

Article

WetVeg-2mm: An Ultra-High-Resolution UAV Dataset for Riparian Vegetation Semantic Segmentation

Guiqi Liu, Runqiao Zhang  and Huapeng Qin *

Eco-Environment and Resource Efficiency Research Laboratory, School of Environment and Energy, Peking University Shenzhen Graduate School, Shenzhen 518055, China; 2401212902@stu.pku.edu.cn (G.L.); zhangrq@stu.pku.edu.cn (R.Z.)

* Correspondence: qinhp@pku.edu.cn; Tel.: +86-755-26035291

Highlights

What are the main findings?

- We developed WetVeg-2mm, an ultra-high-resolution UAV dataset with 2054 image chips, 2 mm ground sampling distance, and 17 semantic classes for fine-grained riparian vegetation segmentation.
- Benchmark experiments with U-Net, Attention U-Net, DeepLabV3+, PSPNet and SegFormer show that SegFormer with ImageNet pretraining achieved the best overall performance on the held-out test set, with 76.23% *mIoU*, 86.01% *mDice*, 85.17% *PA*, 87.30% *Recall* and 85.42% *Precision*.

What are the implications of the main findings?

- WetVeg-2mm provides a reproducible benchmark for evaluating semantic segmentation methods in complex riparian vegetation mosaics using ultra-high-resolution UAV imagery.
- The dataset can support future studies on ecological monitoring, vegetation mapping, rare-class recognition, and robust segmentation under fragmented and visually similar plant communities.

Abstract

Fine-grained mapping of riparian vegetation is important for ecological monitoring, invasive species control, and ecosystem restoration. However, riparian plant communities often exhibit fragmented patches, broad transition zones and high visual similarity among classes, making stable species-level segmentation difficult from conventional satellite imagery or lower-resolution UAV imagery. To address this gap, we present WetVeg-2mm, an ultra-high-resolution UAV dataset for fine-grained riparian vegetation semantic segmentation. Built from UAV surveys over a representative riparian section of the Jiuzhou River in Guangxi, China, the dataset provides 2054 image chips (1024×1024) with pixel-level annotations at 2 mm ground sampling distance. It contains 17 semantic classes in total, including 14 representative wetland plant classes, such as *Colocasia*, *Eichhornia* and *Phragmites*, together with water, bareland and background. Five baseline models, namely U-Net, Attention U-Net, DeepLabV3+, PSPNet and SegFormer, were evaluated using per-class *IoU*, *mIoU*, *mDice*, *PA*, *Precision* and *Recall*. Across all evaluated baseline settings, SegFormer with ImageNet pretraining achieved the best overall performance, with 76.23% *mIoU*, 86.01% *mDice*, 85.17% *PA*, 87.30% *Recall* and 85.42% *Precision* on the test set. Overall, WetVeg-2mm provides a reproducible and challenging benchmark for fine-grained riparian vegetation semantic segmentation.



Academic Editor: Ran Meng

Received: 19 May 2026

Revised: 4 July 2026

Accepted: 6 July 2026

Published: 1 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: riparian vegetation; ultra-high-resolution UAV dataset; semantic segmentation; ecological monitoring

1. Introduction

River riparian zones are ecologically important transition areas between terrestrial and aquatic systems and are often highly sensitive to hydrological fluctuations and environmental disturbances driven by climate change and human activities [1–3]. Fine-grained and repeatable mapping of riparian vegetation is therefore important for ecological monitoring, invasive species control, and ecosystem restoration [4,5]. Satellite remote sensing has been widely used for vegetation and wetland mapping, but its effectiveness is often constrained by spatial resolution and atmospheric conditions. In riparian environments characterized by narrow corridors, fragmented patches and broad transition zones, imagery with spatial resolutions coarser than one meter often suffers from mixed pixels, blurred boundaries and class confusion [4,6–8]. In contrast, unmanned aerial vehicles (UAVs) provide ultra-high spatial resolution, flexible deployment and near-ground observation capabilities, making them more suitable for capturing vegetation texture, canopy boundaries and fine spatial heterogeneity [9–13]. In this study, ultra-high spatial resolution refers to UAV RGB imagery with a millimeter-level ground sampling distance (GSD), which is sufficient to resolve fine vegetation texture, small patches and narrow canopy boundaries.

Before the widespread use of deep learning, remote-sensing image classification and segmentation mainly relied on hand-crafted spectral, textural, shape-based and object-based features, often combined with traditional machine learning classifiers such as support vector machines and random forests [14,15]. These methods remain useful in relatively simple or well-controlled scenes, but their adaptability is limited when natural landscapes contain heterogeneous backgrounds, many categories and visually similar objects. With the development of deep learning, remote-sensing image interpretation has gradually shifted from manually designed features to end-to-end feature learning. Recent reviews show that deep learning has been widely applied to remote-sensing image classification, object detection, semantic segmentation, change detection and environmental monitoring [16,17]. Advanced methods for hyperspectral anomaly detection and maritime ship detection further illustrate the versatility of deep learning across diverse remote-sensing applications [18,19].

Semantic segmentation is a pixel-level image interpretation task that assigns a semantic class label to each pixel. Unlike image classification, which predicts image-level categories, or object detection, which usually outputs bounding boxes, semantic segmentation can represent continuous boundaries, spatial shapes and area extents of land-cover objects. This makes it particularly suitable for mapping vegetation patches, water–land interfaces, bareland edges and mixed transition zones. Deep learning-based segmentation models have evolved around encoder–decoder structures, multi-scale contextual representation, attention mechanisms and boundary recovery [20]. In remote sensing, models such as DeepLabV3+ and ResUNet-a show that multi-scale context, local boundary preservation and class-imbalance-aware learning are important for high-resolution semantic segmentation [21,22]. However, the performance and comparability of these models depend heavily on task-specific datasets with reliable pixel-level annotations and standardized evaluation protocols.

For vegetation segmentation, the increasing availability of UAV imagery has promoted the construction of scene-specific benchmark datasets [16]. In agriculture, datasets such as WeedMap have supported crop and weed mapping from UAV imagery [23]. In forestry,

the FOR-instance dataset provides UAV laser scanning data for semantic and instance segmentation of individual trees [24]. In wetland and aquatic settings, AqUavplant and recent coastal or intertidal vegetation datasets have begun to support UAV-based segmentation of aquatic plants and other coastal biological cover types [25–27]. These studies show that dataset development plays an important role in benchmark evaluation and method development for vegetation segmentation tasks.

However, riparian vegetation mapping presents a distinct set of challenges that are not fully addressed by existing datasets. Plant communities in river riparian zones are often tightly interwoven, with fragmented distributions, complex water–land interfaces, shadows, reflections and frequent transition areas. Practical management also requires fine-grained discrimination among co-occurring riparian plant types rather than coarse vegetation–background separation. Image-level presence and instance-count statistics can describe class occurrence frequency, patch fragmentation and community composition, providing fine-scale structural indicators for future ecological monitoring. Although existing aquatic and coastal vegetation datasets provide valuable starting points [25–27], they do not fully represent the species composition, spatial patterns and annotation requirements of river riparian zones. A public benchmark that combines millimeter-level UAV RGB imagery, fine-grained riparian vegetation classes, pixel-level annotations and an official split for reproducible comparison is therefore still needed.

To address this gap, this study constructs WetVeg-2mm, an ultra-high-resolution UAV RGB dataset for fine-grained riparian vegetation semantic segmentation. WetVeg-2mm was built from UAV surveys over a representative riparian section of the Jiuzhou River in Guangxi, China, and contains 2054 annotated image chips of size 1024×1024 pixels at 2 mm GSD. It contains 17 semantic classes in total, including 14 representative wetland plant classes, such as *Colocasia*, *Eichhornia* and *Phragmites*, together with water, bareland and background. To support reproducible benchmarking, WetVeg-2mm provides pixel-level semantic annotations, an official spatially grouped train/validation/test split and a controlled public-release protocol.

Motivated by these challenges, this study focuses on whether millimeter-level UAV RGB imagery can support fine-grained riparian vegetation segmentation, how representative segmentation architectures perform under fragmented and visually similar riparian scenes, and what class-level and scene-level patterns reveal the main difficulties of this task.

Our contributions are summarized as follows:

1. To address the scarcity of publicly available, accurately annotated ultra-high-resolution UAV datasets for fine-grained riparian vegetation segmentation, this study constructs WetVeg-2mm using RGB imagery acquired over a representative riparian section of the Jiuzhou River in Guangxi, China. The dataset contains 2054 image chips of size 1024×1024 pixels at 2 mm GSD and provides pixel-level annotations for 17 semantic classes, including 14 representative wetland plant classes, water, bareland and background.
2. To support reproducible model training and evaluation, WetVeg-2mm provides an official spatially grouped train/validation/test split and a controlled public-release protocol. This design reduces spatial leakage between data subsets and enables consistent comparison among future segmentation algorithms.
3. To establish an initial benchmark for the proposed dataset, this study evaluates representative CNN-based and transformer-based semantic segmentation models, including U-Net, Attention U-Net, DeepLabV3+, PSPNet and SegFormer. The benchmark results provide a quantitative reference for future studies on fine-grained riparian vegetation segmentation.

4. To further understand the characteristics and challenges of WetVeg-2mm, this study analyzes class-level and scene-level segmentation patterns under fragmented patches, complex water–land interfaces, fine boundaries, class imbalance and high visual similarity among vegetation classes. These analyses reveal the dataset’s inherent challenges and provide guidance for future model development and ecological monitoring applications.

2. Dataset Construction

2.1. Study Area

The dataset was collected from a representative riparian section of the Jiuzhou River in Guangxi, China (Figure 1). Under the combined influence of hydrological processes and human activities, riparian zones along the river often exhibit complex shoreline forms and heterogeneous vegetation patterns. The study area is generally well preserved, covers approximately 5 ha and is representative for fine-grained riparian vegetation mapping and benchmark evaluation.

Climatically, the study area belongs to the subtropical monsoon climate zone with a clear marine influence in Qinzhou, Guangxi. Regional climate information indicates an annual mean air temperature of approximately 21–23 degrees C, annual sunshine duration of about 1800 h and annual precipitation of about 1600 mm [28]. Monthly climate information for Qinzhou further shows clear seasonality, with precipitation lowest in winter and highest in summer; potential solar insolation also varies seasonally, with day length ranging from about 10 h 48 min in late December to about 13 h 29 min in late June [29]. This warm, humid and seasonally wet environment supports vigorous riparian vegetation growth, but it also promotes seasonal water-level fluctuation, mixed water–land transition zones and heterogeneous vegetation patches, thereby increasing the difficulty of fine-grained semantic segmentation.

