.05 % and 1.58 %, respectively, between the CN-Water and the GE-Water datasets. For the SA-CapsFPN, the extraction accuracies with respect to the IoU and F_1-score metrics on the CN-Water dataset were improved by about 3.8 % and 2.01 %, respectively, compared with those on the GE-Water dataset. In fact, among the three datasets, the conditions of the water bodies in the GE-Water dataset are more complicated and challenging than the other two datasets, such as severe shadow covers, dim boundaries, land cover overlappings, and appearance heterogeneities, thereby causing more false alarms due to high semantic similarities and missing more true contents due to low semantic certainties. Consequently, the degradations of the precision and the recall metrics inevitably resulted in the performance decline which can be clearly reflected by the values of the IoU and the F_1-score metrics.

However, through comparative analyses, we confirmed that the proposed WaterHRNet showed significant improvement over all the compared models. For instance, the WaterHRNet achieved a performance gain by about 2.03 % and 1.07 % with regard to the IoU and F_1-score metrics in comparison with the SA-CapsFPN. As well, the performance superiority is even higher compared with the improved U-Net (improved by about 12.7 % and 7.04 % for the IoU and F_1-score metrics). Therefore, we concluded that the proposed WaterHRNet fitted excellently and performed superiorly in water body extraction tasks.

Noteworthily, the scenario complexities of the three datasets are also embodied in the following conditions: water color diversities, water body extension variations, topological structure changes of water bodies, surrounding environment variabilities of water bodies, and pattern differences of water bodies (e.g., closed water bodies and open water bodies). All these tough issues brought remarkable ordeals to examine the model performance. Fortunately, the proposed WaterHRNet handled excellently in extracting the water bodies of varying conditions under diverse scenarios. The advanced properties of the WaterHRNet benefitted from the following design philosophies. First, by stacking an HRNet extractor to abstract feature semantics in different subspaces along with the cross-subspace feature integration from a global perspective, the WaterHRNet is capable of providing multi-resolution, high-quality, and strong-semantic feature representations, which perform supportably to identify water bodies of varying conditions. Second, by inlaying a novel feature attention module to emphasize the significance of the informative and useful feature semantics by comprehensively taking into consideration the channel and spatial feature relevance, the WaterHRNet is further boosted to export semantic-determinate and task-aware feature representations, which support beneficially to recognize true water body regions. In addition, by designing a hierarchical segmentation principle to progressively enhance the target saliencies with category-attentive semantics, the WaterHRNet can finalize finer and finer target-oriented feature representations, which behave superiorly to predict more accurate water body segmentation maps. Overall, the proposed WaterHRNet achieved an average performance of 98.44 %, 97.84 %, 96.35 %, and 98.14 % with regard to the precision, recall, IoU, and F_1-score on the three datasets.

To conduct visual verifications on the performance of the proposed WaterHRNet, Fig. 4 shows a set of water body extraction results from these three test datasets. Elaborative observations show that the water bodies of different types, varying patterns, diverse topologies, various appearances, and multifarious self and surrounding scenarios were correctly recognized and nicely segmented with quite small numbers of omissions and commissions. Note that, some water bodies are elongated or of quite small areas, which occupy a very small portion of the image contents. The complete extraction of such water bodies is not easy since they lack of sufficient feature presences or lack of salient feature semantics. Moreover, some water bodies exhibit ambiguous boundaries with the surrounding environments or covered with shadows at the border regions cast by high-rise land covers, which bring challenges to the accurate delineation of the water bodies. In addition, the water bodies differ greatly in water colors due to diverse water body types or caused by the changes in illumination conditions and the use of different imaging sensors. The texture diversities of the water bodies form another tough issue on the requirement of precise identification. Delightfully, benefitting from the HRNet extractor for high-quality feature semantic exploitation, the feature attention strategy for task-relevant feature semantic promotion, and the hierarchical segmentation principle for category-attentive feature semantic enhancement, the proposed WaterHRNet demonstrated advanced performance on the handling of these challenging water body scenarios. However, as observed in Fig. 4, some water bodies are partially shielded by overhead (see the green boxes marked with #1) or on-surface objects (see the green box marked with #2), which damage the integrities of these water bodies. As a result, such water body regions were incorrectly treated as the background and failed to be inferred. Moreover, some water bodies

