GLH-Water: A Large-Scale Dataset for Global Surface Water Detection in Large-Size Very-High-Resolution Satellite Imagery

Yansheng Li[1], Bo Dang[1*], Wanchun Li[1], Yongjun Zhang[1]

[1]Wuhan University, Wuhan, China
bodang@whu.edu.cn

Abstract

Global surface water detection in very-high-resolution (VHR) satellite imagery can directly serve major applications such as refined flood mapping and water resource assessment. Although achievements have been made in detecting surface water in small-size satellite images corresponding to local geographic scales, datasets and methods suitable for mapping and analyzing global surface water have yet to be explored. To encourage the development of this task and facilitate the implementation of relevant applications, we propose the GLH-water dataset that consists of 250 satellite images and 40.96 billion pixels labeled surface water annotations that are distributed globally and contain water bodies exhibiting a wide variety of types (e.g., rivers, lakes, and ponds in forests, irrigated fields, bare areas, and urban areas). Each image is of the size 12,800 $\times$ 12,800 pixels at 0.3 meter spatial resolution. To build a benchmark for GLH-water, we perform extensive experiments employing representative surface water detection models, popular semantic segmentation models, and ultra-high resolution segmentation models. Furthermore, we also design a strong baseline with the novel pyramid consistency loss (PCL) to initially explore this challenge, increasing IoU by 2.4% over the next best baseline. Finally, we implement cross-dataset generalization and pilot area application experiments, and the superior performance illustrates the strong generalization and practical application value of GLH-water dataset. Project page: https://jack-bo1220.github.io/project/GLH-water.html

Introduction

As one of the fundamental components of the Earth's natural ecosystem, surface water plays a critical role in maintaining biodiversity, ecological balance, and the development of human societies (Vörösmarty et al. 2010). Due to its wide spatial and temporal distribution, using satellite imagery to detect and map the global surface water is a feasible and convenient method, leading to promising breakthroughs that are applied in the flood mapping (Wieland et al. 2023), surface water changes (Donchyts et al. 2016; Pekel et al. 2016), and other assessments of water resources (Wang et al. 2020b).

Previous researches (Pekel et al. 2016) have shown that approximately 90,000 square kilometers of permanent surface water bodies have disappeared over the past 32 years, and there have been significant changes in the geographical distribution of surface water bodies. It is worth noting that the use of satellite data with a spatial resolution of 30 meters can result in the omission of small water bodies. Therefore, more accurate mapping of surface water bodies at a global or regional level is necessary to further explore the complex distribution and changes of surface water bodies on Earth. Similarly, the mapping of flood disaster also requires the introduction of satellite images with higher spatial resolution to provide detailed spatial details (Wieland et al. 2023). This enables accurate identification of flood-affected areas using advanced water extraction methods and supports effective guidance for human rescue efforts.

Compared to synthetic aperture radar (SAR) imagery, very-high-resolution (VHR) optical satellite imagery (Ground Sampling Distance, GSD$<$5m) has the advantage of providing clearer texture and detail information about wa-

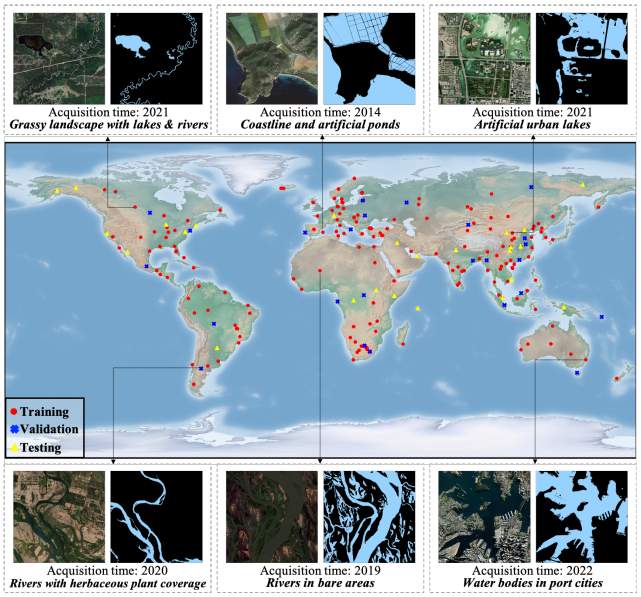

Figure 1: Visualization of the *GLH-water* dataset. We show the geographical coverage of the samples. Several examples from different continents are selected and their image acquisition times and scene descriptions are provided.

*Corresponding author.

ter. In contrast, medium- and low-resolution optical images can only detect large surface water bodies, while small-scale water bodies and their intricate details can only be captured by VHR optical satellite imagery.

To our knowledge, no publicly suitable data has been proposed to facilitate the training and evaluation of surface water detection methods in global VHR optical imagery, which seriously hinders the advancement of global VHR surface water body mapping tasks. In addition, the size of VHR satellite images of the whole scene is typically large. Extracting complete and continuous water bodies from large-size images poses not only a greater challenge that remains unexplored, but also is closer to practical applications such as large-scale surface water mapping.

To promote research on this challenging task and to support applications like higher-resolution global surface water mapping, we propose *GLH-water*, a large-scale dataset for global surface water detection in large-size VHR optical satellite imagery. We collect 250 VHR optical satellite images of 12,800 $\times$ 12,800 pixels containing various water types from the whole world, and create 40.96 billion pixels annotations through manual labeling and expert inspection, as shown in Figure 1. *GLH-water* significantly differs from other existing datasets analyzed in Appendix, with advantages focusing on the following aspects: (1) large-size images; (2) a large number of samples; (3) extensive geographical coverage of the samples; (4) broad temporal span of image acquisition; (5) inclusion of diverse types of surface water. These variations and traits give *GLH-water* its uniqueness and make *GLH-water* well-suited to meet the needs of real-world applications.