From a landscape perspective, the study area is characterized by strong land–water interspersions and diverse vegetation cover (Figure 2). It includes aquatic and emergent vegetation, herbaceous communities, shrubs and trees, together with water and bareland. These components form fragmented patches and broad transition zones, creating a complex riparian scene. Different vegetation types often vary in texture, canopy structure and habitat position, but some classes also share similar visual features and frequently occur along mixed boundaries. This makes the study area a realistic and challenging setting for fine-grained semantic segmentation.

The overall workflow for constructing WetVeg-2mm is illustrated in Figure 3, including UAV data acquisition, orthomosaic generation and tiling, manual annotation, and semantic mask generation.

2.2. Data Acquisition

Data acquisition was conducted in October 2025. The field survey was carried out using a DJI Matrice 350 RTK UAV (DJI, Shenzhen, China) equipped with a SHARE 100M A10 medium-format metric camera (Share UAV, Shenzhen, China) to capture ultra-high-resolution imagery. An RTK positioning scheme was adopted during the survey to improve the stability of the exterior orientation parameters and the geometric accuracy of orthomosaic reconstruction. The survey was conducted under clear weather conditions between 09:00 and 17:00 local time. To balance the need for fine-grained riparian vegetation recognition with field efficiency, the flight altitude was set to 56 m, the flight speed to 4 m/s, the side overlap to 60% and the forward overlap to 70% and all images were acquired in nadir view. Under these settings, the final orthomosaics achieved a ground sampling distance of

2 mm, allowing fine vegetation texture and canopy boundaries to be clearly represented. The sensor parameters are listed in Table 1.

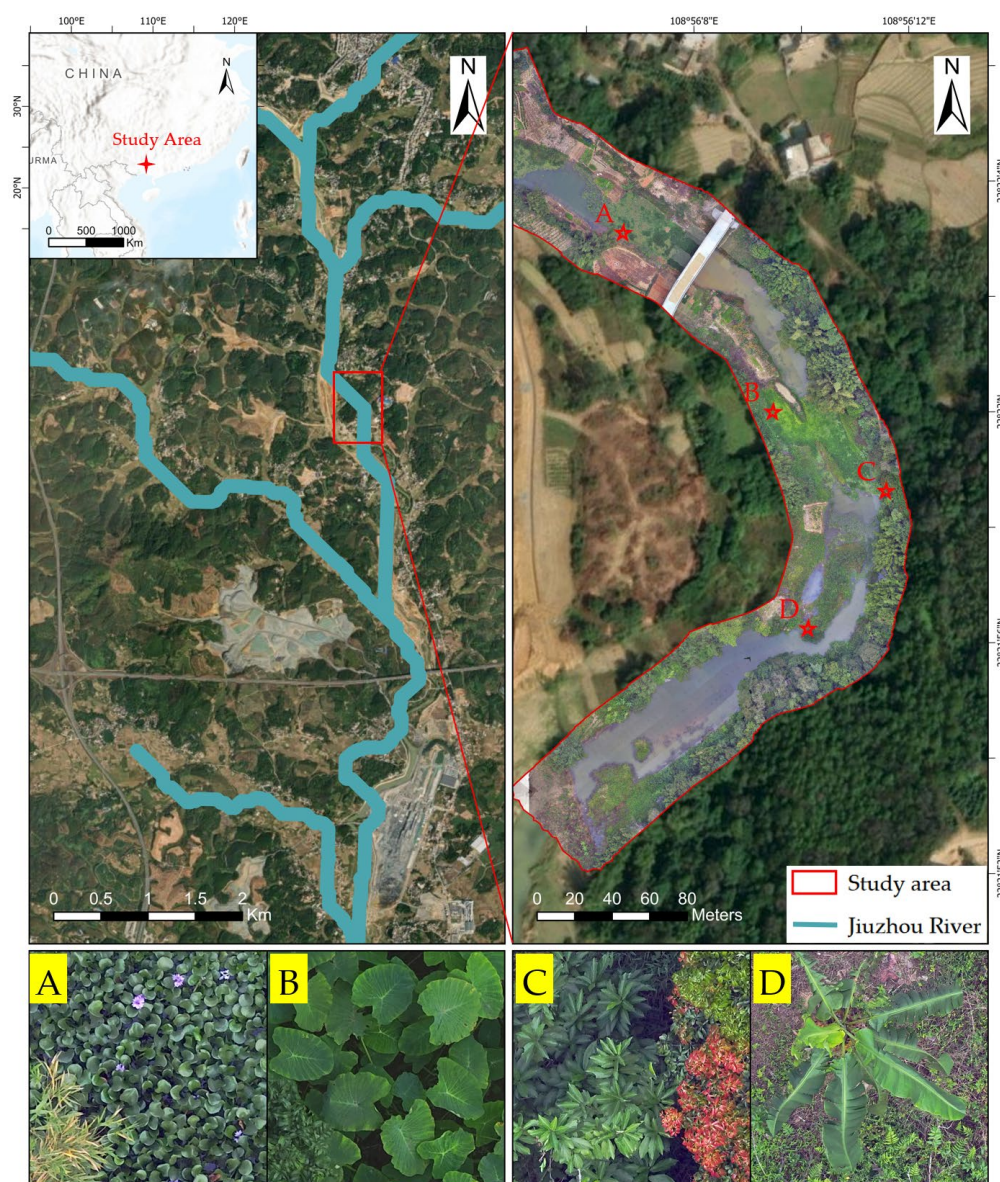


Figure 1. Geographic location of the study area in a representative riparian section of the Jiuzhou River, Guangxi, China. The red stars marked A–D indicate the locations of representative riparian vegetation examples shown in the lower panels.

Table 1. Key specifications of the UAV camera used for data acquisition.

Sensor	SHARE 100 M A10
Resolution (H × V)	11,648 × 8736
Focal Length (mm)	110
Sensor Size (H × V mm)	43.8 × 32.9
Pixel Size (μm)	3.76 × 3.76
Effective Pixels (MP)	102

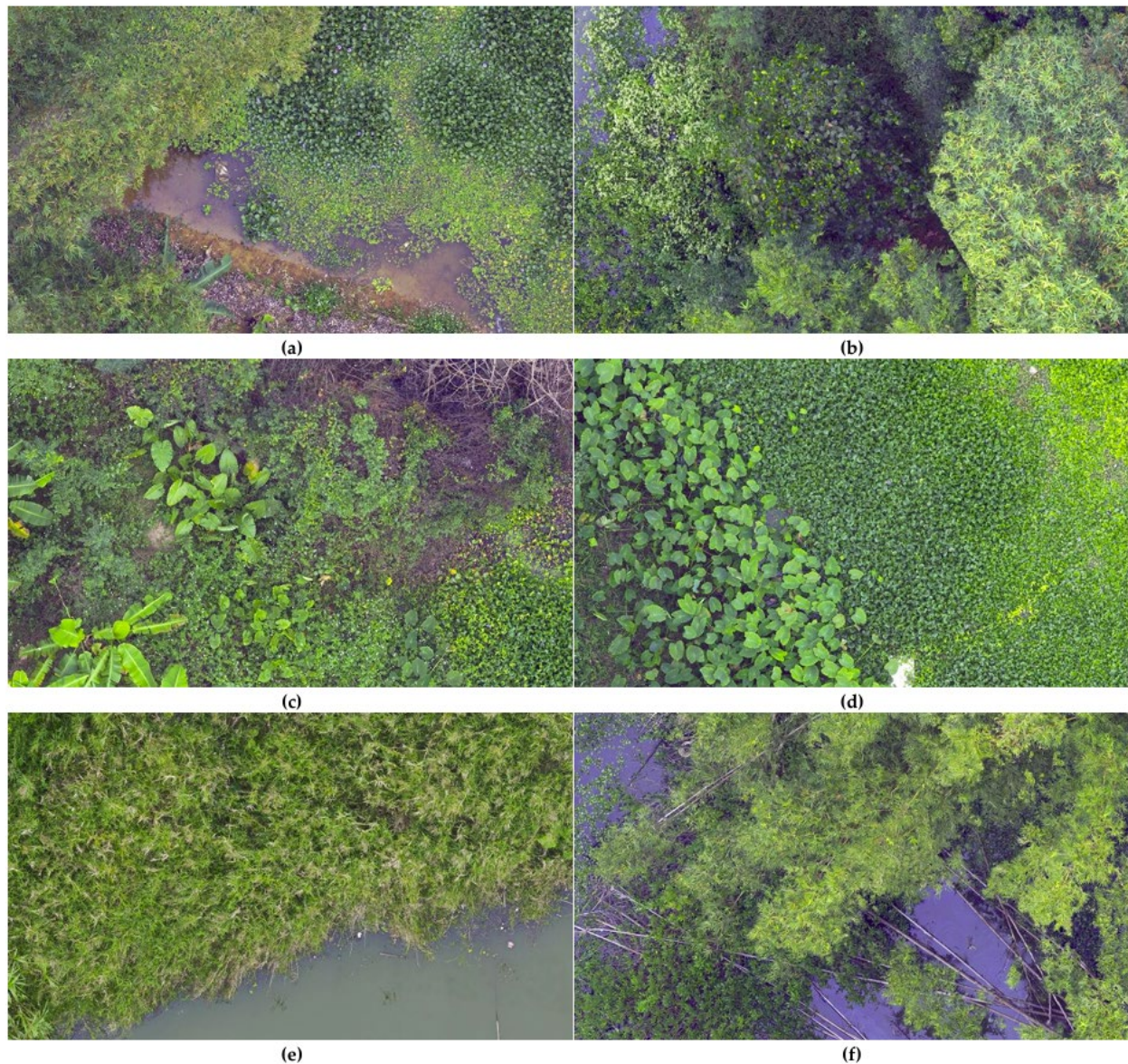


Figure 2. Representative UAV orthomosaic scenes of riparian vegetation. (a) Vegetation in water and land transitional zones. (b) Mixed vegetation under sunlight and shadow. (c) Mixed vegetation with multiple plant types. (d) Contiguous dominant vegetation cover. (e) Emergent herbaceous vegetation along water margins. (f) Bamboo and woody vegetation along water margins.

2.3. Data Preprocessing

During orthomosaic generation, three commonly used photogrammetric software packages were compared, namely Agisoft Metashape Professional v2.2.2, Pix4Dmapper v4.10.0 and DJI Terra v5.1.0. To ensure a fair comparison, repeated stitching and reconstruction were performed using the same raw imagery and the results were evaluated in terms of mosaic completeness and reconstruction accuracy. DJI Terra produced the most complete mosaics and the best geometric quality for the riparian scene and was therefore selected for final orthomosaic generation.

Basic quality control was then applied to ensure that the orthomosaics were suitable for millimeter-level vegetation interpretation and annotation. Images affected by severe blur, abnormal exposure, or insufficient overlap were removed and the resulting orthomosaics were further inspected to ensure the absence of large stitching gaps, obvious

ghosting, or noticeable geometric distortion. Representative orthomosaic examples are shown in Figure 2.