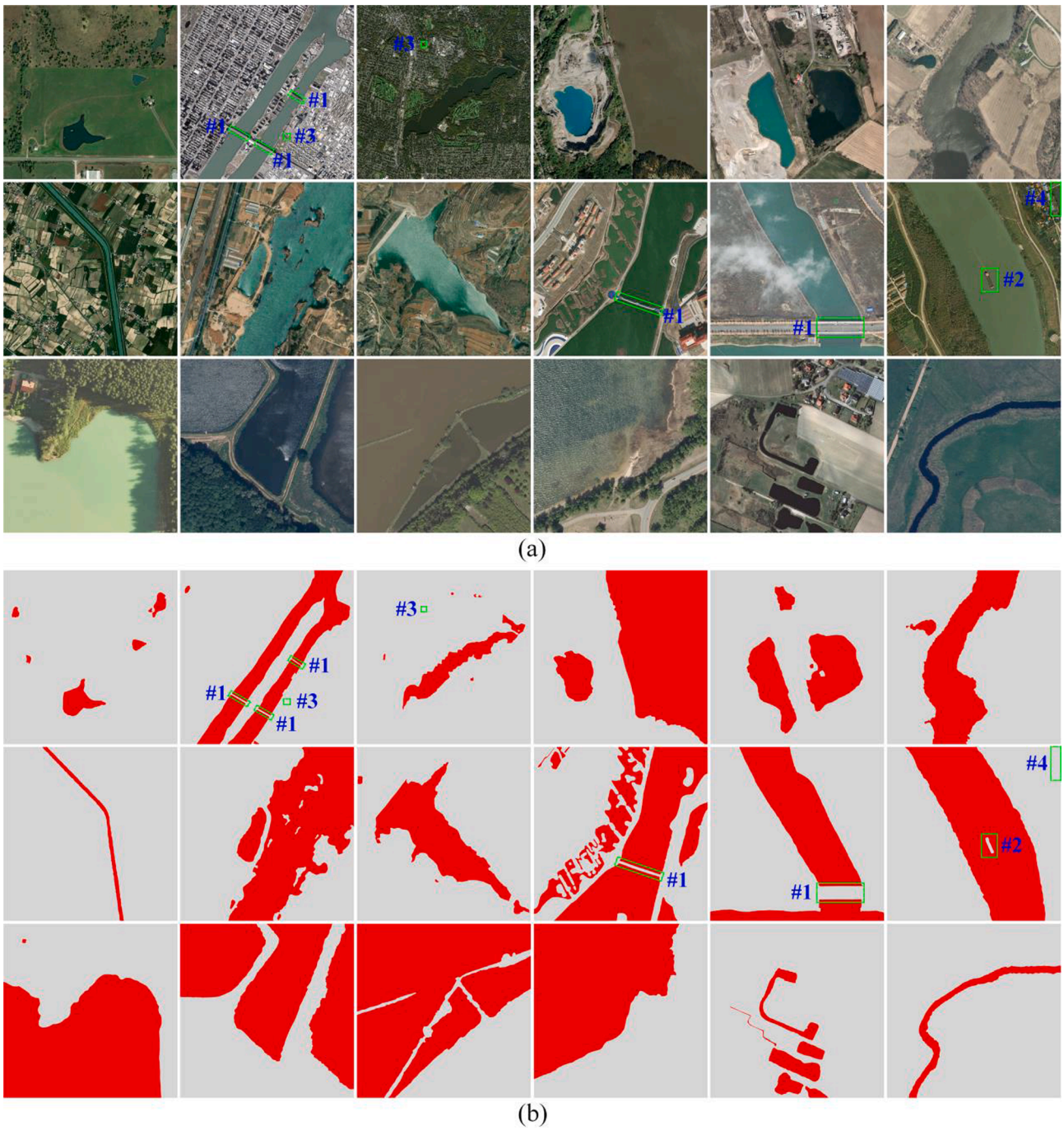


Fig. 4. Examples of water body extraction results from the three test datasets. (a) Test images and (b) extracted water bodies. The regions marked by the green boxes denote the water bodies that are failed to be correctly extracted due to extremely challenging conditions. Specifically, #1 marks the water bodies shielded by overhead objects, #2 marks the water bodies shielded by on-surface objects, #3 marks the water bodies having extremely small sizes, and #4 marks the water bodies exhibiting extremely low contrasts with the surrounding environments.