Furthermore, to evaluate our dataset and explore this challenging task, we evaluate the performance of representative surface water detection models, well-performing general semantic segmentation models and ultra-high resolution segmentation methods on *GLH-water*, and combine their metrics to create a benchmark. Motivated by multi-layer visual field difference of the image pyramid and the topological continuity of surface water in large-size satellite images, we propose a strong baseline with the new pyramid consistency loss (PCL) to offer a promising pipeline for this challenge. Finally, we conduct extensive experiments to demonstrate the strong generalization of *GLH-water* and its application value. In summary, our contributions are as follows:

- We present the first large-scale dataset for global surface water detection in large-size VHR optical satellite imagery. Through the generalization and pilot area application experiments, it offers significant advantages in mapping global surface water using high-resolution optical satellite images.
- We evaluate a variety of semantic segmentation models on *GLH-water*, which can serve as a benchmark for the development of future methods.
- We further propose a novel strong baseline with the PCL, which yields significant improvements and suggests that it will be a competitive pipeline for future development.

In the future, *GLH-water* and strong baseline we proposed are expected to provide reliable training data and models for manufacturing the global high-resolution surface water map.

Related Work

Relevant Datasets

Surface water detection datasets. The upper portion of Table 1 displays existing datasets specifically designed for detecting surface water bodies. The majority of currently available datasets (Hu et al. 2022; Isikdogan et al. 2019; Luo, Tong, and Hu 2021; Seale et al. 2022) for surface water detection based on optical satellite imagery exhibit only low to medium spatial resolution, as seen with Landsat-8 and Sentinel-2 images. The resolution limitations of the images result in the blurring and indistinguishability of small rivers and lakes. To track intricate surface water systems, research endeavors direct their experimental focus towards commercial satellites that offer high resolution of up to 1m or even 0.3m, such as GeoEye and WorldView (Moortgat et al. 2022; Wieland et al. 2023). Regrettably, owing to the policy constraints of commercial satellites, these datasets are unavailable to the public community. In contrast, our *GLH-water* is the first publicly available large-scale dataset for surface water detection from global VHR optical satellite imagery.

Land use and land cover (LULC) datasets. As a fundamental application in the field of remote sensing, LULC datasets are extensively created and utilized, and they commonly encompass the class of water. However, they inadequately fulfill the need of tasks such as refined global water mapping. The reason is that none of them can simultaneously satisfy the trinity of global sampling, VHR, and large-size imagery, as shown in the lower portion of Table 1. For example, FBP (Tong, Xia, and Zhu 2023) originates from the VHR and large-size images of GaoFen-2 satellite. However, it is confined to a limited selection of cities in China, and migration of the trained model to other regions presents a challenge. The image size of DynamicEarthNet (Toker et al. 2022) is only 1,024×1,024 pixels, insufficient to effectively portray the distribution of water bodies, a characteristic with notable geospatial continuity. In contrast, *GLH-water* offers advantages such as global sampling, VHR, large-size imagery. These attributes are essential for executing surface water mapping tasks on a global scale.

Relevant Methods

Surface water detection methods based on non-deep learning algorithms. Normalized Difference Water Index (NDWI) (McFeeters 1996), Modified Normalized Difference Water Index (MNDWI) (Xu 2006), and other threshold-based water indices (Yao et al. 2015; Wu et al. 2018) are commonly proposed and implemented in initial studies. However, their reliance on spectral information result in a lack of consideration for the spatial information present within the images. Additionally, shallow classifiers such as Support Vector Machine (SVM) are employed and show significant improvements (Wu et al. 2018).

Surface water detection methods based on deep learning algorithms. The VHR optical satellite imagery comprises of a range of water bodies, including but not limited to, rivers, lakes, and ponds with diverse sizes and shapes (Li

Dataset	# Images	# Channels	Image size (pixels)	GSD (m)	# Labeled pixels (billion)	Sources	Geographic coverage
Dedicated surface water detection dataset							
DeepWaterMap	>140000	6	-	30	-	Landsat-8	Globe
ESKWB	95	6	545∼1432×625∼1527	10	0.11	Sentinel-2	Globe
SWED	1862	12	256×256	10	0.12	Sentinel-2	Globe
SWB	2841	3	57∼5292×57∼6767	10	-	Sentinel-2	-
C2S-MS Floods	1800	2/13	512×512	10	0.23	Sentinel-1, 2	Globe
2020 GF challenge	1000	3	492∼2000×492∼2000	1 to 4	0.24	GaoFen-2	-
MSRWD	2419	3	512×512	-	0.63	ZY-3, et al.	-
LULC dataset (including water bodies)							
DeepGlobe	803	3	2448×2448	0.5	4.81	DigitalGlobe	-
Agriculture-Vision	94986	4	512×512	0.1/0.15/0.2	24.9	UAV camera	United States
LoveDA	5987	3	1024×1024	0.3	6.27	Google Earth	China
FBP	150	4	7200×6800	4	7.34	GaoFen-2	China
DynamicEarthNet	54750	4	1024×1024	3	1.38	PlanetFusion	Globe
OpenEarthMap	5000	3	1024×1024	0.25-0.5	5.24	-	Globe
GLH-water (our)	250	3	**12800×12800**	**0.3**	**40.96**	Google Earth	**Globe**

Table 1: Comparison among *GLH-water* and other relevant datasets. All datasets are compared on number of images and channels, image size, spatial resolution (GSD), data sources, and geographic coverage.

et al. 2022a; Kang et al. 2023). Numerous studies (Duan and Hu 2019; Yu et al. 2021; Kang et al. 2021) aim to enhance the identification of intricate water bodies by optimizing the deep learning network, thereby enabling more efficient utilization of the multiscale characteristic. The meandering of water body boundaries constitutes a critical hindrance to the precise segmentation of water bodies. (Miao et al. 2018) devise a loss function to derive accurate water body boundaries, considering the distribution of boundary weights.