After orthomosaic generation and quality control, the imagery was tiled into 1024×1024 pixel chips to facilitate subsequent annotation and model training while preserving millimeter-level texture detail. Regular tiling of the study-area orthomosaics produced 10,459 candidate chips. Because many of these chips contained only water, bare ground, or limited useful semantic content, annotating all of them would have greatly increased the labeling workload while contributing little to model development. Therefore, chips with limited annotation value, poor image quality, or unclear target content were excluded and 2054 informative chips were retained for pixel-level annotation to form the final dataset.

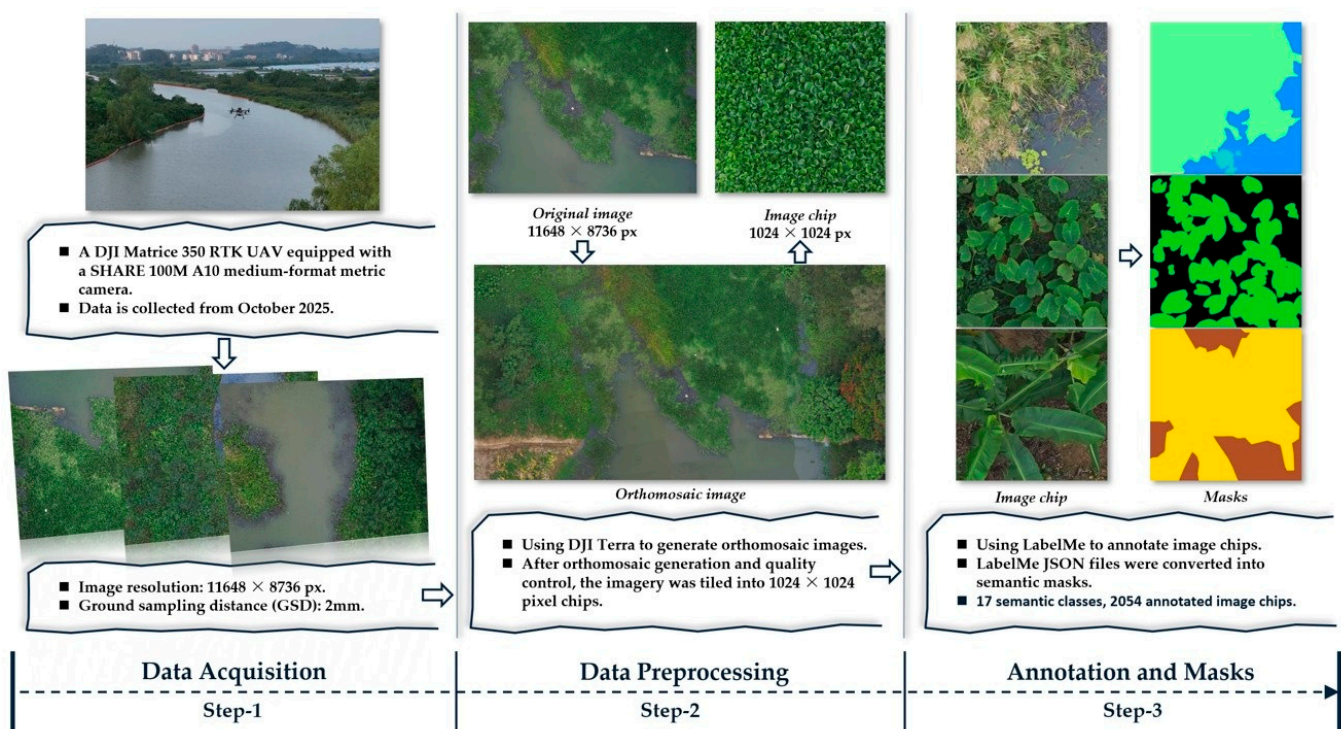


Figure 3. The step-by-step approach to prepare the dataset. Different colors in the masks indicate different semantic classes.

2.4. Label Generation

This dataset adopts pixel-level semantic annotations for fine-grained riparian vegetation segmentation. The label generation process consisted of taxonomy design, manual annotation and multi-round quality control.

2.4.1. Taxonomy and Label Definitions

The dataset contains 17 semantic classes in total, including 14 representative wetland plant classes, such as *Colocasia*, *Eichhornia* and *Phragmites*, together with water, bareland and background. The detailed class names, scientific names and label IDs are listed in Table 2. The taxonomy was designed to balance ecological relevance and annotation feasibility. It covers representative and management-relevant classes commonly found in riparian vegetation communities, while restricting the categories to those with relatively stable visual cues in 2 mm orthomosaics, such as texture, canopy form, color and habitat position. In this study, the 16 non-background classes refer to all classes except background, including water, bareland, and the 14 wetland plant classes. In WetVeg-2mm, background denotes unlabeled regions that were not assigned to any of the 16 non-background classes.

Table 2. Class list with scientific names and label IDs of the proposed dataset.

Class Name	Scientific Name	Class Label
Background *	—	0
Water	—	1
Bareland	—	2
Bambusa	<i>Bambusa multiplex</i>	3
Rhodomyrtus	<i>Rhodomyrtus tomentosa</i>	4
Aeschynomene	<i>Aeschynomene indica</i>	5
Pongamia	<i>Pongamia pinnata</i>	6
Bidens	<i>Bidens alba</i>	7
Syzygium	<i>Syzygium nervosum</i>	8
Eichhornia	<i>Eichhornia crassipes</i>	9
Colocasia	<i>Colocasia esculentum</i> var. <i>antiquorum</i>	10
Alternanthera	<i>Alternanthera philoxeroides</i>	11
Musa	<i>Musa balbisiana</i>	12
Diplazium	<i>Diplazium esculentum</i>	13
Arundo	<i>Arundo donax</i>	14
Pueraria	<i>Pueraria montana</i> var. <i>lobata</i>	15
Phragmites	<i>Phragmites australis</i>	16

* Background is assigned label ID 0 and denotes unlabeled regions.

2.4.2. Annotation Protocol

Pixel-level annotation was performed in LabelMe v5.10 [30], and the polygon annotations were saved in JSON format. During annotation, visible canopy boundaries, texture changes and habitat context in the orthomosaic were used as the main cues and polygons were drawn as closely as possible to the actual patch boundaries. In mixed or transitional areas, where different vegetation types often show similar local textures and gradual boundary changes, a texture-consistency principle was adopted. Regions with continuous and stable texture and shape characteristics were assigned to the same class, whereas clear changes in texture or morphology were treated as class boundaries. Areas affected by reflections, shadows, or partial occlusion were interpreted with reference to surrounding texture continuity and spatial context.

After manual annotation, the LabelMe JSON files were converted into pixel-level semantic masks through a scripted rasterization workflow. In this workflow, manually delineated polygons were rasterized according to the predefined WetVeg-2mm class-ID order, so that each pixel inside an annotated polygon was assigned the corresponding semantic label. The final masks were saved as single-channel indexed PNG files with a color palette for visual inspection. Pixels not covered by any annotated polygon retained the default value 0 and were therefore assigned to the background class, whereas missing images and labels outside the predefined class list were checked before conversion. Representative annotation examples are shown in Figure 4.

2.4.3. Quality Control

To ensure label quality and consistency, a multi-round workflow of annotation, inspection, correction and review was adopted. After initial labeling, 30% of the annotated chips were randomly inspected, with particular attention to class confusion, boundary misalignment, missing labels and over-labeling. Problematic samples were then corrected and the annotation guidelines were refined accordingly. In addition, the annotations were reviewed by personnel with botanical knowledge, especially for easily confused classes. In total, three rounds of revision and review were carried out to ensure stable and consistent labels across categories and scene conditions.



Figure 4. Representative annotation examples from the WetVeg-2mm dataset. The top row shows orthomosaic image chips and the bottom row shows the corresponding semantic masks.

2.5. Official Split and Leakage Prevention Strategy

To prevent spatial leakage caused by neighboring image chips appearing in both the training and test subsets, a geospatial grouping strategy was adopted instead of random chip-level splitting [31,32]. Specifically, according to the spatial positions of the chips in the orthomosaic, the north–south extent of the riparian corridor in the study area was divided into 10 consecutive spatial segments. The split was performed at the segment level and all chips within the same segment were assigned exclusively to either the training, validation, or test subset. This design ensures spatial independence among the subsets. Compared with random splitting, it reduces evaluation bias caused by spatial autocorrelation and provides a more realistic assessment of model generalization to spatially unseen areas [31,32].

Based on this strategy, segments 1, 2, 3, 5, 6 and 7 were assigned to the training set, segments 4 and 8 to the validation set and segments 9 and 10 to the test set, resulting in 1227, 405 and 422 image chips, respectively, with an overall ratio of approximately 6:2:2. Because riparian vegetation is highly patchy and spatially clustered, class distribution is not spatially uniform. Therefore, the validation and test segments were selected to balance coverage of the main scene components and representative classes while preserving spatial separation. The detailed segment allocation is summarized in Table 3.

Table 3. Segment-based official split and sample allocation.

Segment ID (N→S)	Images in Segment	Assigned Split
1	150	Train
2	231	Train
3	344	Train
4	210	Val
5	102	Train
6	355	Train
7	45	Train
8	195	Val
9	177	Test
10	245	Test
Total	2054	—

