remote sensing

MDPI

Article

RLGD: A High-Resolution UAV Dataset and Benchmark for Semantic Segmentation and Revetment Extraction in Waterway Scenes

Aiguo Sun [1,2], Rongjun Xiong [1,3], Rui Zhang [1], Shuang Yang [1,*], Ziwei Zhu [3], Dongsheng Li [3], Tao Xu [1], Yuyao Jiang [1], Zeqiang Chen [1] and Nengcheng Chen [1]

[1] National Engineering Research Center for Geographic Information System, China University of Geosciences, Wuhan 430078, China; sunaiguo@cug.edu.cn (A.S.); xiongrongjun@cug.edu.cn (R.X.); 1202322013@cug.edu.cn (R.Z.); tao.xu@cug.edu.cn (T.X.); yuyaojiang@cug.edu.cn (Y.J.); chenzeqiang@cug.edu.cn (Z.C.); chennengcheng@cug.edu.cn (N.C.)
[2] Changjiang Waterway Institute of Planning and Design, Wuhan 430040, China
[3] Changhang Testing Technology (Wuhan) Co., Ltd., Wuhan 430040, China; cz18171300875@163.com (Z.Z.); lds10000@163.com (D.L.)
* Correspondence: yangshuang@cug.edu.cn

Highlights

What are the main findings?

- RLGD contains 100 high-resolution UAV images (3000 × 4000 pixels), 1.2 billion labeled pixels, and a released eight-class semantic taxonomy tailored to waterway scenes.
- A multi-model benchmark and VDD-to-RLGD transfer experiment reveal a substantial domain gap between general UAV imagery and specialized waterway scenes.

What are the implications of the main findings?

- RLGD supports waterway-scene evaluation through in-domain segmentation, cross-dataset transfer, boundary-sensitive revetment assessment, and visual-completeness screening.
- SAM-based refinement raises revetment IoU from 84.10% to 87.25%, and geometry-based screening attains 81.25% balanced accuracy with expert visual interpretation.

Abstract

High-resolution unmanned aerial vehicle (UAV) imagery is increasingly used for waterway inspection, but automated scene understanding remains constrained by the lack of pixel-level datasets that explicitly represent revetments and their complex interfaces with water, vegetation, roads, and riverside structures. We introduce the Riverway and Landscape Ground Dataset (RLGD), a dedicated UAV semantic segmentation dataset for waterway scenes. RLGD contains 100 images at 3000 × 4000 pixels, corresponding to 1.2 billion labeled pixels. Its public RLGD taxonomy explicitly includes revetment. Seven semantic segmentation models are evaluated under a unified protocol. Across five training runs with different random seeds, mean mIoU ranges from 69.27% to 77.02%, while Hierarchical Joint Network with Boundary Guidance (GHJNet) achieves 77.02 ± 0.41% mIoU and 83.51 ± 1.18% revetment IoU. Direct transfer from Varied Drone Dataset (VDD) to RLGD results in relative mIoU reductions of 11.29–49.81%, whereas adaptation using 20% of the RLGD training source images closes 65.5% of the measured domain gap. For the GHJNet run used in the downstream analysis, object-prior refinement using the Segment Anything Model (SAM) increases revetment IoU from 84.10% to 87.25%. A 42-image visual-completeness screening case study further achieves 83.93% macro-F1 and 81.25%

Check for updates

Academic Editor: Zhenwei Shi

Received: 21 July 2026
Revised: 4 October 2026
Accepted: 18 October 2026
Published: 24 October 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) license.

balanced accuracy against image-level interpretations by three experts. Overall, RLGD provides a densely annotated, task-specific benchmark for dry-season waterway semantic segmentation, revetment extraction, and image-level inspection prioritization within the surveyed Yangtze River corridor.

Keywords: UAV imagery; waterway scenes; semantic segmentation dataset; revetment extraction; infrastructure inspection

1. Introduction

Revetments are engineered bank-protection structures that support riverbank stability and protect adjacent infrastructure. Their spatial distribution and visible condition therefore require periodic inspection for maintenance planning [1,2]. Conventional field inspection is labor-intensive and provides limited spatial continuity along long and heterogeneous river corridors, whereas unmanned aerial vehicle (UAV) photogrammetry and visible–infrared sensing can support spatially continuous inspection of revetment surfaces and embankments [1,3]. These capabilities are also relevant to the continuing development of inland waterway navigation infrastructure [4].

UAV remote sensing offers centimeter-level spatial detail, flexible deployment, and oblique or nadir views that expose both riverbank structures and their surrounding context [5,6]. Semantic segmentation can convert these images into spatially explicit land-cover maps, but model development depends on datasets that reflect the target environment. Models trained primarily on urban or general aerial imagery may not represent the narrow, elongated, partially occluded, and water-adjacent structures that characterize revetments.

Existing datasets provide useful but incomplete precedents. UAVid focuses on urban street scenes [7], whereas the Varied Drone Dataset (VDD) expands variation in viewpoint, location, altitude, and scene content [8]. The Aerial Fluvial Image Dataset (AFID) provides river-oriented imagery and a multi-model benchmark, but its label taxonomy is designed for general fluvial land cover rather than engineered revetments [9]. The High-Resolution SAR Images Dataset (HRSID) addresses a different Synthetic Aperture Radar (SAR) ship-analysis task, yet its use of overlapping image crops, Common Objects in Context (COCO-style) annotations, and target-scale statistics provides useful precedents for specialized remote-sensing dataset design [10]. None of these resources provides a dedicated pixel-level revetment class together with the adjacent waterway context required for revetment extraction.