Application: Global Surface Water Mapping

In the broader context of global water resource monitoring and mapping, the Global Flood Database (Kettner et al. 2021) contains maps at 250m resolution of the extent distribution of 913 flood events that occurred between 2000 and 2018. The European Commission Joint Research Centre (ECJRC) use Landsat-5/7/8 satellite imagery to map global 30m water body data products for the period 1984-2020 (Pekel et al. 2016). Global land cover products which incorporate the water body category are gradually emerging, albeit at a resolution of merely 10m (Jun, Ban, and Li 2014; Karra et al. 2021). In addition, continuous production and application of regional mapping products with low to medium resolution imagery for water bodies (Wang et al. 2022; Li and Niu 2022; Feng et al. 2016) and with high resolution imagery for LULC (Robinson et al. 2019; Li et al. 2022b) is ongoing. The inevitable trend in cartography is to continuously enhance the spatial resolution of products, as this demonstrates the benefits of more detailed information. However, the production of global VHR water cover maps remains a challenging task, due to the difficulty of acquiring and organizing VHR satellite data, the absence of publicly available large-scale surface water detection datasets with manual annotation, and the lack of related models that are suitable for large-size images. Our *GLH-water* and strong baseline are expected to fill these gap.

The GLH-Water Dataset

To fill the lack of pertinent datasets and enhance the generalizability of models in detecting global surface water, we first present the *GLH-water* dataset that contains 250 VHR satellite images with the size of 12,800 × 12,800 pixels. These images are collected from various locations worldwide and manual annotations are included, as illustrated in Figure 1. In the remainder of this section, we provide details on the imagery, annotations, and advantages of our dataset.

Images Collection and Preprocessing

Using the Google Earth platform, we collect a total of 250 satellite images across the globe, each with sizes of 12,800 × 12,800 pixels at about 0.3m (19 level) spatial resolution, and encompassing approximately 3,686 km^2 in geographic coverage. To guarantee the diversity of the dataset, we handpick the geographic coordinates and acquisition time of the sample data to ensure an accurate representation of the various attributes of the global water bodies.

Annotation Method and Inspection

The annotation labeling process includes three distinct stages: fine labeling, fine checking and correction, and random checking conducted by experts. After scrutiny and revision, no apparent errors are found in the *GLH-water* dataset. Some annotated samples are shown in Figure 2. More details about annotation and inspection are provided in appendix.

Advantage Analysis

To the best of our knowledge, the *GLH-water* dataset is the first publicly available and largest dedicated dataset for global-scale surface water detection from large-size VHR satellite imagery. A comparison with other existing datasets is shown in Table 1.

Specifically, our *GLH-water* dataset has five remarkable and important advantages:

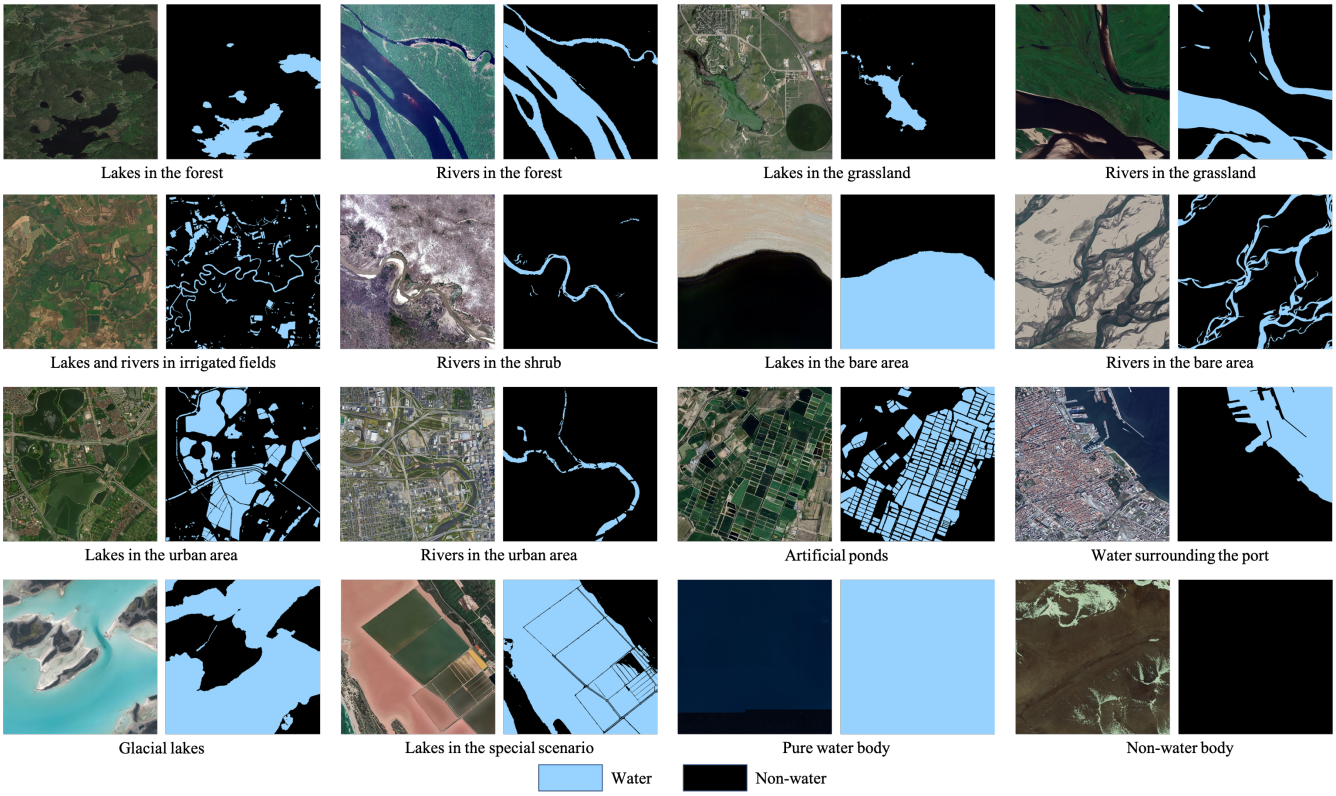

Figure 2: Visualization of various types of surface water bodies in different scenarios on *GLH-water* dataset. It can comprehensively reflect the diversity of the global surface water system.