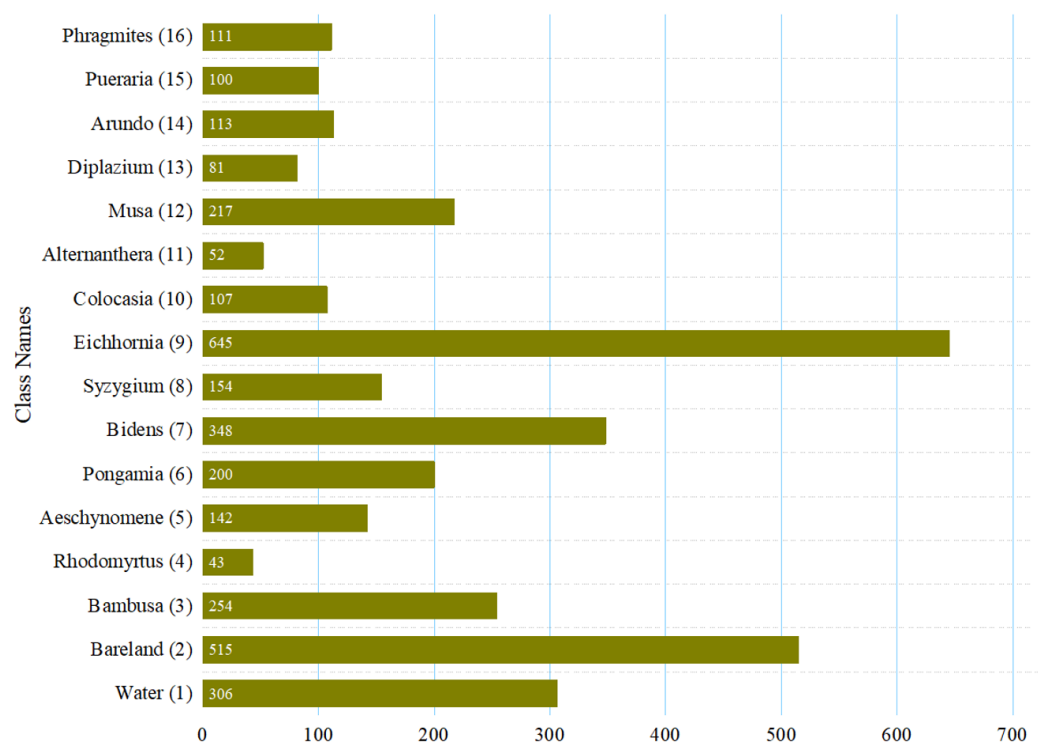
2.6. Dataset Statistics

To characterize the composition of the dataset and the imbalance among the 16 non-background classes (i.e., all classes except background), three complementary statistics were used: image-level presence, instance counts and pixel-area proportions. Image-level presence is defined as the number of image chips containing a given class; instance counts represent the number of annotated patches for each class and reflect the degree of patch fragmentation; and pixel-area proportions indicate the fraction of labeled pixels belonging to each class, thereby describing class dominance in terms of area. Because a single image chip may contain multiple classes, image-level presence is a non-exclusive statistic and the summed counts across classes generally exceed the number of images in a given subset. The dataset contains 2054 annotated image chips in total, which were split into 1227 training images, 405 validation images, and 422 testing images, while the per-class image-level presence across the three subsets is reported in Table 4.

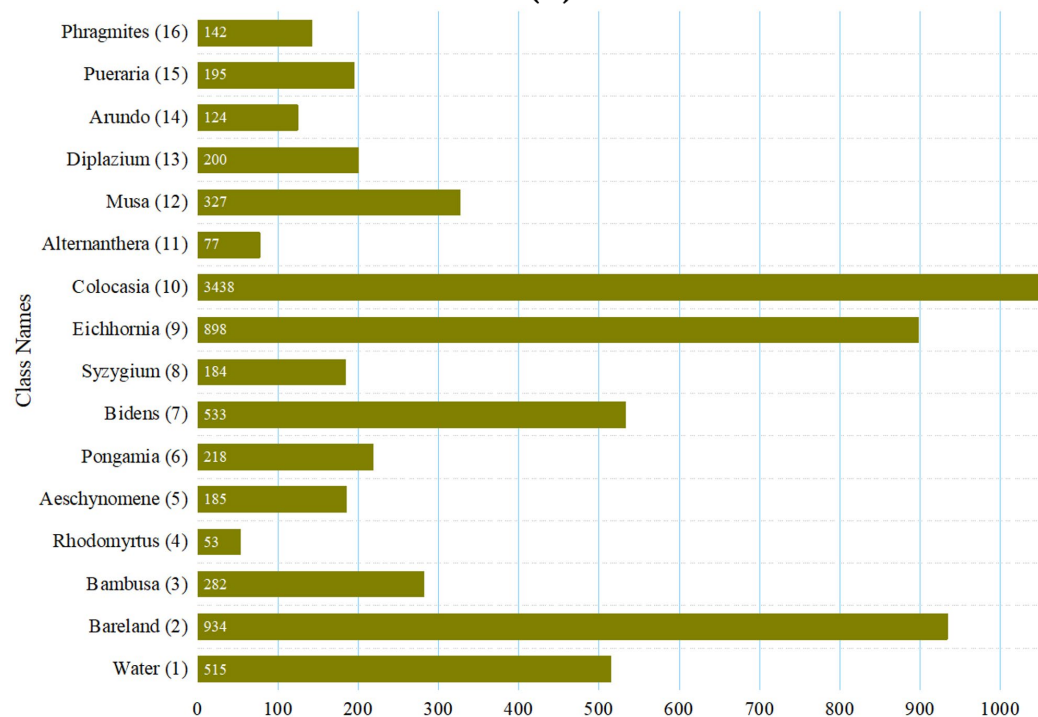
Table 4. Image-level class presence in the training, validation and testing subsets.

Class Name	Training Samples	Validation Samples	Testing Samples	Total Samples
Water	182	66	58	306
Bareland	309	105	101	515
Bambusa	156	48	50	254
Rhodomyrtus	25	8	10	43
Aeschynomene	88	26	28	142
Pongamia	124	39	37	200
Bidens	206	76	66	348
Syzygium	92	30	32	154
Eichhornia	384	126	135	645
Colocasia	68	15	24	107
Alternanthera	29	11	12	52
Musa	129	46	42	217
Diplazium	51	15	15	81
Arundo	69	21	23	113
Pueraria	59	20	21	100
Phragmites	64	24	23	111

Overall, the dataset shows clear class imbalance and substantial variation in patch scale. In terms of image-level presence, Eichhornia is the most frequently occurring class, appearing in 645 image chips, followed by bareland (515) and Bidens (348). In contrast, Rhodomyrtus (43), Alternanthera (52) and Diplazium (81) appear in relatively few chips, indicating a clear long-tailed distribution. In terms of instance counts, Colocasia has by far the largest number of instances (3438), even though its image-level presence and pixel-area proportion are relatively low, suggesting that this class mainly occurs as numerous small and fragmented patches. In terms of pixel-area proportion, Eichhornia is dominant, accounting for 24.86% of all labeled non-background pixels, followed by bareland (11.44%) and Bambusa (9.48%). By contrast, Diplazium (0.66%), Alternanthera (1.23%) and Rhodomyrtus (1.50%) occupy much smaller areas. These statistics indicate that the dataset contains both dominant classes with large spatial coverage and underrepresented classes with fine-scale fragmentation, which together increase the difficulty of fine-grained semantic segmentation. The three statistics are further visualized in Figure 5.



(a)



(b)

Figure 5. Cont.

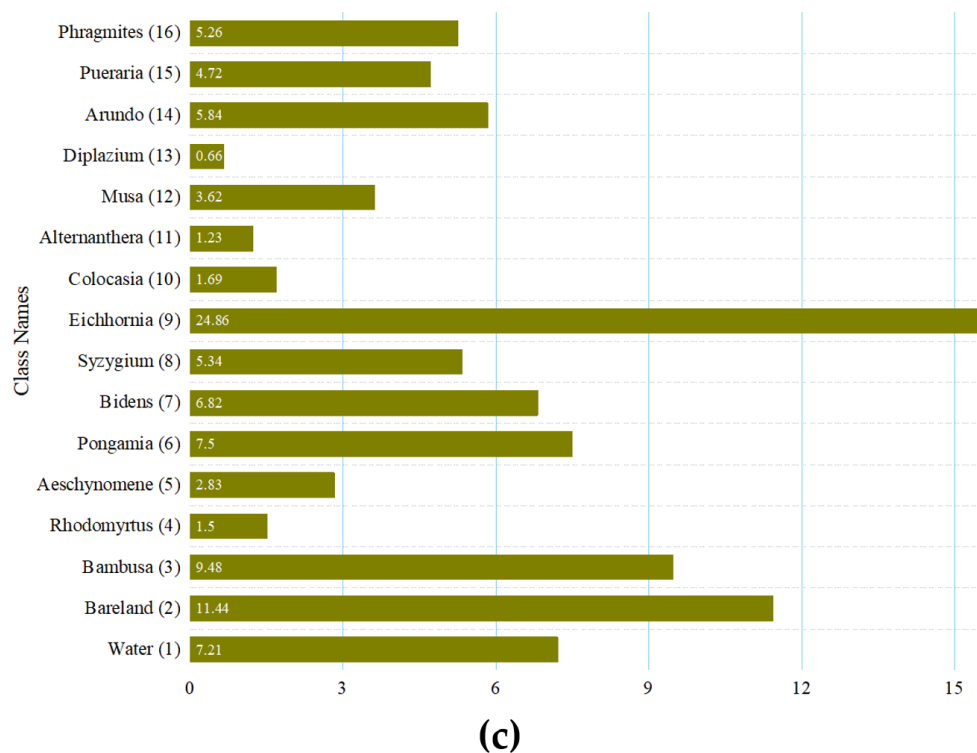


Figure 5. Distribution of the 16 non-background classes in the WetVeg-2mm dataset: (a) image-level presence, (b) instance counts and (c) pixel-area proportions (%).

2.7. Data Format and Release

The full internal WetVeg-2mm dataset is organized into original image chips, semantic segmentation masks, original polygon annotations, official split files, class information, and benchmark code, as illustrated in Figure 6. The image chips are stored in the JPEGImages folder as 1024×1024 pixel JPEG files, corresponding to a ground sampling distance of 2 mm. The semantic labels are stored in the SegmentationClass folder as single-channel PNG masks, with class IDs ranging from 0 to 16. The original polygon annotations are provided in LabelMe JSON format in the json folder. The ImageSets folder contains the official training, validation, and test split files, while Class_list.xlsx, README.md, and the code folder provide the class mapping, usage instructions, and benchmark training scripts, respectively.

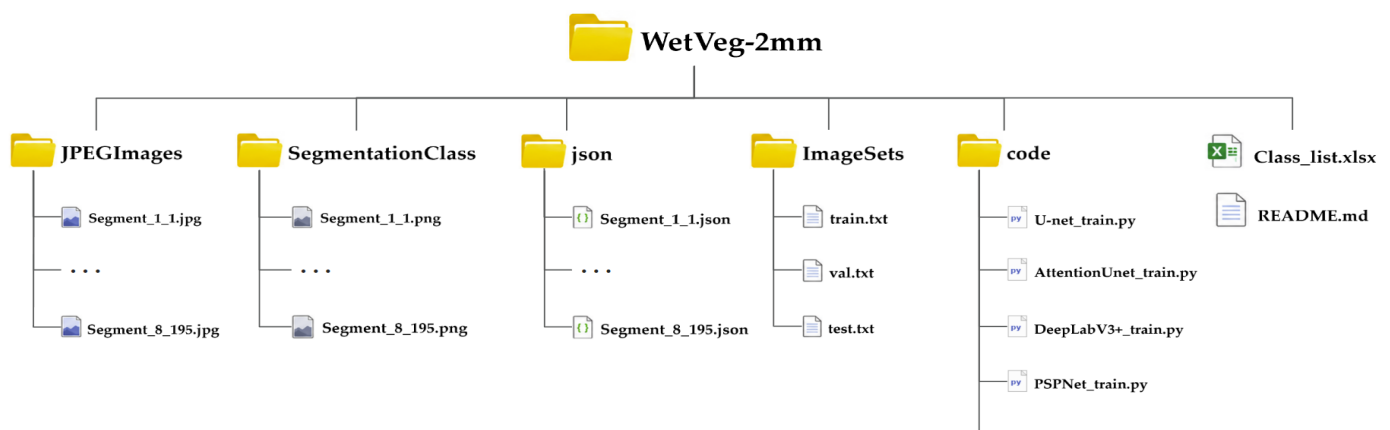


Figure 6. Folder structure of the WetVeg-2mm dataset.