Reviews of UAV semantic segmentation summarize variation in platforms, viewpoints, regions, and acquisition conditions [11]. Ensemble knowledge transfer addresses shifts in scene structure, viewpoint, and object statistics by transferring knowledge from multiple source domains to aerial semantic segmentation [12]. LoveDA, OpenEarthMap, and FLAIR explicitly address geographic or domain diversity [13–15], while iSAID and SkyScapes provide detailed high-resolution object or scene annotations [16,17]. FloodNet demonstrates task-oriented UAV annotation for a specialized environmental event [18], and LandCover.ai and DeepGlobe provide reproducible land-cover benchmarks at larger spatial scales [19,20]. SAMRS demonstrates foundation-model-assisted scaling of remote-sensing segmentation resources through an automated mask-generation pipeline [21]. These studies motivate RLGD's emphasis on documented taxonomy, pixel-level statistics, fixed benchmark protocols, and explicit cross-dataset evaluation.

RLGD complements VDD and AFID by introducing a dedicated revetment label and preserving the adjacent waterway context required for engineering-object extraction (Table 1). The public dataset is distributed under the eight-class taxonomy.

Table 1. Comparison of the principal UAV semantic segmentation datasets.

Dataset	Images	Released Benchmark Classes	Primary Scene	View Diversity	Dedicated Revetment Label	Role
VDD [8]	400	7	Varied general UAV scenes	30°/60°/90°	No	Source-domain training and transfer
AFID [9]	816	8	Two fluvial systems	Nadir and oblique	No	Related river-oriented reference
RLGD	100	8	Yangtze waterway corridor	Nadir and oblique	Yes	Released in-domain benchmark and downstream tasks

A second challenge concerns evaluation. Rigorous dataset evaluation requires a transparent task definition, reproducible segmentation models, domain-transfer evidence, and representative downstream analyses. Modern mask-based segmentation frameworks broaden the choice of benchmark architectures [22], while dense CRF refinement, boundary-aware losses, and point-based rendering illustrate complementary ways to improve local detail [23–25]. In waterway scenes, region-level mIoU alone is insufficient because a small displacement along a long revetment boundary can materially change the extracted object even when the overall pixel score remains high. IoU-oriented optimization is therefore a useful complement to conventional region metrics [26]. Region, boundary, efficiency, transfer, and downstream comparison measures are defined in Section 2.2.4.

This study presents a dataset–benchmark–application framework. The contributions are as follows:

First, we construct RLGD, a high-resolution UAV dataset dedicated to waterway scene segmentation. The dataset contains 100 images and a released eight-class taxonomy with an explicit revetment category, together with annotation rules and statistical analyses of class frequency, object area, and spectral similarity.

Second, we establish two complementary evaluation protocols. An eight-class in-domain benchmark evaluates seven semantic segmentation architectures, while a seven-class VDD-to-RLGD protocol quantifies transfer degradation under a shared taxonomy. Five-run mIoU and revetment IoU summaries, boundary accuracy, and computational efficiency provide complementary measures of segmentation performance.

Third, we demonstrate two downstream uses of the dataset. Object-prior refinement using the Segment Anything Model (SAM) evaluates whether instance-level masks improve revetment boundaries produced by semantic segmentation. The refined masks are subsequently characterized through connectivity and geometry for visual-completeness screening. These categories support image-level inspection prioritization.

The remainder of the paper is organized as follows. Section 2 describes RLGD, the benchmark protocols, object-prior refinement, and visual-completeness screening. Section 3 reports the in-domain benchmark, cross-dataset transfer, revetment-refinement results, and visual-completeness analysis. Section 4 discusses the benchmark results, domain transfer, object-prior refinement, and current limitations. Section 5 concludes the study.

2. Materials and Methods

This section describes the study area and data acquisition, image-quality screening, the RLGD taxonomy and annotation protocol, benchmark design, object-prior refinement, and the visual-completeness screening case study.

2.1. RLGD Construction

2.1.1. Study Area and Data Acquisition

The imagery is acquired during the dry season (low-water period) along a representative Yangtze River waterway corridor (Jingjiang Section) in Hubei Province, China (Figure 1). The available collection therefore characterizes one corridor under dry-season hydrological and vegetation conditions rather than year-round variability. The study area contains urban built-up land, rural riparian zones, vegetation, roads, water, ships, and multiple forms of engineered revetment. These co-occurring classes create visually complex land–water interfaces in which revetments may be narrow, discontinuous, or partially covered by vegetation.