occupy extremely small areas and show extremely small sizes in the images (see the green boxes marked with #3). The feature semantics of these water bodies were extremely insufficient even in the highest-resolution feature map. Therefore, they were not successfully recognized. In addition, some water bodies exhibit extremely low contrasts with the surrounding environments (see the green box marked with #4). These water bodies had quite similar texture properties to the surrounding scenarios and showed extremely low feature saliencies in the feature maps. As a result, they were also incorrectly identified as the background.

Aiming at further visually comparing the water body extraction performances between the proposed WaterHRNet and the other models, Fig. 5 also presents a set of water body extraction results in some challenging scenarios. As observed by the green boxes in Fig. 5, some small-size, lathy, blur-boundary, low-contrast, or shadow-contaminated water bodies were not correctly recognized and extracted by some of the compared models, resulting in some omission errors. Moreover, some background contents showing similar spectral properties to the water bodies were falsely extracted by some models, leading to some commission errors. In contrast, the proposed WaterHRNet performed competitively with quite low omission and commission errors in handling these challenging water body scenarios.

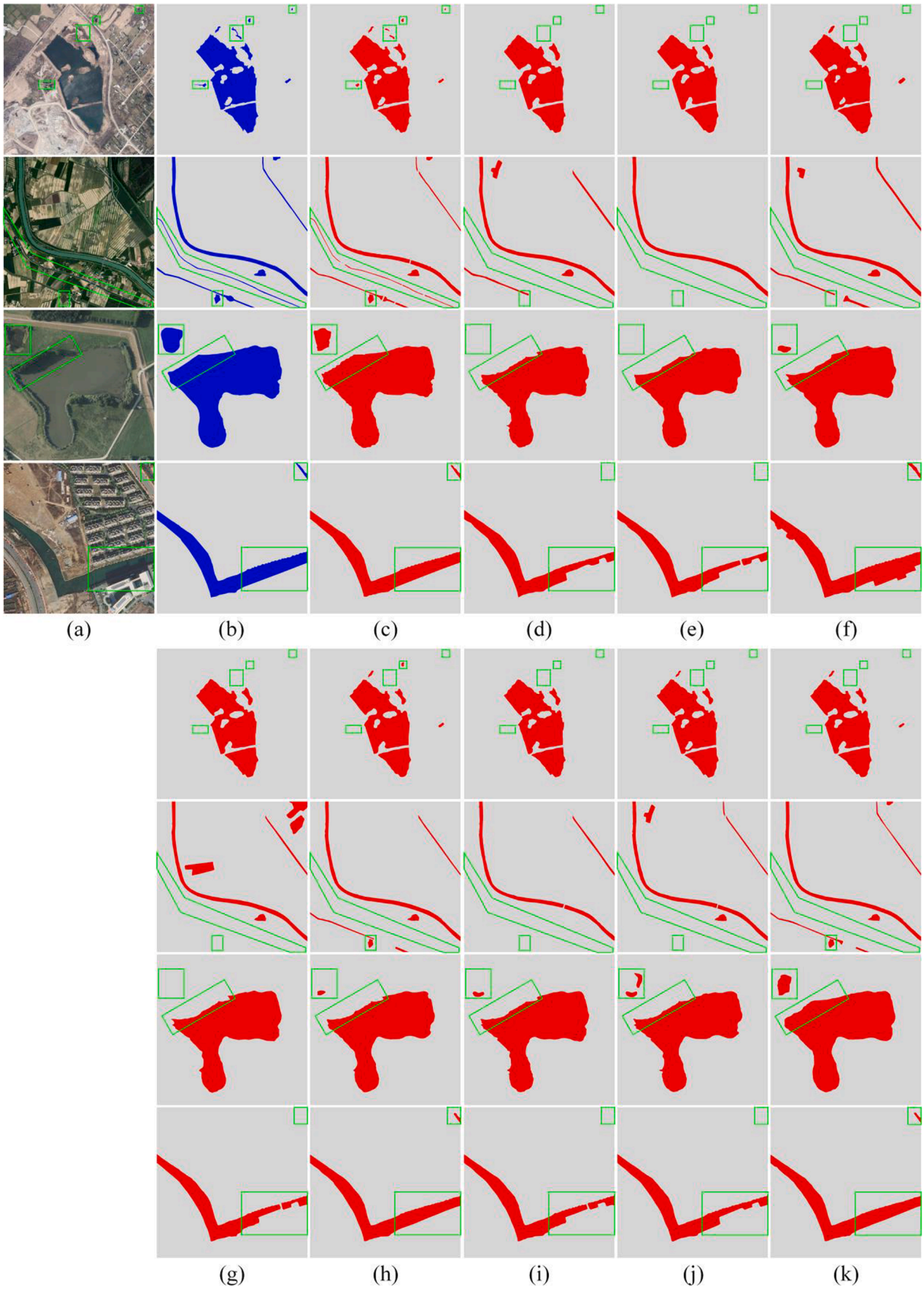


Fig. 5. Visual comparisons of water body extraction results obtained by different models. (a) Test images, (b) ground truths, (c) WaterHRNet, (d) MSLWENet, (e) DLFC-Net, (f) MSR-Net, (g) MSCENet, (h) MECNet, (i) Improved U-Net, (j) W-Net, and (k) SA-CapsFPN. The regions marked by the green boxes show the water body extraction results of different models in some challenging scenarios, including small-size, lathy, blur-boundary, low-contrast, and shadow-contaminated water bodies.

To further evaluate the processing efficiencies of the WaterHRNet and the eight compared models, the frames per second (FPS) indicator was employed for conducting computational performance comparisons. The FPS indicator was defined as the number of image frames being successfully processed each second. At the test stage, the execution time on each dataset was recorded for each of the models to compute the FPS indicator. As reported in Tables 1, 2, and 3, given the input images with 800 × 800 pixels in size on each of the three datasets, the WaterHRNet achieved a processing efficiency of about 18 FPS, which behaved almost equally-matched with the MSCENet model. Comparatively, the MSR-Net achieved the fastest processing efficiency of about 25 FPS due to the lightweight network architecture, whereas the SA-CapsFPN processed less efficiently with an FPS of about 7 caused by the dynamic routing process. Through computational performance evaluations, we concluded that the WaterHRNet behaved promisingly with acceptable processing efficiency in view of the superior water body extraction accuracies.