The public release is distinguished from the full internal dataset to support controlled benchmarking. Upon publication, the publicly released version will be made available through Zenodo (DOI: 10.5281/zenodo.19849119). It will include the image chips, training and validation annotations, class information, official split files, and benchmark code. The semantic masks and polygon annotations for the test set will be withheld from public release; evaluation on the held-out test set can be requested from the authors. The benchmark code and usage instructions will also be available through GitHub at <https://github.com/sdsadsdd/WetVeg-2mm> (accessed on 5 July 2026). The dataset will be distributed under the CC BY 4.0 license.

3. Benchmark Experiments

3.1. Dataset Description

WetVeg-2mm is an ultra-high-resolution UAV dataset designed for fine-grained riparian vegetation semantic segmentation. The dataset comprises 2054 annotated image chips of size 1024×1024 pixels. In total, it contains 17 semantic classes, including 14 representative wetland plant classes, such as *Colocasia*, *Eichhornia* and *Phragmites*, together with water, bareland and background. For all benchmark experiments, the official split described in Section 2.5 was used, yielding 1227 training chips, 405 validation chips and 422 testing chips, with an overall ratio of approximately 6:2:2.

3.2. Evaluation Methods

To systematically evaluate the usability and challenge of the WetVeg-2mm dataset for fine-grained riparian vegetation semantic segmentation, a unified benchmark evaluation framework was established. Specifically, five representative semantic segmentation models were selected as baselines, namely U-Net [33], Attention U-Net [34], DeepLabV3+ [35], PSPNet [36] and SegFormer [37]. These models represent the classical encoder–decoder architecture, an attention-enhanced encoder–decoder architecture, a multi-scale context modeling framework based on atrous convolution, a global context aggregation framework based on pyramid pooling and a transformer-based architecture with hierarchical self-attention, respectively. Therefore, they provide a comprehensive basis for assessing how different architectural biases adapt to millimeter-level riparian vegetation segmentation.

All models were trained, validated and tested using the official spatial split provided in this study. Model parameters were learned from the training set, model selection was performed on the validation set and final quantitative results were obtained from the test set by comparing predictions with pixel-level annotations. To further examine the effect of pretrained initialization in ultra-high-resolution remote-sensing segmentation, each model was evaluated under two settings: Scratch, in which the model was randomly initialized and trained only on WetVeg-2mm, and PT, in which the encoder was initialized with ImageNet-pretrained weights [38] and then trained on WetVeg-2mm. Final performance was compared using the metrics defined in Section 3.4, including *IoU*, *mIoU*, *mDice*, *PA*, *Recall* and *Precision*.

3.2.1. U-Net

U-Net, originally proposed by Ronneberger et al., for biomedical image segmentation [33], adopts a symmetric encoder–decoder architecture with skip connections between corresponding stages. By fusing high-level semantic features with low-level spatial details, it is able to recover fine object boundaries while maintaining strong semantic representation. Owing to its simple design, stable training behavior and strong performance under limited training data, U-Net has become one of the most widely used baseline models in semantic

segmentation. In this study, it serves as a reference baseline for evaluating local detail preservation and boundary recovery on millimeter-resolution riparian vegetation imagery.

3.2.2. Attention U-Net

Attention U-Net is an extension of U-Net that introduces attention gates into the network architecture [34]. These gates selectively filter encoder features before they are passed to the decoder, suppressing irrelevant responses and enhancing target-related regions. Compared with standard U-Net, Attention U-Net places greater emphasis on selective feature refinement. In the riparian scenes considered in this study, where mixed vegetation patches and visually similar categories are common, Attention U-Net provides a useful baseline for examining whether attention mechanisms can improve non-background localization and fine-grained category discrimination.

3.2.3. DeepLabV3+

DeepLabV3+, proposed by Chen et al. [21], is an advanced semantic segmentation framework that combines atrous convolution with the atrous spatial pyramid pooling (ASPP) module to capture multi-scale contextual information more effectively. It further incorporates a decoder structure to refine object boundaries and recover spatial detail. Compared with conventional convolution-based models, DeepLabV3+ generally offers stronger contextual modeling ability in scenes with complex spatial structures and scale variation. For WetVeg-2mm, which contains fragmented vegetation patches, water–land transitions and intricate boundaries, DeepLabV3+ serves as a strong baseline for evaluating the contribution of multi-scale contextual features.

3.2.4. PSPNet

PSPNet, proposed by Zhao et al. [36], is designed to aggregate global contextual information through a pyramid pooling module operating at multiple spatial scales. This design enables the network to capture scene-level priors while preserving local detail, thereby improving segmentation performance in complex scenes. In riparian vegetation mapping along river corridors, different classes often exhibit strong patchiness, scale variation and heterogeneous local texture. PSPNet is therefore well suited as a baseline for assessing the role of global context modeling in millimeter-level fine-grained vegetation segmentation.

3.2.5. SegFormer

SegFormer, proposed by Xie et al. [37], is a transformer-based semantic segmentation framework that combines a hierarchical Transformer encoder with a lightweight all-MLP decoder. Its encoder produces multi-scale feature representations, while the MLP decoder aggregates features from different stages without relying on a complex convolutional decoder. Different from CNN-based architectures, SegFormer models long-range dependencies through self-attention and does not require fixed positional encodings, which improves its flexibility for inputs with different spatial layouts. In this study, SegFormer serves as a modern transformer-based baseline for evaluating whether global dependency modeling and multi-scale token representations can improve the segmentation of fragmented riparian vegetation and visually similar classes.

Overall, the five baseline models encode different architectural and mathematical assumptions. U-Net relies on local convolutional feature extraction and skip connections, making it suitable for recovering local spatial details and fine boundaries. Attention U-Net further introduces attention gates to selectively emphasize target-related features and suppress irrelevant background responses. DeepLabV3+ uses atrous convolution and atrous spatial pyramid pooling to enlarge the receptive field and capture multi-scale context. PSPNet aggregates scene-level priors through pyramid pooling, which is useful when class

recognition depends on broader spatial context. SegFormer represents a transformer-based alternative that uses self-attention to model long-range dependencies and hierarchical token representations. Based on the physical characteristics of WetVeg-2mm, U-Net-like models are expected to be effective for dense local structures and narrow boundaries, Attention U-Net may improve localization in mixed backgrounds, DeepLabV3+ and PSPNet are expected to perform well in scenes with scale variation and water–land contextual dependence, and SegFormer is expected to be competitive in fragmented and interwoven scenes where nonlocal contextual cues are important. Nevertheless, all models are expected to remain challenged by class imbalance and high visual similarity among fine-grained vegetation categories.

3.3. Implementation Details

As shown in Table 5, all experiments were conducted on a graphical workstation equipped with an Intel(R) Core(TM) i9-14900K processor @ 3.20 GHz, 160 GB RAM and an NVIDIA GeForce RTX 4090D GPU (24 GB VRAM), running Windows 11 Pro for Workstations. The implementation was carried out using Python 3.10.16, with model training and inference performed within the PyTorch 2.3.0 framework [39].

Table 5. The experimental environment and common training configuration used in this study.

Configuration Item	Setting
Operating system	Windows 11 Pro for Workstations
Processor	Intel(R) Core(TM) i9-14900K @ 3.20 GHz
RAM	160 GB
GPU	NVIDIA GeForce RTX 4090D (24 GB VRAM)
Programming language	Python 3.10.16
Deep learning framework	PyTorch 2.3.0
Input size	1024 × 1024 pixels
Input bands	RGB
Epochs	100
Batch size	2
Optimizer	AdamW
Initial learning rate	0.0001
Weight decay	0.0001
Random seed	42

Using the common configuration listed in Table 5, all baseline models were trained, validated, and tested using the official spatial split of WetVeg-2mm. The training set was used for parameter optimization, the validation set was used for model selection, and the held-out test set was used only for final quantitative evaluation. All RGB image chips were resized to 1024 × 1024 pixels, with bilinear interpolation for images and nearest-neighbor interpolation for semantic masks. Pixel values were converted to floating-point tensors and scaled to the range [0, 1]. During training, random horizontal flipping with a probability of 0.5 and random vertical flipping with a probability of 0.2 were applied, whereas no random augmentation was used during validation or testing. The objective function was the sum of cross-entropy loss and soft Dice loss. No learning-rate scheduler was used, and the best checkpoint for each model was selected according to the highest validation *mIoU*.

For the pretrained (PT) setting, the encoder was initialized with ImageNet-pretrained weights; for the Scratch setting, the corresponding model was trained from random initialization. In addition, the encoder/backbone settings were specified consistently across the benchmark. U-Net, Attention U-Net, DeepLabV3+ and PSPNet used ResNet-50 encoders/backbones in both Scratch and PT settings. SegFormer was evaluated with its

native MiT-B0 hierarchical Transformer encoder in both settings, with random initialization for Scratch and ImageNet-pretrained initialization for PT.

3.4. Evaluation Metrics

To evaluate the baseline models on the WetVeg-2mm dataset, five commonly used semantic segmentation metrics were adopted: mean Intersection over Union (*mIoU*), mean Dice coefficient (*mDice*), pixel accuracy (*PA*), *Recall* (*Rec*) and *Precision* (*Pre*) [40–43]. Among them, *mIoU* and *mDice* were used as the primary overall metrics, since they better reflect class-level overlap between predictions and ground-truth labels in fine-grained multi-class segmentation [41,42]. *PA* measures the proportion of correctly classified pixels, while *Precision* and *Recall* describe the reliability and completeness of the predictions, respectively [43]. In addition, *IoU* was further reported for each non-background class to analyze category-wise performance differences [40,41].

$$IoU_c = \frac{TP_c}{TP_c + FP_c + FN_c} \quad (1)$$

$$mIoU = \frac{1}{C} \sum_{c=0}^{C-1} \frac{TP_c}{TP_c + FP_c + FN_c} \quad (2)$$

$$mDice = \frac{1}{C} \sum_{c=0}^{C-1} \frac{2TP_c}{2TP_c + FP_c + FN_c} \quad (3)$$

$$PA = \frac{\sum_{c=0}^{C-1} TP_c}{\sum_{c=0}^{C-1} (TP_c + FN_c)} \quad (4)$$

$$Precision = \frac{1}{C} \sum_{c=0}^{C-1} \frac{TP_c}{TP_c + FP_c} \quad (5)$$

$$Recall = \frac{1}{C} \sum_{c=0}^{C-1} \frac{TP_c}{TP_c + FN_c} \quad (6)$$

For each class c , TP_c , FP_c and FN_c denote the numbers of true-positive, false-positive and false-negative pixels, respectively. The overall *mIoU*, *mDice*, *Precision* and *Recall* were obtained by averaging the class-wise results over all classes, whereas *PA* was computed over all labeled pixels. Pixels with ignore labels were excluded from all metric calculations.