- **The size of samples is large.** The size of each image is up to 12,800 × 12,800 pixels, which is more in line with the size of a whole scene image acquired by satellite. It is also a challenge for existing methods.
- **Inclusion of a large number of samples.** After non-overlapping cropping, a total of 156,250 tiles of 512 × 512 pixels in size and 40.96 billion labeled pixels are included, which is the largest dataset for global surface water detection from large-size VHR satellite imagery.
- **The geographical coverage of the samples is extensive.** Figure 1 illustrates the detailed geographic distribution, showing that data points are present on all continents except Antarctica. The geographic distribution is reasonable, and can represent the features of surface water body worldwide. Therefore, models trained on our dataset are expected to have stronger generalizability in the geographical dimension.
- **The temporal span of image acquisition is broad.** The range of acquisition time of data spans from 2011 to 2022, and each year contains a certain amount of data. Models exhibit greater temporal generalization ability.
- **Inclusion of a diverse type of surface water landscapes**, as depicted in Figure 2. These include, but are not limited to, *lakes and rivers in the forest, grassland, field, shrub, bare area, and urban area*, *pools, glacial lakes, and water in the special scenario*. The wide types serves as a representation of various geographic landscapes, land cover conditions, water body shapes, and color tone types, thus providing a comprehensive reflection of the diversity of the global surface water system.

In summary, the above five advantages drive *GLH-water* dataset to be unique and advanced.

Dataset Splits

To ensure that the training and test data are roughly equally distributed, we randomly select 80% of the original images as the training set, 10% as the validation set, and 10% as the test set. From the geographical distribution range shown in Figure 1, we can find that the validation and test sets are randomly distributed in various regions, which can well reflect the actual performance of the model trained by this dataset.

Method

Baseline Models

Many models are developed to consider the characteristics of water bodies in VHR satellite images in the field of remote sensing, as outlined in related work. We choose three representative models (i.e., MECNet (Zhang et al. 2021), MSResNet (Dang and Li 2021), and MSCENet (Kang et al. 2021)) as baseline models. In the realm of computer vision, numerous sophisticated semantic segmentation models are

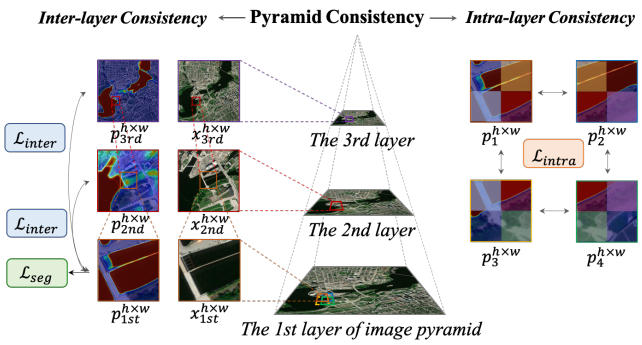

Figure 3: An overview of our strong baseline with the PCL.

perpetually created, which can be adapted for satellite images. We use five advanced models to construct the benchmark results, namely FCN (Long, Shelhamer, and Darrell 2015), PSPNet (Zhao et al. 2017), DeepLab v3+ (Chen et al. 2018), HRNet (Wang et al. 2020a), and STDC (Fan et al. 2021). In addition, given the VHR and the large size of images in *GLH-water*, we also evaluate several ultra-high resolution segmentation methods (i.e., FCtL (Li et al. 2021), MagNet (Huynh et al. 2021), and ISDNet (Guo et al. 2022)).

A Strong Baseline With the PCL

We develop a competitive strong baseline with the new PCL that is specifically designed to explore the detection performance of surface water in large-size VHR satellite images. The pyramid consistency encompasses the inter-layer consistency (i.e., visual field consistency between pyramidal layers) and the intra-layer consistency (i.e., spatial consistency within a pyramidal layer), as illustrated in Figure 3. To construct the image pyramid, we downsample each original large-size image $X^{H\times W}$ at varying rates $\{\sigma_i, i = n\}$, resulting in a multi-layer representation. Considering computational cost, we adopt downsampling rates of 1, 1/5, and 1/25, generating an image pyramid comprising three layers.

Inter-layer consistency. As mentioned in (Min et al. 2022), the human brain may be influenced by the varying sizes of the visual field being observed, potentially resulting in divergent interpretations. The differences in attention maps of tiles with distinct visual fields displayed in Figure 4(a) show that the model is influenced by context information associated with the visual field. Motivated by this idea, we propose the inter-layer consistency loss to calculate the discriminative variances of the model resulting from dissimilarities in the visual range of patches. Specifically, we define the small tiles in the original image (i.e., tiles located in the first layer of the pyramid) $x_{1st}^{h\times w}$ as the fundamental units and establish inter-layer tile groups $\{x_{1st}, x_{2nd}, x_{3rd}\}^{h\times w}$ by upwardly mapping the corresponding tiles from various layers. It is pertinent to note that while the size of each tile within the tile group remains consistent and same, the visual field they contained is gradually increasing as shown in Figure 3. The tile groups are trained by the encoder and decoder to obtain the corresponding $sigmoid$ normalized confidence maps