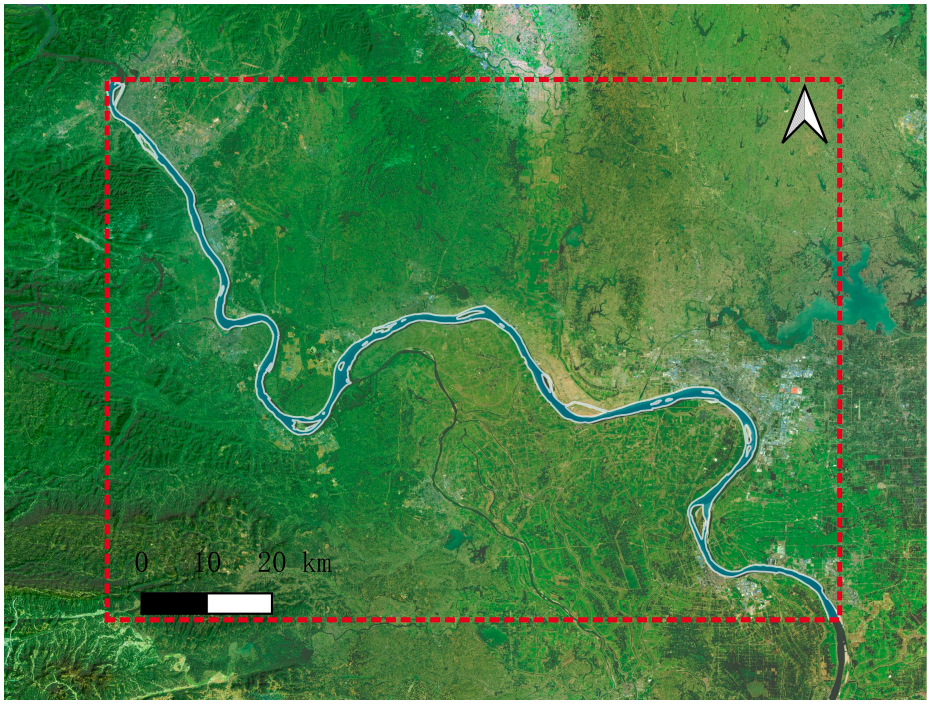


Figure 1. Study area and the scope of the waterway improvement corridor. The red dashed box indicates the approximate extent of the study area along the Jingjiang section of the Yangtze River in Hubei Province, China.

Images are collected using a DJI Mavic 3 Enterprise UAV (SZ DJI Technology Co., Ltd., Shenzhen, China) equipped with a 4/3 complementary metal-oxide-semiconductor (CMOS) wide-angle camera and a 12 mm fixed-focal-length lens (24 mm full-frame equivalent), as shown in Table 2. Nominal flights are conducted at 100 to 120 m above ground level; selected visually complex or occluded revetment areas are additionally observed at approximately 60 m. The ground sampling distance at 100 m is approximately 3 cm. The nominal forward and side overlaps are 80% and 70%, respectively, and the flight speed is 6 to 8 m/s. Flights follow strip-like paths along both riverbanks and cover approximately 50 m inland from each bank. Source images are standardized to 3000 × 4000 pixels for annotation and model development.

Table 2. UAV and image-acquisition settings used to construct RLGD.

Category	Parameter	Setting	Purpose
Platform	UAV	DJI Mavic 3 Enterprise	Industrial multirotor acquisition
Camera	Sensor/lens	4/3 CMOS; 12 mm (24 mm equivalent)	High-resolution visible imagery
Flight	Nominal altitude	100–120 m	Corridor-scale coverage
	Supplementary altitude	Approximately 60 m	Occluded or visually complex revetments
	Forward/side overlap	80%/70%	Continuous coverage
	Speed	6–8 m/s	Balance efficiency and image quality
Coverage	Riverbank corridor	Approximately 50 m inland per bank	Include surrounding scene context
Environment	Acquisition time	09:00–18:00; front or side-front lighting	Reduce severe backlighting

More than one thousand source frames are screened before annotation. Images with severe overexposure, insufficient scene content, or weak object interpretability are removed, and 100 representative images are retained. Figure 2 shows examples of an accepted image, an overexposed image, and an invalid low-information image.



(a) (b) (c)

Figure 2. Examples from image-quality screening: (**a**) accepted sample; (**b**) overexposed sample; and (**c**) invalid low-information sample.

The selection objective is to construct a controlled, fully annotated benchmark for revetment-centered waterway interpretation, rather than to retain every frame acquired along a highly overlapping flight path. Screening is performed at the source-image level before annotation, and it removes frames that cannot support reliable class or boundary interpretation. The resulting 100 images retain the high spatial detail needed to delineate narrow revetments and adjacent classes. RLGD should therefore be understood as a dense, task-specific benchmark within one surveyed corridor.

2.1.2. Semantic Class Definition and Annotation Protocol

Table 3 defines the eight classes of the released RLGD: other, building, road, vegetation, vehicle, water, ship, and revetment. During annotation, visible building walls and roofs are initially distinguished to support manual boundary interpretation under different UAV viewing geometries. These labels are merged into building before release and in-domain evaluation. The finer nine-label scheme, RLGD-9, denotes the internal annotation stage.

For annotation, revetment is operationally defined in this study as the visible above-water portion of an engineered bank-protection structure, including gravity-type structures made of masonry or concrete, slope-type structures such as ecological blocks and riprap, and other engineered forms. The engineering terminology is informed by the Construction Code of Channel Regulation Works (JTS 224-2016) [27]. Naturally exposed soil, beaches, and unmanaged vegetation are not labeled as revetment.

Table 3. Released RLGD semantic classes and annotation definitions.