For all reported evaluation metrics, higher values indicate better model performance, because they correspond to stronger spatial overlap, higher pixel-level correctness or better Precision–Recall agreement between model predictions and reference annotations.

4. Results

4.1. Overall Quantitative Results

To systematically evaluate the usability and challenge of WetVeg-2mm for fine-grained riparian vegetation segmentation, five representative baseline models, namely U-Net, Attention U-Net, DeepLabV3+, PSPNet and SegFormer, were compared under two initialization settings: Scratch, in which the model was randomly initialized and trained only on WetVeg-2mm, and PT, in which the encoder was initialized with ImageNet-pretrained weights and then trained on WetVeg-2mm. This resulted in ten benchmark experiments, all conducted using the official spatial split of the dataset. The quantitative results on the test set are summarized in Table 6.

Table 6. Quantitative comparison of baseline models on the test set.

Model		<i>mIoU</i> (%)	<i>mDice</i> (%)	<i>PA</i> (%)	<i>Rec</i> (%)	<i>Pre</i> (%)
U-Net	Scratch	29.23	38.65	57.72	38.40	49.34
	PT	44.93	56.04	70.27	53.63	74.48
Att. U-Net *	Scratch	29.73	38.30	59.58	37.42	59.74
	PT	43.93	55.52	69.22	53.08	73.12
DeepLabV3+	Scratch	35.01	47.46	61.94	45.46	60.58
	PT	69.50	80.61	83.57	78.69	84.48
PSPNet	Scratch	62.84	75.56	79.09	75.17	78.36
	PT	69.57	81.32	83.51	80.01	83.91
SegFormer	Scratch	65.91	77.67	80.62	78.90	77.07
	PT	76.23	86.01	85.17	87.30	85.42

* Att. U-Net = Attention U-Net; bold values indicate the best performance for each metric.

Overall, SegFormer with ImageNet pretraining achieves the best performance across all five overall metrics, reaching 76.23% *mIoU*, 86.01% *mDice*, 85.17% *PA*, 87.30% *Recall* and 85.42% *Precision*. Compared with the strongest pretrained CNN baselines, SegFormer-PT improves *mIoU* by 6.66 percentage points over PSPNet-PT and improves *PA* by 1.60 percentage points over DeepLabV3+-PT. These results indicate that transformer-based long-range dependency modeling and hierarchical multi-scale representations provide a strong advantage for the complex riparian scenes in WetVeg-2mm.

The differences are also clear in the Scratch setting. Without ImageNet pretraining, SegFormer-Scratch already achieves the highest *mIoU* (65.91%), *mDice* (77.67%), *PA* (80.62%) and *Recall* (78.90%) among the Scratch models, while PSPNet-Scratch obtains the highest *Precision* (78.36%). Both SegFormer-Scratch and PSPNet-Scratch clearly outperform U-Net, Attention U-Net and DeepLabV3+ trained from Scratch, suggesting that strong global context modeling is especially beneficial for millimeter-level riparian orthomosaics, where fragmented patches, mixed boundaries and high visual similarity among classes are common.

Pretraining improves all five architectures, but the magnitude of improvement varies substantially. The benefit is particularly pronounced for DeepLabV3+, whose *mIoU* and *mDice* increase by 34.49 and 33.15 percentage points, respectively. U-Net and Attention U-Net also show clear gains, whereas PSPNet and SegFormer improve more moderately because their Scratch versions are already comparatively strong. SegFormer-PT nevertheless remains the best-performing setting overall, showing that ImageNet-pretrained initialization further enhances an architecture that is already effective for this dataset. Taken together, these results show that WetVeg-2mm can support stable training with mainstream CNN-based and transformer-based segmentation models while still leaving clear room for further improvement.

4.2. Per-Class Performance Analysis

Table 7 reports the *IoU* values of all baseline models for each non-background class. The results reveal substantial variation across classes, confirming that WetVeg-2mm poses a clear fine-grained classification challenge. SegFormer-PT obtains the best *IoU* in 11 of the 16 non-background classes, including water (88.49%), Bambusa (77.52%), Aeschynomene (59.87%), Pongamia (80.77%), Bidens (71.74%), Colocasia (84.13%), Alternanthera (65.87%), Musa (89.40%), Diplazium (70.09%), Pueraria (91.09%) and Phragmites (81.41%). DeepLabV3+-PT remains the best model for Rhodomyrtus (69.30%), Syzygium (83.19%) and Eichhornia (89.92%), PSPNet-PT performs best for bareland (72.04%), and SegFormer-Scratch obtains the highest *IoU* for Arundo (94.06%).

Table 7. Per-class *IoU* (%) of baseline models on the test set.

Class	U-Net		Att. U-Net *		DeepLabV3+		PSPNet		SegFormer	
	Scratch	PT	Scratch	PT	Scratch	PT	Scratch	PT	Scratch	PT
Water	77.18	78.66	79.00	78.17	77.89	80.53	81.75	83.63	83.55	88.49
Bareland	61.37	63.15	63.44	65.95	48.78	66.86	63.47	72.04	70.06	65.12
Bambusa	0.00	37.11	29.26	36.56	34.13	76.48	58.93	75.24	57.48	77.52
Rhodomyrtus	0.00	0.07	0.00	5.05	0.00	69.30	50.17	64.19	55.28	56.63
Aeschynomene	0.00	48.12	23.10	49.45	22.20	42.24	51.16	57.73	48.99	59.87
Pongamia	8.24	44.56	21.23	64.63	47.52	80.04	67.76	74.20	71.58	80.77
Bidens	38.39	63.70	50.43	40.55	50.32	52.07	41.25	62.11	56.18	71.74
Syzygium	45.59	63.48	4.08	65.19	45.46	83.19	80.30	77.68	79.57	79.44
Eichhornia	70.36	82.70	76.20	75.05	62.88	89.92	88.46	88.61	85.33	89.37
Colocasia	42.56	75.41	60.33	49.45	30.57	78.43	45.99	49.05	73.03	84.13
Alternanthera	0.00	0.00	0.00	0.00	0.00	25.83	28.80	48.97	24.72	65.87
Musa	51.55	71.74	58.33	70.85	44.46	76.48	74.83	81.71	74.38	89.40
Diplazium	0.00	4.56	0.00	4.71	0.55	53.44	40.62	47.73	29.91	70.09
Arundo	23.01	48.91	0.01	65.57	5.59	93.41	92.03	87.89	94.06	87.94
Pueraria	43.08	35.01	0.00	0.07	33.37	81.22	87.50	78.47	87.76	91.09
Phragmites	0.00	0.95	0.00	30.75	49.86	72.82	64.72	75.31	76.92	81.41

* Att. U-Net = Attention U-Net; bold values indicate the best performance for each class.

Despite the strong overall performance of SegFormer-PT, several categories remain difficult for many models. *Rhodomyrtus*, *Aeschynomene*, *Alternanthera* and *Diplazium* show low or near-zero *IoU* values under several Scratch or weaker CNN settings, indicating that the available training signal is often insufficient for stable recognition without strong pretrained or contextual representations. These difficulties are likely related to limited sample coverage, small and fragmented patches and high visual similarity to neighboring herbaceous or shrub-like classes.

Pretraining substantially improves many difficult and long-tailed categories. For instance, DeepLabV3+-PT reaches 69.30% on *Rhodomyrtus* and 83.19% on *Syzygium*, while SegFormer-PT reaches 65.87% on *Alternanthera*, 70.09% on *Diplazium*, 91.09% on *Pueraria* and 81.41% on *Phragmites*. The strong performance of SegFormer-PT across many classes suggests that hierarchical token representations and long-range contextual modeling help distinguish visually similar vegetation types. However, the fact that DeepLabV3+-PT and PSPNet-PT remain best for several classes also shows that CNN-based multi-scale context aggregation is still competitive for specific category structures.

Figure 7 further shows that the five pretrained models exhibit different strengths across scene types rather than following a single fixed ranking in all cases. In the highly fragmented sample (Figure 7a) and the multi-class interwoven sample (Figure 7b), all models recover the main non-background regions to some extent, but their boundary stability and small-patch preservation differ. In the single-dominant-class sample with intra-class variation (Figure 7c) and the shadow-affected sample (Figure 7e), context-aware models such as DeepLabV3+, PSPNet and SegFormer generally produce smoother large-region predictions. By contrast, in the dense local-structure sample (Figure 7d), U-Net-like encoder–decoder models retain some fine local arrangements, whereas context-heavy models may merge or suppress small adjacent patches. In the challenging failure-case sample (Figure 7f), all models show clear limitations, and even SegFormer can be affected by dominant-class confusion and missing small targets.