4.4. Ablation studies

In the proposed WaterHRNet, the dual-path feature attention module and the hierarchical segmentation scheme contributed positively and significantly to the enhancement of the feature encoding quality and the quality of the extracted water bodies. The dual-path feature attention module served to emphasize the saliencies of the task-oriented and semantic-important features in both the channel and spatial domains to enhance the feature representation quality. The hierarchical segmentation scheme functioned to exploit category-attentive feature semantics to augment the feature difference between the foreground and background contents and to employ region localization cues to improve the segmentation accuracy and detail. As ablation studies, we further examined the advanced superiorities of these two components to the performance promotions of the water body extraction task.

First, we evaluated the contribution of the dual-path feature attention module to the performance gain of the WaterHRNet. To this end, we removed all the feature attention modules from the HRNet feature extractor to cancel the feature semantic recalibration mechanism. The modified architecture was termed as WaterHRNet-N. Tables 4, 5, and 6 record the water body extraction performances obtained by the modified architecture on the three test datasets. Likewise, the precision, recall, IoU, and F_1-score metrics were also adopted for quantitative analyses and comparisons. As reflected in these tables, by embedding no feature attention modules for feature semantic enhancement, the WaterHRNet-N performed less promisingly with significant accuracy declines on all the three datasets compared with the WaterHRNet. This is because, without the feature attention module to focus on the informative channel features and the task-relevant spatial features, the representation quality of the extracted feature maps from the HRNet feature extractor were weakened and lowered, thereby resulting in the performance degradation on the cases of challenging conditions, such as low-contrast and spectral-heterogeneous water bodies. On average, the extraction accuracies of the WaterHRNet-N with regard to the IoU and F_1-score metrics were declined by about 6.29 % and 3.37 %, respectively, compared with those of the WaterHRNet. In conclusion, we confirmed that the dual-path feature attention module contributed significantly to the performance gain of the WaterHRNet.