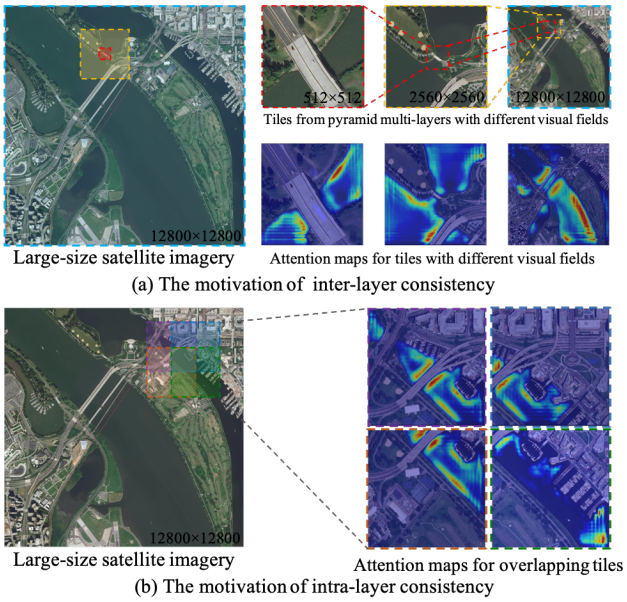

Figure 4: The motivation of pyramid consistency. Attention maps of tiles with different visual fields and overlapping tiles are distinct in the same areas. Attention maps are obtained by the GradCAM (Selvaraju et al. 2017), using the ResNet-50 model trained on the *GLH-water* dataset.

$\{p_{1st}, p_{2nd}, p_{3rd}\}^{h\times w}$. Minimizing the differences between same regions in the confidence maps, which are caused by differences in the visual field, alleviates the visual field bias that arises due to limited contextual information.

Intra-layer consistency. Slicing the original image for processing may lead to the loss of contextual information and interdependence between adjacent tiles, which potentially disrupts the topological continuity of water bodies in remote sensing images with a large size. Figure 4(b) implies that the models exhibit differentiated attention for adjacent tiles with overlaps. We argue that enforcing consistency of the overlapping region on tiles with different contextual information helps to resume the continuity of water bodies. Based on this challenge and motivation, we further develop the intra-layer consistency loss to effectively model the continuous relationship between neighboring tiles and compensate for the information gap induced by image slicing. Specifically, we define four adjacent and overlapping tiles as an intra-layer tile group $\{x_1, x_2, x_3, x_4\}^{h\times w}$, and their spatial relationships are depicted in Figure 3. All tile pairs $\{(x_i, x_j), 1 \leqslant i < j \leqslant 4\}$ are processed by the encoder and decoder to obtain $sigmoid$ normalized confidence map pairs (p_i, p_j). The overlapping part between them is used to calculate the intra-layer consistency loss.

Loss function. Inspired by the focal loss (Lin et al. 2017), we modify and present a novel consistency loss function to effectively calculate both inter-layer consistency $\mathcal{L}_{inter}$ and intra-layer consistency $\mathcal{L}_{intra}$ abovementioned. Overall optimization objective function can be defined as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{seg} + \alpha_{inter}\mathcal{L}_{inter} + \alpha_{intra}\mathcal{L}_{intra}, \quad (1)$$

where $\mathcal{L}_{seg}$ denotes the regular semantic segmentation loss

using binary cross entropy loss ℓ_{bce}. $\alpha_{inter}, \alpha_{intra}$ are trade-off weights. $\mathcal{L}_{inter}, \mathcal{L}_{intra}$ are formulated as

$$\mathcal{L}_{inter} = \frac{1}{w \times h} \sum^{w \times h} (1 - p_{1st})^r (1 - \lambda) y_{1st} \ell_2 (p_{1st}, \tilde{p}_{2nd}) + \lambda p_{1st}^r (1 - y_{1st}) \ell_2 (p_{1st}, \tilde{p}_{2nd}) + \frac{1}{w \times h} \sum^{w \times h} (1 - p_{1st})^r (1 - \lambda) y_{1st} \ell_2 (p_{1st}, \tilde{p}_{3rd}) + \lambda p_{1st}^r (1 - y_{1st}) \ell_2 (p_{1st}, \tilde{p}_{3rd}), \quad (2)$$

$$\mathcal{L}_{intra} = \frac{1}{\tilde{w} \times \tilde{h}} \sum^{\tilde{w} \times \tilde{h}} \sum_{1 \leqslant i < j \leqslant 4} (1 - \tilde{p}_i)^r (1 - \lambda) \tilde{y}_i \ell_2 (\tilde{p}_i, \tilde{p}_j) + \lambda \tilde{p}_i^r (1 - \tilde{y}_i) \ell_2 (\tilde{p}_i, \tilde{p}_j), \quad (3)$$

where y_{1st} denotes corresponding binary annotations of tiles in first layer. The value of 1 denotes water type in the pixel, while the value of 0 indicates non-water type. p_{1st} denotes the confidence map of tiles in first layer, and $\tilde{p}_{2nd}, \tilde{p}_{3rd}$ represent the confidence maps of the overlapping regions in the inter-layer tile group of other layers (after upsampling). $\ell_2 (p_{1st}, \tilde{p}_{2nd}) = \|p_{1st} - \tilde{p}_{2nd}\|_2^2$ calculates the square of the euclidean distance. Similarly, $\tilde{p}_i, \tilde{p}_j$ represent the confidence map of the overlapping regions in the intra-layer tile group. $\tilde{w}$ and $\tilde{h}$ represent the width and height of the overlapping region. r, λ are tunable parameters, which help the model to focus on learning hard-to-distinguish samples.

Benchmark and Experiment

Setup

Implementation details. To ensure the fairness of the evaluation, the implemention setting of all methods is as similar as possible. More details are in the appendix.

Evaluation metrics. We use the intersection-over-union (IoU) metric to evaluate the quantitative performance of detecing surface water, following previous related work. In addition, we also use Frames Per Second (FPS) to evaluate the computational efficiency of different models.

Evaluation Results

As described in baseline models, we evaluate 12 of the popular methods shown in Table 2. The accuracy of generic semantic segmentation models is overall higher than that of models designed for surface water detection, which proves that surface water detection in large-size VHR satellite imagery is challenging and still needs further development. Furthermore, our PCL outperforms other methods by leveraging the multi-layer field of visual information present in large-size images and the topological continuity of water bodies. Nonetheless, it suffers from low efficiency and high computational cost, which are common issues faced by other ultra-high resolution segmentation methods (i.e., FCtL (Li et al. 2021) and ISDNet (Guo et al. 2022)). Thus, striking a balance between accuracy and efficiency should be considered a crucial research priority in this task. The qualitative visualization results and additional metric results are provided in the appendix.