Class	Definition and Boundary Convention	Source
Revetment	Visible above-water engineered bank-protection structures, including masonry, concrete, ecological blocks, and riprap	Engineering terminology informed by JTS 224-2016 [27]
Building	Visible building surfaces, including walls and roofs	Adapted from VDD [8] and consolidated in this study
Road	Paved vehicle-access surfaces, including crest and wharf-front roads	Adapted from VDD [8]
Vegetation	Trees, shrubs, grass, and managed green areas	Adapted from VDD [8]
Vehicle	Motorized and non-motorized land vehicles	Adapted from VDD [8]
Water	Navigable channels, rivers, ponds, or other visible water surfaces	Adapted from VDD [8]
Ship	Navigating or berthed vessels above the visible waterline	Defined in this study
Other	Regions not covered by the named class definitions	Adapted from VDD [8]

Ambiguous boundaries are resolved using both material transitions and object geometry. A clear curb, pavement change, or structural interface defines the revetment–road boundary. When no sharp interface is visible, roads are interpreted as relatively planar traffic surfaces, whereas revetments follow the riverbank slope or the land–water edge. Under partial vegetation cover, only visually supported structural regions are labeled; annotators do not extrapolate an occluded revetment through vegetation without visible evidence (Figure 3).

Annotation follows a three-stage workflow. A model pretrained on VDD [8] first generates coarse masks for common classes. One annotator then corrects class identities and boundaries at the native image resolution in LabelMe, with particular attention to revetment, ship, wall, roof, and narrow transition zones. A second annotator reviews the masks, and disagreements are resolved by joint visual inspection. The annotators' research focuses on UAV image segmentation in waterway scenes. Before annotation, they receive training and technical guidance through exchanges with staff of the collaborating institute, who have extensive experience in waterway research. The annotators are familiar with the visual characteristics of typical waterway objects. This consensus workflow reduces repetitive labeling while retaining manual control over all final masks. The release contains no separate ignore class: pixels outside the named class definitions are assigned to other classes, while vegetation-occluded revetment is not extrapolated beyond visible evidence.

To provide a quantitative check distinct from this consensus review, a representative subset of 12 full-resolution source images is independently annotated a second time from the RGB imagery without access to the first masks. The independent 12-image subset is purposefully selected to cover revetment-road boundaries, vegetation-obscured structures, ships, and vehicles. The unadjudicated masks are compared at 3000×4000 resolution using pixel agreement, class-wise IoU, eight-class mIoU, and a revetment boundary F-score evaluated over different tolerances. This repeat annotation audit complements, but does not replace, the original second-person review and joint adjudication.

2.1.3. Dataset Characteristics and Statistical Analysis

The dataset split used in the experiments contains 70 training, 20 validation, and 10 test images (Table 4). Because each standardized image contains 12 million pixels, the three subsets contain 840, 240, and 120 million labeled pixels, respectively. Revetment accounts for 8.91%, 8.46%, and 5.89% of the pixels in the three subsets, and appears in

74.29%, 75.00%, and 70.00% of the images. For N_S in a data subset S, pixel entropy is calculated as $E_S = \frac{1}{N_S}\sum_{i\in S} E_i$, where E_i is the Entropy of image i. It is calculated as $E_i = -\sum_{c=1}^{8} \frac{n_{i,c}}{H_i W_i} \ln\left(\frac{n_{i,c}}{H_i W_i}\right)$, where $n_{i,c}$ is the number of pixels in the image labeled as class c. Table 4 reports the E_S over the source images in each subset.