Table 4
Water body extraction performances achieved by different modified architectures on the GE-Water dataset.

Model	Precision (%)	Recall (%)	IoU (%)	F_1-score (%)	FPS
WaterHRNet	**97.72**	**96.94**	**94.80**	**97.33**	18
WaterHRNet-N	94.27	93.31	88.30	93.79	21
WaterHRNet-H	96.55	95.76	92.59	96.15	19
WaterHRNet-P	96.84	96.03	93.11	96.43	18
WaterHRNet-S	92.83	92.12	86.00	92.47	22

Table 5
Water body extraction performances achieved by different modified architectures on the CN-Water dataset.

Model	Precision (%)	Recall (%)	IoU (%)	F_1-score (%)	FPS
WaterHRNet	**98.97**	**98.86**	**97.85**	**98.91**	18
WaterHRNet-N	96.14	95.22	91.71	95.68	21
WaterHRNet-H	98.62	97.66	96.34	98.14	19
WaterHRNet-P	98.79	97.87	96.71	98.33	18
WaterHRNet-S	94.29	93.94	88.88	94.11	22

Table 6
Water body extraction performances achieved by different modified architectures on the LandCover.ai dataset.

Model	Precision (%)	Recall (%)	IoU (%)	F_1-score (%)	FPS
WaterHRNet	**98.62**	**97.73**	**96.41**	**98.17**	18
WaterHRNet-N	95.32	94.36	90.18	94.84	21
WaterHRNet-H	97.21	96.33	93.74	96.77	19
WaterHRNet-P	97.76	96.82	94.72	97.29	18
WaterHRNet-S	94.08	93.41	88.22	93.74	22

Second, we evaluated the contribution of the hierarchical segmentation scheme to the performance gain of the WaterHRNet. To be specific, first, we removed the SLEM along with the hierarchical segmentation structure and directly applied the feature map generated in the highest-resolution branch (Branch 1) of the HRNet feature extractor to conduct water body extraction. The modified architecture was termed as WaterHRNet-H. Second, as a simplified version of the hierarchical segmentation scheme, we removed only the SLEM and directly concatenated the upsampled prediction map in the lower-resolution branch to the feature map in the upper higher-resolution branch to conduct category-attentive feature augmentation. The modified architecture was termed as WaterHRNet-P. The water body extraction performances obtained by these modified architectures on the three test datasets are reported in Tables 4, 5, and 6. As reflected in these tables, without the hierarchical segmentation scheme, the WaterHRNet-H showed significantly lower extraction accuracies than the WaterHRNet on all the three test datasets. The reason is that, by exploiting no category-attentive feature semantics to promote the contrasts between the foreground and background feature representations, the quality of the output feature map used for water body prediction was not further augmented, thereby leading to the performance decline in some challenging scenarios, such as blur-boundary and texture-inconsistent water bodies. On average, the extraction accuracies of the WaterHRNet-H in terms of the IoU and F_1-score metrics were degraded by about 2.13 % and 1.12 %, respectively, compared with those of the WaterHRNet. Note that, the WaterHRNet-P showed better extraction performances than the WaterHRNet-H. It proved that the inclusion of the segmentation cues from lower-resolution branches to supervise the segmentation in the higher-resolution branches was very beneficial to provide valuable localization evidences to upgrade the segmentation accuracy and detail. In conclusion, we confirmed that the hierarchical segmentation scheme contributed positively to the performance gain of the WaterHRNet.