Method	Backbone	IoU(%) (↑)	FPS (↑)
Segmentation methods proposed for surface water detection			
MECNet	-	44.67	3.44
MSResNet	Res-34	69.76	4.03
MSCENet	Res2-50	74.81	2.60
Generic segmentation methods in Computer Vision			
FCN8s	VGG-16	73.66	6.70
PSPNet	Res-50	75.19	5.98
DeepLab v3+	Res-50	79.80	4.48
HRNet-48	-	78.60	3.03
STDC-1446	-	75.82	**26.50**
Ultra-high Resolution Segmentation methods			
MagNet	FPN-Res-50	62.77	13.33
FCtL	FCN8s-VGG16	74.92	0.112
ISDNet	DeepLab v3-Res-18	53.04	2.09
Our PCL	PSPNet -Res-50	**82.26**	1.34

Table 2: Benchmark results on *GLH-water* `test` set. FPS is measured in training settings with batchsize=2. F1-score and GPU memory metrics are released in appendix.

Table 3 demonstrates the consistent improvement of our strong baseline compared to the fair baseline approach across different segmentation model settings. This indicates that our approach is an effective pipeline and is driving progress in this task.

Method	Seg model	IoU (%)
Baseline	FCN8s-VGG16	73.66
FCtL		74.92
Our PCL		**75.78** (+2.12)
Baseline	PSPNet-Res50	75.19
Our PCL		**82.26** (+7.07)
Baseline	DeepLabV3+-Res50	79.80
Our PCL		**81.33** (+1.53)

Table 3: Performance of baseline with different networks or the existing method with the fair network and our proposed model on *GLH-water* `test` set. The results show that our model outperforms common models when using various segmentation models.

Ablation Study on the Strong Baseline

Effectiveness of components in PCL. Exps. II and III in Table 4 show that both key components of PCL (i.e., $\mathcal{L}_{inter}$ and $\mathcal{L}_{intra}$) outperform the baseline by a large margin (+5.89% and +5.50%), and their combination can further improve the performance of the model (Exp. VII).

Effectiveness of loss function of PCL. We conduct an ablation study using the vanilla L2 loss function to measure the pyramid consistency (Exp. IV), and find that the loss function we designed (Eqs. (2) and (3)) has superior capacity

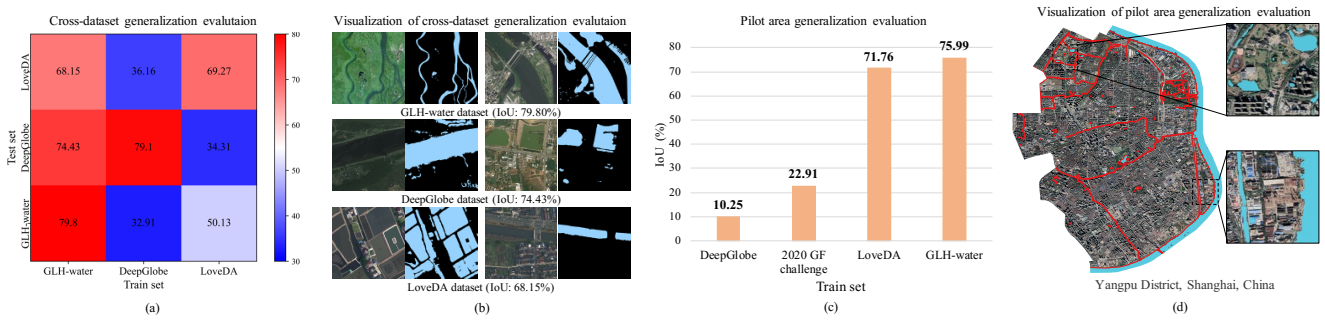

Figure 5: Generalization and application experiment results on *GLH-water* dataset. (a) IoU (%) of cross-dataset evaluation. (b) Visualization results of cross-dataset evaluation. The displayed results are all predicted by the model trained on *GLH-water*. (c) IoU (%) of pilot area evalution with the HRNet-48 models trained on different datasets. (d) Visualization results of the model trained by our *GLH-water* on the pilot area. Red lines represent ground truth, and cyan masks are predictions.

ID	Configuration	IoU (%)	Δ (%)
I	Baseline ($\mathcal{L}_{seg}$)	75.19	-
II	$\mathcal{L}_{seg} + \mathcal{L}_{inter}$	81.08	+5.89
III	$\mathcal{L}_{seg} + \mathcal{L}_{intra}$	80.69	+5.50
IV	Vanilla ℓ_2	71.46	−3.73
V	w/o the 2rd layer of image pyramid	80.13	+4.94
VI	w/o the 3nd layer of image pyramid	81.37	+6.18
VII	**Our PCL**	**82.26**	+7.07

Table 4: Ablation study on key components of strong baseline (seg model: PSPNet-Res-50). Vanilla ℓ_2 means using vanilla L2 loss to measure the pyramid loss rather than loss function (Eqs.(2) and (3)) we designed.

to facilitate learning on hard-to-distinguish samples, thereby resulting in performance improvement.

Impact of the number of image pyramid layers. We observe that if we only apply two layers of the image pyramid to participate in training, our PCL will bring a performance improvement of over 4.9% (Exps. V and VI). When building three layers of the image pyramid, we can see that there is a 7.07% improvement (Exp. VII), indicating that more layers being considered may be more beneficial.

Cross-Dataset Generalization Evaluation

Considering the similar resolution, data source, and large image size, we choose the *LoveDA* (Wang et al. 2021) and *DeepGlobe* dataset (Demir et al. 2018) to implement the cross-dataset generalization evaluation. Following the data split of (Wang et al. 2021; Guo et al. 2022), we use DeepLab v3+-Res-50 as the segmentation model to train the models and evaluate the cross-dataset performance.