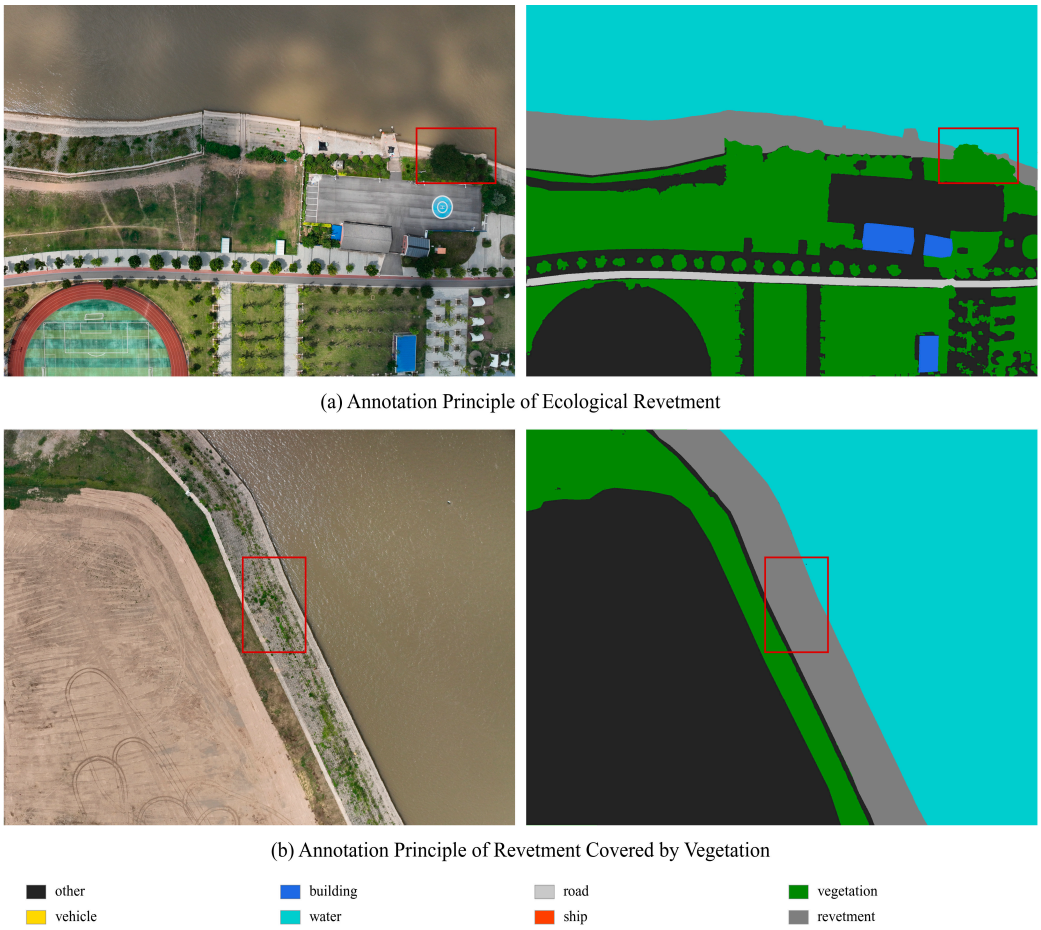


Figure 3. Example of revetment annotation under partial vegetation occlusion. The annotation-stage class colors are: revetment, gray; road, light gray; vegetation, green; vehicle, yellow; building, blue; water, cyan; ship, orange-red; and other, dark gray. The red boxes highlight revetment areas with vegetation cover and their corresponding semantic annotations.

Table 4. Summary statistics of the RLGD splits.

Split	Images	Total Pixels	Mean Pixel Entropy	Mean Classes/Image	Revetment Pixels (%)	Images with Revetment (%)
Train	70	840,000,000	1.10	5.56	8.91	74.29
Validation	20	240,000,000	1.18	5.85	8.46	75.00
Test	10	120,000,000	1.11	6.20	5.89	70.00

RLGD contains 1.2 billion labeled pixels. Vegetation is the dominant category (43.0% of all pixels), followed by water (27.3%), other (15.1%), and revetment (8.5%). Building, road, vehicle, and ship account for 2.1%, 2.8%, 0.1%, and 1.0% of the pixels, respectively. This distribution reflects practical waterway imagery, in which broad natural backgrounds surround comparatively sparse engineered and mobile objects (Figures 4 and 5).

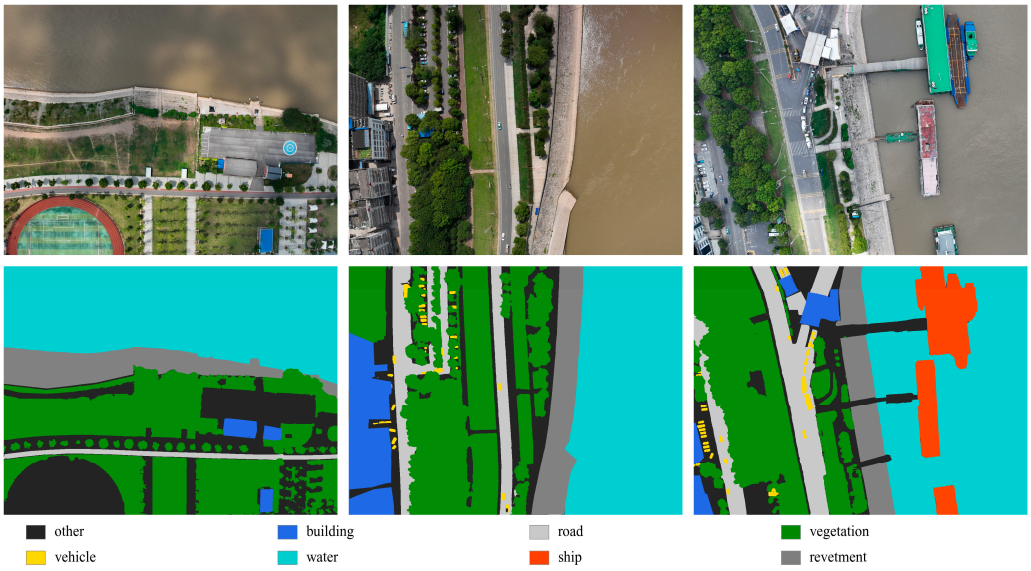


Figure 4. Representative RLGD images and pixel-level annotations. The annotation-stage class colors are: revetment, gray; road, light gray; vegetation, green; vehicle, yellow; building, blue; water, cyan; ship, orange-red; and other, dark gray.