As the last set of ablation experiments, we removed all the feature attention modules from the HRNet feature extractor to cancel the feature semantic recalibration mechanism and removed the hierarchical segmentation structure along with the SLEM from the segmentation head to cancel the category-attentive feature semantic augmentation mechanism. The modified architecture was termed as WaterHRNet-S. As shown by the water body extraction performances in Tables 4, 5, and 6, without the feature attention module and the hierarchical segmentation scheme for feature semantic promotions, the water body extraction accuracies of the WaterHRNet-S declined significantly on all the three test datasets compared with those of the WaterHRNet. To be specific, the

average extraction accuracies of the WaterHRNet-S with respect to the IoU and F_1-score metrics were declined by about 8.65 % and 4.70 %, respectively, compared with those of the WaterHRNet. In conclusion, we confirmed that the combination of the feature attention module and the hierarchical segmentation scheme contributed significantly and positively to the performance gain of the WaterHRNet.

At the test stage, the execution time of each of the modified architectures was also recorded to evaluate their computational performances. Likewise, the FPS indicator was also employed for reflecting the processing efficiency of each of the modified architectures. As shown by the values of the FPS indicator in Tables 4, 5, and 6, most of the modified architectures behaved relatively more efficiently than the WaterHRNet due to the simplifications of some architecture components. Noteworthily, the WaterHRNet-S achieved a higher FPS indicator than the WaterHRNet-N. Similarly, the WaterHRNet-H achieved a higher FPS indicator than the WaterHRNet-P.

5. Conclusion

This paper has designed a high-performance fully convolutional hierarchical architecture, abbreviated as WaterHRNet, for water body extraction issues. The WaterHRNet involved three functional units: a feature extractor, a feature attention module, and a segmentation head. Specially, assembled with an HRNet formulation assisted by cross-subspace feature semantic augmentation, the WaterHRNet can provide high-quality, strong-semantic feature abstractions, which support significantly to handle water bodies under diverse scenarios. Moreover, integrated with a powerful feature attention module by investigating spatial and channel feature importance, the WaterHRNet can attend to the contributions of the semantic-relevant, task-oriented feature encodings, which favor positively to separate water bodies from the complex environments. In addition, designed with a hierarchical segmentation scheme by considering category-attentive feature semantics in different subspaces, the WaterHRNet can finalize a semantic-salient, target-sensitive feature representation, which performs effectively to attain accurate water body segmentations. The proposed WaterHRNet has been elaborately verified and quantitatively analyzed on three datasets. Experimental analyses demonstrated that the WaterHRNet behaved excellently and competitively with an average precision of 98.44 %, average recall of 97.84 %, average IoU of 96.35 %, and average F_1-score of 98.14 %. Comparative tests also proved the practical feasibility and advanced superiority of the WaterHRNet in water body extraction tasks. Since there are still some limitations in the proposed WaterHRNet, which impede the accurate localization and segmentation of the water bodies in different-scenario remote sensing images, in our future works, we will explore new advanced techniques to further improve the water body extraction accuracy. To be specific, first, we will investigate super-resolution techniques to enhance the extraction quality of the extremely small-size and lathy water bodies. Second, we will investigate effective spatial context characterization strategies to promote the extraction integrity of the occluded water body regions. Finally, we will investigate weakly-supervised or few-shot learning architectures to alleviate the requirement of large-amount annotated samples for model training.

CRediT authorship contribution statement

Yongtao Yu: Conceptualization, Funding acquisition, Methodology, Writing – original draft. **Long Huang:** Conceptualization, Methodology, Writing – original draft. **Weibin Lu:** Data curation, Methodology, Writing – original draft. **Haiyan Guan:** Funding acquisition, Formal analysis, Writing – review & editing. **Lingfei Ma:** Software, Visualization. **Shenghua Jin:** Formal analysis, Validation. **Changhui Yu:** Investigation, Validation. **Yongjun Zhang:** Data curation, Methodology. **Peng Tang:** Investigation, Validation. **Zuojun Liu:** Software, Visualization. **Wenhao Wang:** Supervision, Writing – review & editing. **Jonathan Li:** Supervision, Writing – review & editing.

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.

Data availability

The authors do not have permission to share data.

Acknowledgements

This work was supported by the National Natural Science Foundation of China [grant numbers 62076107, 41971414, 51975239]; the Six Talent Peaks Project in Jiangsu Province [grant number XYDXX-098]; and the Natural Science Foundation of Jiangsu Province [grant numbers BK20211365, BK20191214].