As shown in Figure 5(a), there is little difference between the results of the model trained on *GLH-water* and the model trained on *LoveDA* on the *LoveDA* test set (68.15% vs. 69.27%). However, the performance of the model trained on *LoveDA* is significantly diminished when transferred directly to the *GLH-water* test set (50.13% vs. 79.80%). A similar situation occurrs between *DeepGlobe* and *GLH-water*. Results in Figure 5(a) and (b) confirm the strong generalization of our *GLH-water*.

Pilot Area Application Evaluation

Providing data support for VHR global surface water mapping is one of the motivations for constructing the *GLH-water*. We select the Yangpu District of Shanghai, China, which is independent of the dataset and annotated by experts, as a pilot area (60.61 km^2) to further discuss the surface water mapping application of *GLH-water*.

Based on the results presented in Figure 5(c) and (d), it is evident that the model trained on *GLH-water* exhibits superior performance (75.99%) in the surface water mapping task in the pilot area, surpassing the models trained on other datasets. These findings suggest that *GLH-water* holds significant potential for global-scale VHR surface water mapping, owing to its strong generalization.

Conclusions and Future Work

We present a global large-scale dataset for surface water detection in large-size VHR satellite imagery, which is the first publicly available dataset for this task. Unlike existing datasets, we collect 250 large-size satellite images containing various surface water scenes across the whole earth and carefully annotate 40.96 billion pixel labels. Considering the advantages of *GLH-water* over other datasets, cross-dataset generalization, and pilot area application evaluation results, we believe that this dataset is more suitable for practical applications. Additionally, we build a benchmark to evaluate advanced segmentation models in the fields of remote sensing and computer vision. We also propose a strong baseline with PCL, which is a promising research pipeline to advance this task and related applications.

In future research, *GLH-water* is expected to serve not only as an evaluation tool for algorithm advancements but also as a supportive resource for global high-resolution surface water mapping and related environmental sustainability topics, such as global water resource conservation and management.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 42371321; in part by the State Key Program of the National Natural Science Foundation of China under Grant 42030102, in part by the Wuhan University-Huawei Geoinformatics Innovation Laboratory. The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

References

Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 801–818.

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

Demir, I.; Koperski, K.; Lindenbaum, D.; Pang, G.; Huang, J.; Basu, S.; Hughes, F.; Tuia, D.; and Raskar, R. 2018. Deepglobe 2018: A challenge to parse the earth through satellite images. In *CVPR Workshops*, 172–181.

Donchyts, G.; Baart, F.; Winsemius, H.; Gorelick, N.; Kwadijk, J.; and Van De Giesen, N. 2016. Earth's surface water change over the past 30 years. *Nature Climate Change*, 6(9): 810–813.

Duan, L.; and Hu, X. 2019. Multiscale refinement network for water-body segmentation in high-resolution satellite imagery. *IEEE Geoscience and Remote Sensing Letters*, 17(4): 686–690.

Fan, M.; Lai, S.; Huang, J.; Wei, X.; Chai, Z.; Luo, J.; and Wei, X. 2021. Rethinking bisenet for real-time semantic segmentation. In *CVPR*, 9716–9725.

Feng, M.; Sexton, J. O.; Channan, S.; and Townshend, J. R. 2016. A global, high-resolution (30-m) inland water body dataset for 2000: First results of a topographic–spectral classification algorithm. *International Journal of Digital Earth*, 9(2): 113–133.

Guo, S.; Liu, L.; Gan, Z.; Wang, Y.; Zhang, W.; Wang, C.; Jiang, G.; Zhang, W.; Yi, R.; Ma, L.; and Xu, K. 2022. ISDNet: Integrating Shallow and Deep Networks for Efficient Ultra-High Resolution Segmentation. In *CVPR*, 4361–4370.

Hu, K.; Li, M.; Xia, M.; and Lin, H. 2022. Multi-scale feature aggregation network for water area segmentation. *Remote Sensing*, 14(1): 206.

Huynh, C.; Tran, A. T.; Luu, K.; and Hoai, M. 2021. Progressive Semantic Segmentation. In *CVPR*, 16755–16764.

Isikdogan, L. F.; Bovik, A.; Passalacqua, P.; and Passalacqua, P. 2019. Seeing through the clouds with deepwatermap. *IEEE Geoscience and Remote Sensing Letters*, 17(10): 1662–1666.

Jun, C.; Ban, Y.; and Li, S. 2014. Open access to Earth land-cover map. *Nature*, 514(7523): 434–434.

Kang, J.; Guan, H.; Ma, L.; Wang, L.; Xu, Z.; and Li, J. 2023. WaterFormer: A coupled transformer and CNN network for waterbody detection in optical remotely-sensed imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 206: 222–241.

Kang, J.; Guan, H.; Peng, D.; and Chen, Z. 2021. Multi-scale context extractor network for water-body extraction from high-resolution optical remotely sensed images. *International Journal of Applied Earth Observation and Geoinformation*, 103: 102499.

Karra, K.; Kontgis, C.; Statman-Weil, Z.; Mazzariello, J. C.; Mathis, M.; and Brumby, S. P. 2021. Global land use/land cover with Sentinel 2 and deep learning. In *IGARSS*, 4704–4707. IEEE.

Kettner, A.; Tellman, B.; Kuhn, C.; Doyle, C.; Slayback, D.; Brakenridge, G. R.; Sullivan, J.; and Erickson, T. A. 2021. Satellite observations indicate increasing proportion of population exposed to floods. *Nature*, 596: 80–86.

Li, Q.; Yang, W.; Liu, W.; Yu, Y.; and He, S. 2021. From Contexts to Locality: Ultra-High Resolution Image Segmentation via Locality-Aware Contextual Correlation. In *ICCV*, 7252–7261.