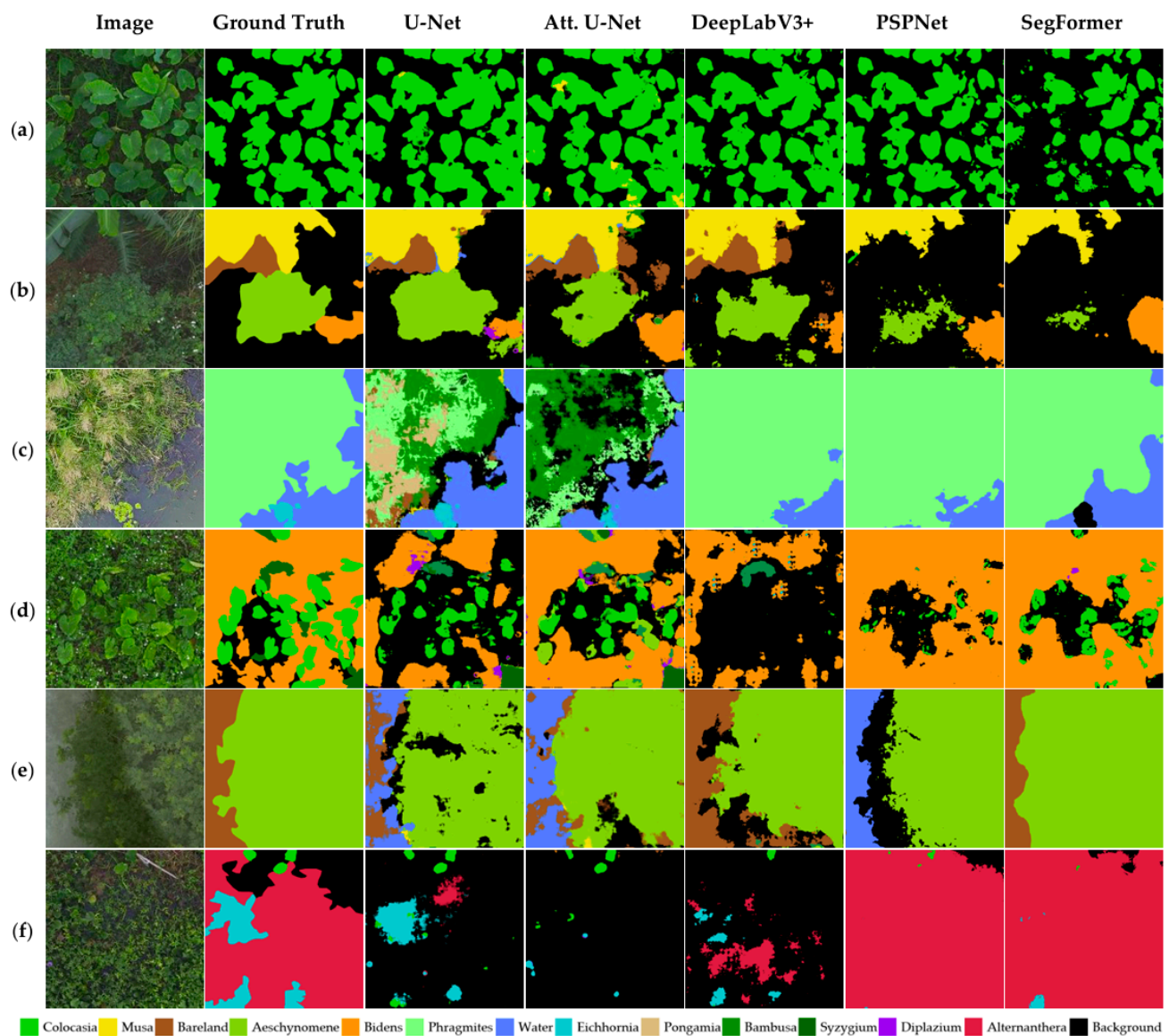


Figure 7. Qualitative comparison of pretrained baseline models on representative test samples from WetVeg-2mm. From left to right are the orthomosaic chip, the ground-truth mask and the predictions of U-Net, Attention U-Net, DeepLabV3+, PSPNet and SegFormer. The selected samples illustrate different scene characteristics, including (a) highly fragmented patches, (b) multi-class interwoven regions, (c) a single-dominant-class case with intra-class variation, (d) a dense local-structure sample, (e) a shadow-affected sample and (f) a challenging failure case. The class-color legend is shown at the bottom.

Taken together, the qualitative and quantitative results suggest that SegFormer-PT provides the strongest overall benchmark performance, but no architecture is uniformly optimal for every class or scene pattern. Different models still show different strengths and weaknesses depending on whether the dominant challenge lies in fragmented patches, local-structure preservation, intra-class variation, contextual dependence or degraded imaging conditions.

5. Discussion

5.1. Discussion of Benchmark Results

The benchmark results show that WetVeg-2mm is both usable and challenging as a fine-grained riparian vegetation segmentation dataset. On the one hand, all five baseline

architectures can be trained stably under a unified setting, and the best pretrained model, SegFormer-PT, reaches 76.23% *mIoU* and 86.01% *mDice*, indicating that the dataset provides effective supervision for semantic segmentation. On the other hand, the large differences across models, initialization strategies and classes show that the task is far from saturated. Significant room remains for improving rare-class recognition, boundary delineation and robustness in complex vegetation mosaics.

From the perspective of model characteristics, the results suggest that architectural suitability depends on both overall context modeling and scene type. SegFormer achieves the best overall benchmark performance, highlighting the value of long-range dependency modeling and hierarchical multi-scale token representations for fragmented and visually similar riparian vegetation. DeepLabV3+ and PSPNet remain strong CNN-based baselines for several classes, indicating that atrous convolution, pyramid pooling and multi-scale contextual aggregation are still important for complex riparian scenes. Meanwhile, the qualitative comparison shows that U-Net and Attention U-Net are not uniformly inferior: in some dense local-structure samples, direct encoder–decoder fusion can preserve local arrangements more faithfully. Therefore, the main difference among models lies not only in overall accuracy, but also in how each architecture adapts to different spatial patterns and visual conditions.

Another important finding is the strong role of pretraining. The clear improvements observed for all five models, especially DeepLabV3+, indicate that pretrained visual priors remain highly valuable even for millimeter-level UAV imagery [44,45]. At the same time, the strong Scratch performance of SegFormer and PSPNet suggests that architectural design can partly compensate for the lack of pretrained initialization when global context aggregation is effective. The remaining class-level variation further indicates that transferability depends on the interaction among class morphology, sample scale, texture variation and spatial context. In this sense, WetVeg-2mm is useful not only for reporting leaderboard-style results, but also for studying how different segmentation architectures respond to fine-grained ecological patterns.

5.2. Limitations and Future Work

Although WetVeg-2mm provides an ultra-high-resolution benchmark for fine-grained riparian vegetation segmentation, several limitations remain. First, the dataset is currently derived from one representative riparian section of the Jiuzhou River and one acquisition campaign, which means that geographic, seasonal and phenological variability are still limited. Second, the dataset shows clear class imbalance, with some classes appearing in relatively few chips or occupying only a small fraction of the labeled area. This makes rare-class learning particularly difficult and may affect the stability of class-wise evaluation.

Future work may expand the dataset to additional sites, seasons and growth stages, thereby improving its geographic coverage and temporal diversity. Future benchmark updates could evaluate foundation-model-based and promptable remote-sensing segmentation approaches, such as SAMRS and RingMo-SAM [46,47]. Hyperbolic-space semantic segmentation methods, such as HyperR3SNet and EAAHNet [48,49], are also worth exploring because they are well suited to modeling hierarchical, fine-grained and imbalanced semantic categories. In addition, future studies may explore class reweighting, long-tailed learning, boundary-aware optimization and multi-scale feature fusion to further improve performance on difficult and underrepresented classes [50–52].

6. Conclusions

This study presented WetVeg-2mm, an ultra-high-resolution UAV RGB dataset for fine-grained riparian vegetation semantic segmentation. The dataset contains 2054 image chips

of size 1024×1024 pixels at 2 mm GSD, covers 17 semantic classes, and provides pixel-level annotations, an official spatially grouped train/validation/test split, and a controlled public-release protocol. Benchmark experiments with U-Net, Attention U-Net, DeepLabV3+, PSPNet and SegFormer show that the dataset supports stable training under a unified setting while remaining challenging because of class imbalance, fragmented vegetation patches, fine boundaries, mixed water–land interfaces and high visual similarity among classes. SegFormer-PT achieves the strongest overall performance, but class-level and scene-level analyses indicate that no architecture is uniformly optimal across all categories and scene patterns. WetVeg-2mm therefore provides a reproducible data foundation for riparian vegetation mapping and future development of robust fine-grained segmentation methods.

Author Contributions: Conceptualization, H.Q., G.L. and R.Z.; methodology, G.L. and R.Z.; software, G.L. and R.Z.; validation, G.L.; formal analysis, G.L.; investigation, G.L. and R.Z.; resources, H.Q.; data curation, G.L.; writing—original draft preparation, G.L.; writing—review and editing, G.L.; visualization, G.L.; supervision, H.Q.; project administration, G.L.; funding acquisition, H.Q. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key R&D Program of China (No. 2023YFC3208801).

Data Availability Statement: The publicly released version of the WetVeg-2mm dataset will be made available on Zenodo upon publication of this paper (DOI: <https://doi.org/10.5281/zenodo.19849119>). The public release will include the image chips, the training and validation annotations, the class mapping file, the official split files, and the original polygon annotations for the publicly accessible subsets. To support controlled benchmark evaluation, the test semantic masks and polygon annotations will not be publicly released; evaluation on the test set can be requested from the authors. The benchmark code, official split files, and usage instructions will be made publicly available through GitHub at <https://github.com/sdsadsdd/WetVeg-2mm> (accessed on 5 July 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Han, Z.; Yin, D.; Zhao, R.; Liu, Z.; Cheng, R.; Lin, G.; Jia, H. The Ecological Effect of Ship Canals on Wetlands. *Water* **2024**, *16*, 3324. [CrossRef]
2. Wan, D.; Yang, J.; Liu, X.; Liu, Q.; Chen, B.; Chen, Y.; Zhao, D.; Liu, J. Identifying Freshwater Wetland Suitable Habitat through a Hydrological-Biological Connectivity Framework: A Case Study of Naoli River Wetlands, China. *Ecol. Indic.* **2025**, *170*, 112993. [CrossRef]
3. Naiman, R.J.; Décamps, H. The Ecology of Interfaces: Riparian Zones. *Annu. Rev. Ecol. Evol. Syst.* **1997**, *28*, 621–658. [CrossRef]
4. Ballanti, L.; Byrd, K.B.; Woo, I.; Ellings, C. Remote Sensing for Wetland Mapping and Historical Change Detection at the Nisqually River Delta. *Sustainability* **2017**, *9*, 1919. [CrossRef]
5. Taddeo, S.; Dronova, I. Indicators of Vegetation Development in Restored Wetlands. *Ecol. Indic.* **2018**, *94*, 454–467. [CrossRef]
6. Bair, E.H.; Dozier, J.; Rittger, K.; Stillinger, T.; Kleiber, W.; Davis, R.E. How Do Tradeoffs in Satellite Spatial and Temporal Resolution Impact Snow Water Equivalent Reconstruction? *Cryosphere* **2023**, *17*, 2629–2643. [CrossRef]
7. Wu, Q.; Lane, C.R.; Li, X.; Zhao, K.; Zhou, Y.; Clinton, N.; DeVries, B.; Golden, H.E.; Lang, M.W. Integrating LiDAR Data and Multi-Temporal Aerial Imagery to Map Wetland Inundation Dynamics Using Google Earth Engine. *Remote Sens. Environ.* **2019**, *228*, 1–13. [CrossRef] [PubMed]
8. Huylenbroeck, L.; Laslier, M.; Dufour, S.; Georges, B.; Lejeune, P.; Michez, A. Using Remote Sensing to Characterize Riparian Vegetation: A Review of Available Tools and Perspectives for Managers. *J. Environ. Manag.* **2020**, *267*, 110652. [CrossRef] [PubMed]
9. Guo, M.; Li, J.; Sheng, C.; Xu, J.; Wu, L. A Review of Wetland Remote Sensing. *Sensors* **2017**, *17*, 777. [CrossRef] [PubMed]
10. Sun, Z.; Wang, X.; Wang, Z.; Yang, L.; Xie, Y.; Huang, Y. UAVs as Remote Sensing Platforms in Plant Ecology: Review of Applications and Challenges. *J. Plant Ecol.* **2021**, *14*, 1003–1023. [CrossRef]
11. Morgan, G.R.; Hodgson, M.E.; Wang, C.; Schill, S.R. Unmanned Aerial Remote Sensing of Coastal Vegetation: A Review. *Ann. GIS* **2022**, *28*, 385–399. [CrossRef]