References

Abid, N., Shahzad, M., Malik, M.I., Schwanecke, U., Ulges, A., Kovács, G., Shafait, F., 2021. UCL: Unsupervised curriculum learning for water body classification from remote sensing imagery. Int. J. Appl. Earth Observ. Geoinform. 105, 102568.

Boguszewski, A., Batorski, D., Ziemba-Jankowska, N., Dziedzic, T., Zambrzycka, A., 2021. LandCover.ai: Dataset for automatic mapping of buildings, woodlands, water and roads from aerial imagery. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. Workshops, Virtual, pp. 1102-1110.

Chen, H., Liang, Q., Liang, Z., Liu, Y., Ren, T., 2020. Extraction of connected river networks from multi-temporal remote sensing imagery using a path tracking technique. Remote Sens. Environ. 246, 111868.

Chu, H., Kong, S., Chang, C., 2018. Spatio-temporal water quality mapping from satellite images using geographically and temporally weighted regression. Int. J. Appl. Earth Observ. Geoinform. 65, 1–11.

Dang, B., Li, Y., 2021. MSResNet: Multiscale residual network via self-supervised learning for water-body detection in remote sensing imagery. Remote Sens. 13 (16), 3122.

Duan, L., Hu, X., 2020. Multiscale refinement network for water-body segmentation in high-resolution satellite imagery. IEEE Geosci. Remote Sens. Lett. 17 (4), 686–690.

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 (4), 618–622.

He, Y., Yao, S., Yang, W., Yan, H., Zhang, L., Wen, Z., Zhang, Y., Liu, T., 2021. An extraction method for glacial lakes based on landsat-8 imagery using an improved U-Net network. IEEE J Sel. Topics Appl. Earth Observ. Remote Sens. 14, 6544–6558.

Jin, S., Liu, Y., Fagherazzi, S., Mi, H., Qiao, G., Xu, W., Sun, C., Liu, Y., Zhao, B., Fichot, C.G., 2021. River body extraction from sentinel-2A/B MSI images based on an adaptive multi-scale region growth method. Remote Sens. Environ. 255, 112297.

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 Observ. Geoinform. 103, 102499.

Kim, J., Kim, H., Jeon, H., Jeong, S., Song, J., Vadivel, S.K.P., Kim, D., 2021. Synergistic use of geospatial data for water body extraction from sentinel-1 images for operational flood monitoring across southeast Asia using deep neural networks. Remote Sens. 13 (23), 4759.

Li, Y., Dang, B., Zhang, Y., Du, Z., 2022a. Water body classification from high-resolution optical remote sensing imagery: Achievements and perspectives. ISPRS J. Photogramm. Remote Sens. 187, 306–327.

Li, W., Li, Y., Gong, J., Feng, Q., Zhou, J., Sun, J., Shi, C., Hu, W., 2021a. Urban water extraction with UAV high-resolution remote sensing data based on an improved U-Net model. Remote Sens. 13 (16), 3165.

Li, J., Wang, C., Xu, L., Wu, F., Zhang, H., Zhang, B., 2021b. Multitemporal water extraction of Dongting lake and Poyang lake based on an automatic water extraction and dynamic monitoring framework. Remote Sens. 13 (5), 865.

Li, J., Meng, Y., Li, Y., Cui, Q., Yang, X., Tao, C., Wang, Z., Li, L., Zhang, W., 2022b. Accurate water extraction using remote sensing imagery based on normalized difference water index and unsupervised deep learning. J. Hydrol. 612, 128202.

Li, Z., Wang, R., Zhang, W., Hu, F., Meng, L., 2019a. Multiscale features supported DeepLabV3+ optimization scheme for accurate water semantic segmentation. IEEE Access 7, 155787–155804.

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

Li, L., Yan, Z., Shen, Q., Cheng, G., Gao, L., Zhang, B., 2019b. Water body extraction from very high spatial resolution remote sensing data based on fully convolutional networks. Remote Sens. 11 (10), 1162.

Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection. In: Proc. IEEE Int. Conf. Comput. Vis., Venice, Italy, pp. 2999-3007.