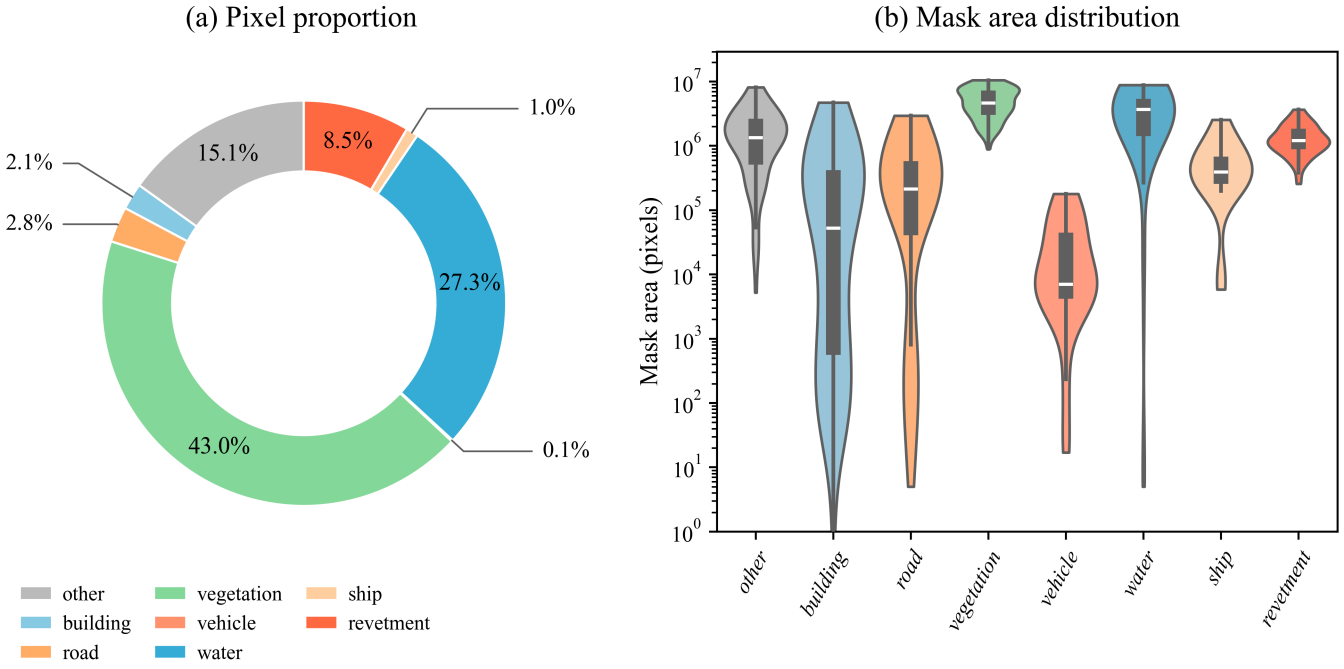


Figure 5. Pixel-level class distribution and object-area characteristics of RLGD.

The long-tailed distribution makes overall accuracy an inadequate standalone measure: a model may perform well on vegetation and water while failing on vehicles, ships, or revetments. The principal multi-run benchmark therefore reports mIoU together with the application-critical revetment IoU, while boundary measures further characterize revetment delineation.

Per-image class-mask areas also vary substantially. Water and vegetation generally occupy large parts of source images, whereas vehicle, ship, and building class-mask areas are comparatively small. Revetment class-mask areas span a broad range because visible bank-protection structures may cover long strips, fragmented local regions, or partially visible areas under vegetation. These statistics describe class occupancy within individual source images and motivate multiscale segmentation baselines.

The class-wise RGB statistics in Figure 6a indicate that color cues alone cannot reliably separate revetment from road and heterogeneous background surfaces. Revetment appearance changes with construction material, weathering, moisture, illumination, and viewing direction, while adjacent roads and exposed surfaces may show similar visible-spectrum responses.

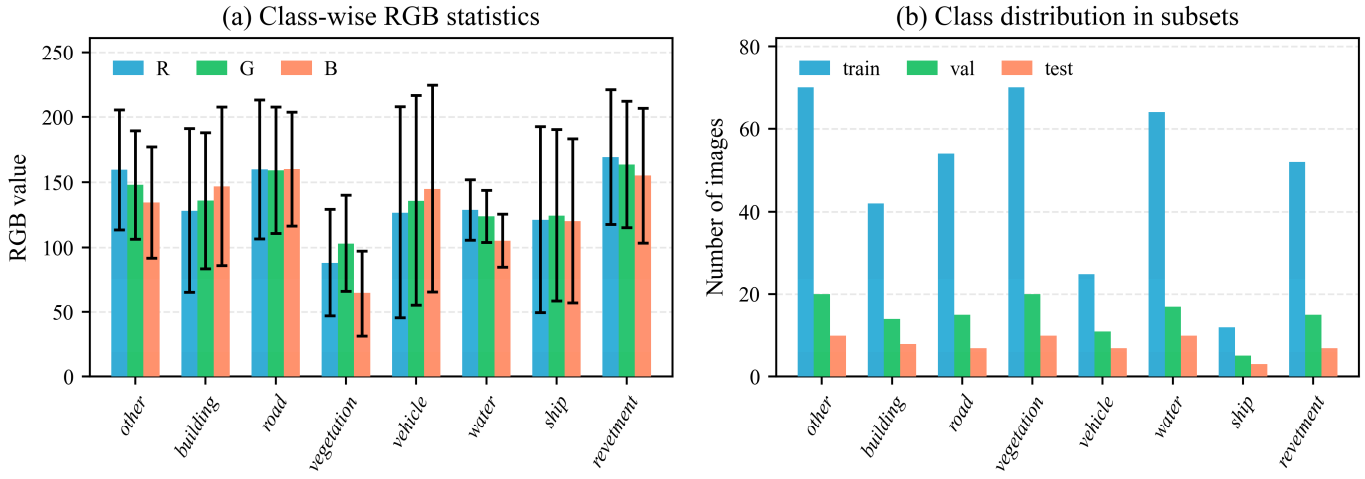


Figure 6. Class-wise RGB statistics and image-level category occurrence across the fixed subsets.