12. Zhou, Y.; Qiu, C.; Li, Y.; Wang, C.; Zhang, Y.; Huang, W.; Li, L.; Liu, H.; Zhang, D. Integrating UAV Data to Explore the Relationship between Microtopographic Variation and *Spartina Alterniflora* Expansion during Its Early Invasion. *Ecol. Indic.* **2023**, *154*, 110633. [\[CrossRef\]](#)
13. Anderson, K.; Gaston, K.J. Lightweight Unmanned Aerial Vehicles Will Revolutionize Spatial Ecology. *Front. Ecol. Environ.* **2013**, *11*, 138–146. [\[CrossRef\]](#)
14. Blaschke, T. Object Based Image Analysis for Remote Sensing. *ISPRS J. Photogramm. Remote Sens.* **2010**, *65*, 2–16. [\[CrossRef\]](#)
15. Mountrakis, G.; Im, J.; Ogole, C. Support Vector Machines in Remote Sensing: A Review. *ISPRS J. Photogramm. Remote Sens.* **2011**, *66*, 247–259. [\[CrossRef\]](#)
16. Ma, L.; Liu, Y.; Zhang, X.; Ye, Y.; Yin, G.; Johnson, B.A. Deep Learning in Remote Sensing Applications: A Meta-Analysis and Review. *ISPRS J. Photogramm. Remote Sens.* **2019**, *152*, 166–177. [\[CrossRef\]](#)
17. Yuan, Q.; Shen, H.; Li, T.; Li, Z.; Li, S.; Jiang, Y.; Xu, H.; Tan, W.; Yang, Q.; Wang, J.; et al. Deep Learning in Environmental Remote Sensing: Achievements and Challenges. *Remote Sens. Environ.* **2020**, *241*, 111716. [\[CrossRef\]](#)
18. Wang, D.; Zhuang, L.; Gao, L.; Sun, X.; Zhao, X. Global Feature-Injected Blind-Spot Network for Hyperspectral Anomaly Detection. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 5509305. [\[CrossRef\]](#)
19. Ying, L.; Miao, D.; Zhang, Z.; Zhang, H.; Pedrycz, W. Multi-Granularity Feature Enhancement Network for Maritime Ship Detection. *CAAI Trans. Intell. Technol.* **2024**, *9*, 649–664. [\[CrossRef\]](#)
20. Minaee, S.; Boykov, Y.; Porikli, F.; Plaza, A.; Kehtarnavaz, N.; Terzopoulos, D. Image Segmentation Using Deep Learning: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 3523–3542. [\[CrossRef\]](#) [\[PubMed\]](#)
21. Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In Proceedings of the Computer Vision—ECCV 2018: 15th European Conference, Munich, Germany, 8–14 September 2018; pp. 801–818.
22. Diakogiannis, F.I.; Waldner, F.; Caccetta, P.; Wu, C. ResUNet-a: A Deep Learning Framework for Semantic Segmentation of Remotely Sensed Data. *ISPRS J. Photogramm. Remote Sens.* **2020**, *162*, 94–114. [\[CrossRef\]](#)
23. Sa, I.; Popović, M.; Khanna, R.; Chen, Z.; Lottes, P.; Liebisch, F.; Nieto, J.; Stachniss, C.; Walter, A.; Siegwart, R. WeedMap: A Large-Scale Semantic Weed Mapping Framework Using Aerial Multispectral Imaging and Deep Neural Network for Precision Farming. *Remote Sens.* **2018**, *10*, 1423. [\[CrossRef\]](#)
24. Puliti, S.; Pearse, G.; Surový, P.; Wallace, L.; Hollaus, M.; Wielgosz, M.; Astrup, R. FOR-Instance: A UAV Laser Scanning Benchmark Dataset for Semantic and Instance Segmentation of Individual Trees. *arXiv* **2023**. [\[CrossRef\]](#)
25. Istiak, M.A.; Khan, R.H.; Rony, J.H.; Syeed, M.M.M.; Ashrafuzzaman, M.; Karim, M.R.; Hossain, M.S.; Uddin, M.F. AqUavplant Dataset: A High-Resolution Aquatic Plant Classification and Segmentation Image Dataset Using UAV. *Sci. Data* **2024**, *11*, 1411. [\[CrossRef\]](#) [\[PubMed\]](#)
26. Wei, Y.; Guo, Z.; Li, C.; Li, W.; Wang, S. SAM-Based Few-Shot Learning for Coastal Vegetation Segmentation in UAV Imagery via Cross-Matching and Self-Matching. *Remote Sens.* **2025**, *17*, 3404. [\[CrossRef\]](#)
27. Martínez-Movilla, A.; Román, M.; Fontenla-Carrera, G.; Rodríguez-Somoza, J.L.; Olabarria, C.; Martínez-Sánchez, J. Dataset of High-Resolution Aerial Images for Intertidal Macroalgae. *Sci. Data* **2025**, *12*, 1848. [\[CrossRef\]](#) [\[PubMed\]](#)
28. Li, S.; Huang, H.; Dai, Z. Research on Climate Change and Adaptation in the Beibu Gulf, Guangxi, over the Past 60 Years. *Ocean Dev. Manag.* **2017**, *34*, 50–55. [\[CrossRef\]](#)
29. Qinzhou Climate, Weather by Month, Average Temperature (China)—Weather Spark. Available online: <https://weatherspark.com/y/119990/Average-Weather-in-Qinzhou-China-Year-Round> (accessed on 2 July 2026).
30. Russell, B.C.; Torralba, A.; Murphy, K.P.; Freeman, W.T. LabelMe: A Database and Web-Based Tool for Image Annotation. *Int. J. Comput. Vis.* **2008**, *77*, 157–173. [\[CrossRef\]](#)
31. Valavi, R.; Elith, J.; Lahoz-Monfort, J.J.; Guillera-Aroita, G. blockCV: An r Package for Generating Spatially or Environmentally Separated Folds for k-Fold Cross-Validation of Species Distribution Models. *Methods Ecol. Evol.* **2019**, *10*, 225–232. [\[CrossRef\]](#)
32. Kattenborn, T.; Schiefer, F.; Frey, J.; Feilhauer, H.; Mahecha, M.D.; Dormann, C.F. Spatially Autocorrelated Training and Validation Samples Inflate Performance Assessment of Convolutional Neural Networks. *ISPRS Open J. Photogramm. Remote Sens.* **2022**, *5*, 100018. [\[CrossRef\]](#)
33. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*; Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F., Eds.; Springer International Publishing: Cham, Switzerland, 2015; pp. 234–241.
34. Oktay, O.; Schlemper, J.; Folgoc, L.L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N.Y.; Kainz, B.; et al. Attention U-Net: Learning Where to Look for the Pancreas. *arXiv* **2018**. [\[CrossRef\]](#)
35. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 834–848. [\[CrossRef\]](#) [\[PubMed\]](#)

36. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid Scene Parsing Network. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
37. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. *Adv. Neural Inf. Process Syst.* **2021**, *34*, 12077–12090.
38. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [[CrossRef](#)]
39. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2019; Volume 32.
40. Csurka, G.; Larlus, D.; Perronnin, F. What Is a Good Evaluation Measure for Semantic Segmentation? In *Proceedings of the British Machine Vision Conference 2013*; British Machine Vision Association: Bristol, UK, 2013; pp. 32.1–32.11.
41. Everingham, M.; Van Gool, L.; Williams, C.; Winn, J.; Zisserman, A. The Pascal Visual Object Classes (VOC) Challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [[CrossRef](#)]
42. Dice, L.R. Measures of the Amount of Ecologic Association Between Species. *Ecology* **1945**, *26*, 297–302. [[CrossRef](#)]
43. Sokolova, M.; Lapalme, G. A Systematic Analysis of Performance Measures for Classification Tasks. *Inf. Process Manag.* **2009**, *45*, 427–437. [[CrossRef](#)]
44. Yosinski, J.; Clune, J.; Bengio, Y.; Lipson, H. How Transferable Are Features in Deep Neural Networks? In Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS 2014), Montreal, QC, Canada, 8–13 December 2014; pp. 3320–3328.
45. Penatti, O.; Nogueira, K.; dos Santos, J. Do Deep Features Generalize from Everyday Objects to Remote Sensing and Aerial Scenes Domains? In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Boston, MA, USA, 7–12 June 2015; pp. 44–51.
46. Wang, D.; Zhang, J.; Du, B.; Xu, M.; Liu, L.; Tao, D.; Zhang, L. SAMRS: Scaling-up Remote Sensing Segmentation Dataset with Segment Anything Model. *arXiv* **2023**. [[CrossRef](#)]
47. Yan, Z.; Li, J.; Li, X.; Zhou, R.; Zhang, W.; Feng, Y.; Diao, W.; Fu, K.; Sun, X. RingMo-SAM: A Foundation Model for Segment Anything in Multimodal Remote-Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5625716. [[CrossRef](#)]
48. Fu, J.; Wang, C.; Liu, M.; Li, X.; Liu, Y.; Shi, W.; Wang, R. HyperR3SNet: Leveraging Hyperbolic Space and Vision Foundation Models for Remote Sensing Semantic Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2026**, *64*, 5620016. [[CrossRef](#)]
49. Li, X.; Xu, F.; Liu, F.; Lyu, X.; Gao, H.; Zhou, J.; Kaup, A. A Euclidean Affinity-Augmented Hyperbolic Neural Network for Semantic Segmentation of Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5636718. [[CrossRef](#)]
50. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollar, P. Focal Loss for Dense Object Detection. In Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence, Los Alamitos, CA, USA, 23 July 2018; pp. 2980–2988.
51. Cui, Y.; Jia, M.; Lin, T.-Y.; Song, Y.; Belongie, S. Class-Balanced Loss Based on Effective Number of Samples. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 9268–9277.
52. Kervadec, H.; Bouchtiba, J.; Desrosiers, C.; Granger, E.; Dolz, J.; Ayed, I.B. Boundary Loss for Highly Unbalanced Segmentation. In Proceedings of the 2nd International Conference on Medical Imaging with Deep Learning, London, UK, 8–10 July 2019; pp. 285–296.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.