Lu, M., Fang, L., Li, M., Zhang, B., Zhang, Y., Ghamisi, P., 2022. NFANet: A novel method for weakly supervised water extraction from high-resolution remote sensing imagery. IEEE Trans. Geosci. Remote Sens. 60, 5617114.
Miao, Z., Fu, K., Sun, H., Sun, X., Yan, M., 2018. Automatic water-body segmentation from high-resolution satellite images via deep networks. IEEE Geosci. Remote Sens. Lett. 15 (4), 602–606.
Nones, M., 2021. Remote sensing and GIS techniques to monitor morphological changes along the middle-lower Vistula river. Poland. Int. J. River Basin Manag. 19 (3), 345–357.
Qin, P., Cai, Y., Wang, X., 2022. Small waterbody extraction with improved U-Net using Zhuhai-1 hyperspectral remote sensing images. IEEE Geosci. Remote Sens. Lett. 19, 5502705.
Qin, X., Zhang, Z., Huang, C., Gao, C., Dehghan, M., Jagersand, M., 2019. BASNet: Boundary-aware salient object detection. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog., Long Beach, USA, pp. 7479-7489.
Rishikeshan, C.A., Ramesh, H., 2018. An automated mathematical morphology driven algorithm for water body extraction from remotely sensed images. ISPRS J. Photogramm. Remote Sens. 146, 11–21.
Sarp, G., Ozcelik, M., 2017. Water body extraction and change detection using time series: A case study of Lake Burdur. Turkey. J. Taibah Univ. Sci. 11 (3), 381–391.
Tambe, R.G., Talbar, S.N., Chavan, S.S., 2021. Deep multi-feature learning architecture for water body segmentation from satellite images. J. Vis. Commun. Image Rep. 77, 103141.
Wang, Z., Gao, X., Zhang, Y., Zhao, G., 2020a. MSLWENet: A novel deep learning network for lake water body extraction of Google remote sensing images. Remote Sens. 12 (24), 4140.
Wang, Z., Gao, X., Zhang, Y., 2021a. HA-Net: A lake water body extraction network based on hybrid-scale attention and transfer learning. Remote Sens. 13 (20), 4121.
Wang, Y., Li, Z., Zeng, C., Xia, G., Shen, H., 2020b. An urban water extraction method combining deep learning and Google earth engine. IEEE J Sel. Topics Appl. Earth Observ. Remote Sens. 13, 769–782.
Wang, Y., Li, S., Lin, Y., Wang, M., 2021b. Lightweight deep neural network method for water body extraction from high-resolution remote sensing images with multisensors. Sens. 21 (21), 7397.
Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., Liu, W., Xiao, B., 2021c. Deep high-resolution representation learning for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. 43 (10), 3349–3364.
Wang, G., Wu, M., Wei, X., Song, H., 2020c. Water identification from high-resolution remote sensing images based on multidimensional densely connected convolutional neural networks. Remote Sens. 12 (5), 795.
Xue, W., Yang, H., Wu, Y., Kong, P., Xu, H., Wu, P., Ma, X., 2021. Water body automated extraction in polarization SAR images with dense-coordinate-feature-concatenate network. IEEE J Sel. Topics Appl. Earth Observ. Remote Sens. 14, 12073–12087.
Yu, Y., Yao, Y., Guan, H., Li, D., Liu, Z., Wang, L., Yu, C., Xiao, S., Wang, W., Chang, L., 2021. A self-attention capsule feature pyramid network for water body extraction from remote sensing imagery. Int. J. Remote Sens. 42 (5), 1801–1822.
Yuan, K., Zhuang, X., Schaefer, G., Feng, J., Guan, L., Fang, H., 2021. Deep-learning-based multispectral satellite image segmentation for water body detection. IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens. 14, 7422–7434.
Zhang, Z., Lu, M., Ji, S., Yu, H., Nie, C., 2021a. Rich CNN features for water-body segmentation from very high resolution aerial and satellite imagery. Remote Sens. 13 (10), 1912.
Zhang, J., Xing, M., Sun, G., Chen, J., Li, M., Hu, Y., Bao, Z., 2021b. Water body detection in high-resolution SAR images with cascaded fully-convolutional network and variable focal loss. IEEE Trans. Geosci. Remote Sens. 59 (1), 316–332.