Image-level occurrence statistics (Figure 6b) show that the dominant natural classes are present throughout the split and that revetment occurs in approximately 70–75% of the images. The smaller test-set pixel proportion of revetment reflects scene composition rather than the absence of the target class.

Overall, RLGD combines high spatial resolution, class imbalance, scale variation, spectral overlap, and complex land–water adjacency. These properties make the dataset relevant to both general waterway scene understanding and the specialized extraction of engineered bank-protection structures.

To quantify the spectral separability between revetment and adjacent classes, we compute the Bhattacharyya distance (BD) [28] between class-wise RGB distributions. For two classes i and j with mean vectors μ_i, μ_j and covariance matrices Σ_i, Σ_j, the BD is defined as

$$BD_{ij} = \frac{1}{8}(\mu_i - \mu_j)^{\mathrm{T}}\Sigma^{-1}(\mu_i - \mu_j) + \frac{1}{2}\ln\left(\frac{|\Sigma|}{\sqrt{|\Sigma_i||\Sigma_j|}}\right), \Sigma = \frac{\Sigma_i + \Sigma_j}{2} \quad (1)$$

The Bhattacharyya distances in Table 5 confirm that revetment is spectrally close to other surfaces (0.17) and road (0.25), while it is more distinct from water (0.45) and vegetation (1.11). Consequently, local RGB appearance is insufficient by itself; reliable segmentation also requires spatial context, multiscale structure, and land–water relationships.

Table 5. Bhattacharyya distance matrix between semantic classes.

Class	Revetment	Road	Other	Vegetation	Water
Revetment	/	0.25	0.17	1.11	0.45
Road	0.25	/	0.42	2.08	1.14
Other	0.17	0.42	/	1.16	0.39
Vegetation	1.11	2.08	1.16	/	0.92
Water	0.45	1.14	0.39	0.92	/

2.2. Benchmark Design

2.2.1. Evaluation Tasks and Label Mappings

RLGD uses one public evaluation mapping derived from the annotation workflow. It consolidates the annotation-stage wall and roof labels into building. RLGD-9 refers only to the finer annotation-stage taxonomy and is not released or presented as a separate public benchmark.

2.2.2. Benchmark Models

The benchmark is designed to cover complementary feature-extraction and decoding paradigms rather than minor variants of a single architecture. Six publicly available models are selected to represent convolutional encoder–decoders, pyramid context aggregation, multilevel feature fusion, object-context modeling, hierarchical Transformers, and modern convolutional attention. GHJNet is evaluated as a seventh architecture using the same RLGD split and optimization schedule, which is an internally developed, task-specific reference model included to examine boundary guidance on RLGD.

DeepLabV3+ [29] combines atrous spatial pyramid pooling with an encoder–decoder structure; the benchmark uses a ResNet-50 backbone and an output stride of 16 to preserve spatial detail while aggregating multiscale context. PSPNet [30] uses pyramid pooling at several spatial scales to combine local features with a global scene prior and is also implemented with ResNet-50. UPerNet [31] couples pyramid pooling at the highest feature level with top-down feature-pyramid fusion and lateral connections, allowing semantic information from deep layers to be combined with spatial detail from shallower layers. OCRNet [32] first estimates soft object regions, aggregates class-level object representations, and then redistributes the resulting object context to individual pixels. The benchmark uses HRNet-18 to retain high-resolution features for small objects and fragmented boundaries.

SegFormer-B0 [33] employs a hierarchical Mix Transformer encoder with overlapping patch embedding and a lightweight multilayer perceptron decoder that aligns and fuses multiscale features. The B0 variant is used in the in-domain benchmark, while B1 is additionally evaluated in the cross-dataset experiment to examine the effect of model scale. SegNeXt [34] returns to a convolutional design and uses multiscale convolutional attention, implemented with efficient depthwise convolutions of different receptive fields, to encode contextual information while retaining the spatial inductive bias of convolution.

The GHJNet is an internally developed, task-oriented reference architecture for the RLGD benchmark and the downstream revetment case study. As illustrated in Figure 7, its SegFormer-B0 encoder produces four feature levels. The integration I branch aligns and fuses all four levels into a multiscale semantic representation, whereas the detail (D) branch combines the first three levels to recover local structure and boundary cues. The detail pathway uses depth-wise-separable and local convolutions, horizontal and vertical strip operations, and directional guidance with criss-cross contextual interaction. It predicts a single-channel boundary logit. After sigmoid activation, the resulting guidance map F_D multiplicatively enhances the main semantic representation, followed by convolutional refinement and multiclass prediction. GHJNet is assessed here as a complete configuration; the benchmark does not attribute its performance to an individual decoder component.