Li, Y.; Dang, B.; Zhang, Y.; and Du, Z. 2022a. Water body classification from high-resolution optical remote sensing imagery: Achievements and perspectives. *ISPRS Journal of Photogrammetry and Remote Sensing*, 187: 306–327.

Li, Y.; and Niu, Z. 2022. Systematic method for mapping fine-resolution water cover types in China based on time series Sentinel-1 and 2 images. *International Journal of Applied Earth Observation and Geoinformation*, 106: 102656.

Li, Z.; Zhang, H.; Lu, F.; Xue, R.; Yang, G.; and Zhang, L. 2022b. Breaking the resolution barrier: A low-to-high network for large-scale high-resolution land-cover mapping using low-resolution labels. *ISPRS Journal of Photogrammetry and Remote Sensing*, 192: 244–267.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *ICCV*, 2980–2988.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *CVPR*, 3431–3440.

Luo, X.; Tong, X.; and Hu, Z. 2021. An applicable and automatic method for earth surface water mapping based on multispectral images. *International Journal of Applied Earth Observation and Geoinformation*, 103: 102472.

McFeeters, S. K. 1996. The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features. *International journal of remote sensing*, 17(7): 1425–1432.

Miao, Z.; Fu, K.; Sun, H.; Sun, X.; and Yan, M. 2018. Automatic water-body segmentation from high-resolution satellite images via deep networks. *IEEE geoscience and remote sensing letters*, 15(4): 602–606.

Min, J.; Zhao, Y.; Luo, C.; and Cho, M. 2022. Peripheral Vision Transformer. In *NeurIPS*.

Moortgat, J.; Li, Z.; Durand, M.; Howat, I.; Yadav, B.; and Dai, C. 2022. Deep learning models for river classification at sub-meter resolutions from multispectral and panchromatic commercial satellite imagery. *Remote Sensing of Environment*, 282: 113279.

Pekel, J.-F.; Cottam, A.; Gorelick, N.; and Belward, A. S. 2016. High-resolution mapping of global surface water and its long-term changes. *Nature*, 540(7633): 418–422.

Robinson, C.; Hou, L.; Malkin, K.; Soobitsky, R.; Czawlytko, J.; Dilkina, B.; and Jojic, N. 2019. Large Scale High-Resolution Land Cover Mapping With Multi-Resolution Data. In *CVPR*.

Seale, C.; Redfern, T.; Chatfield, P.; Luo, C.; and Dempsey, K. 2022. Coastline detection in satellite imagery: A deep learning approach on new benchmark data. *Remote Sensing of Environment*, 278: 113044.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 618–626.

Toker, A.; Kondmann, L.; Weber, M.; Eisenberger, M.; Camero, A.; Hu, J.; Hoderlein, A. P.; Şenaras, c.; Davis, T.; Cremers, D.; Marchisio, G.; Zhu, X. X.; and Leal-Taixé, L. 2022. DynamicEarthNet: Daily Multi-Spectral Satellite Dataset for Semantic Change Segmentation. In *CVPR*, 21158–21167.

Tong, X.-Y.; Xia, G.-S.; and Zhu, X. X. 2023. Enabling country-scale land cover mapping with meter-resolution satellite imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 196: 178–196.

Vörösmarty, C. J.; McIntyre, P. B.; Gessner, M. O.; Dudgeon, D.; Prusevich, A.; Green, P.; Glidden, S.; Bunn, S. E.; Sullivan, C. A.; Liermann, C. R.; et al. 2010. Global threats to human water security and river biodiversity. *nature*, 467(7315): 555–561.

Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. 2020a. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10): 3349–3364.

Wang, J.; Zheng, Z.; Ma, A.; Lu, X.; and Zhong, Y. 2021. LoveDA: A Remote Sensing Land-Cover Dataset for Domain Adaptive Semantic Segmentation. In Vanschoren, J.; and Yeung, S., eds., *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Wang, X.; Xiao, X.; Qin, Y.; Dong, J.; Wu, J.; and Li, B. 2022. Improved maps of surface water bodies, large dams, reservoirs, and lakes in China. *Earth System Science Data*, 14(8): 3757–3771.

Wang, X.; Xiao, X.; Zou, Z.; Dong, J.; Qin, Y.; Doughty, R. B.; Menarguez, M. A.; Chen, B.; Wang, J.; Ye, H.; et al. 2020b. Gainers and losers of surface and terrestrial water resources in China during 1989–2016. *Nature communications*, 11(1): 3471.

Wieland, M.; Martinis, S.; Kiefl, R.; and Gstaiger, V. 2023. Semantic segmentation of water bodies in very high-resolution satellite and aerial images. *Remote Sensing of Environment*, 287: 113452.

Wu, W.; Li, Q.; Zhang, Y.; Du, X.; and Wang, H. 2018. Two-step urban water index (TSUWI): A new technique for high-resolution mapping of urban surface water. *Remote sensing*, 10(11): 1704.

Xu, H. 2006. Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery. *International journal of remote sensing*, 27(14): 3025–3033.

Yao, F.; Wang, C.; Dong, D.; Luo, J.; Shen, Z.; and Yang, K. 2015. High-resolution mapping of urban surface water using ZY-3 multi-spectral imagery. *Remote Sensing*, 7(9): 12336–12355.

Yu, Y.; Yao, Y.; Guan, H.; Li, D.; Liu, Z.; Wang, L.; Yu, C.; Xiao, S.; Wang, W.; and Chang, L. 2021. A self-attention capsule feature pyramid network for water body extraction from remote sensing imagery. *International Journal of Remote Sensing*, 42(5): 1801–1822.

Zhang, Z.; Lu, M.; Ji, S.; Yu, H.; and Nie, C. 2021. Rich CNN features for water-body segmentation from very high resolution aerial and satellite imagery. *Remote Sensing*, 13(10): 1912.

Zhao, H.; Shi, J.; Qi, X.; Wang, X.; and Jia, J. 2017. Pyramid scene parsing network. In *CVPR*, 2881–2890.