2.2.3. Training and Evaluation Protocol

All models use the same source-image split, input crop, data augmentation, iteration budget, optimizer family, and validation schedule. The RLGD benchmark within the domain utilizes the results of five runs with different random seeds. Cross-dataset experiments train on VDD and evaluate directly on the RLGD common-label subset.

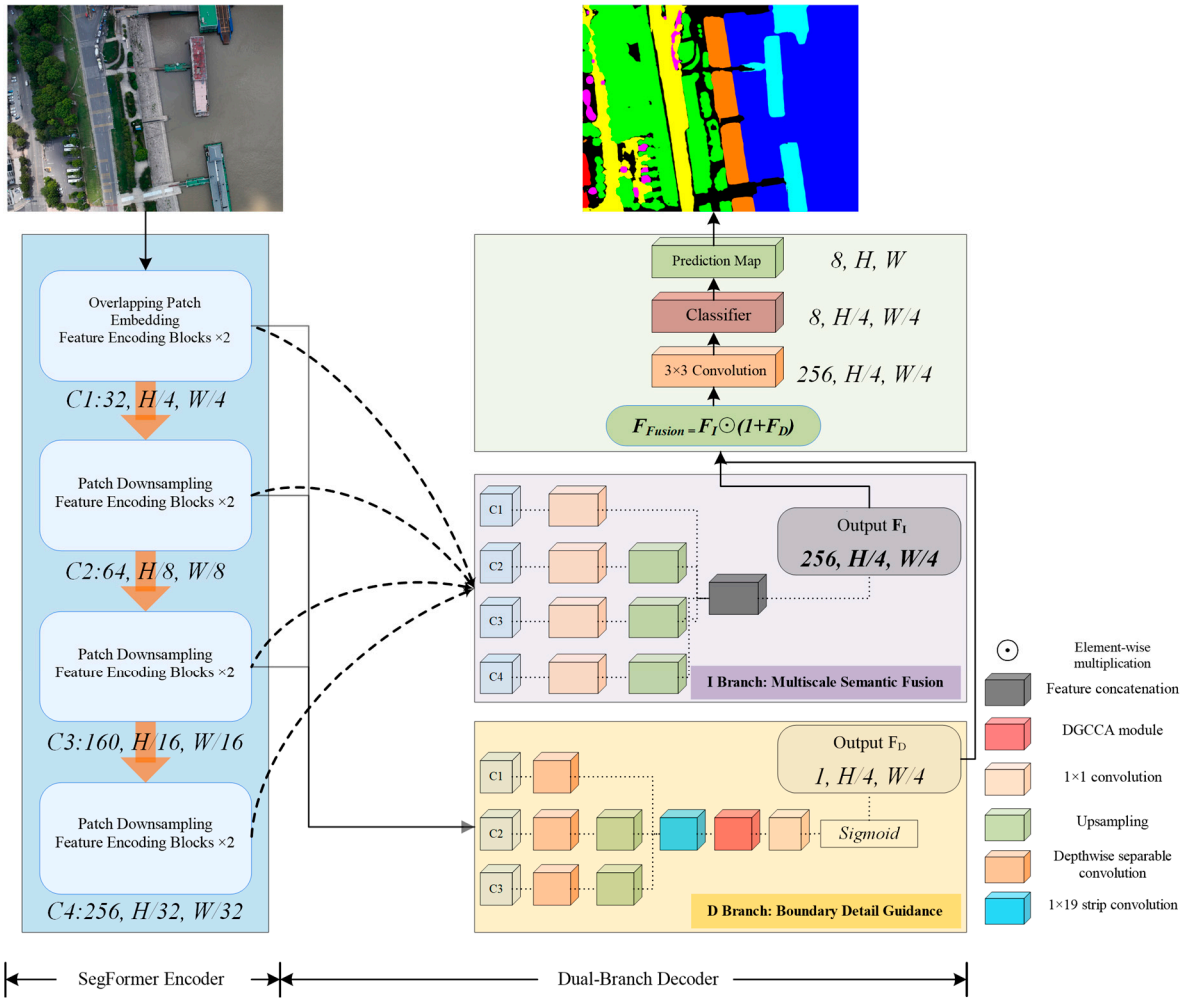


Figure 7. Schematic architecture of GHJNet. The I branch fuses four SegFormer-B0 feature levels, and the D branch uses the first three levels to produce boundary guidance for the final semantic prediction. The F_D denotes the sigmoid-activated guidance map. DGCCA denotes the directional-guidance module with criss-cross contextual interaction.

The 70/20/10 split is performed at the source-image level before patch extraction, so all 12 tiles from a source image remain in one subset and source-ID overlap across subsets is zero. To assess residual cross-subset similarity, we perform a post hoc audit on all source images in the training and test sets. Exact duplicate detection, perceptual hashing, a frozen ImageNet-pretrained Swin-T cosine representation, and SIFT matching with RANSAC homography are used; geometric overlap is treated as the primary evidence, and failed homographies are recorded as invalid rather than zero overlap. No exact duplicate source images are found. Among 700 train–test source-image pairs, three yield geometrically valid positive matches. Visual review distinguishes one short-interval same-sortie pair (81 s and 60 frame numbers apart) with repeated coverage from two cross-date reobservations of the same site, acquired five days apart. We count only the former as repeated same-session coverage, not as an exact duplicate. The cross-date pairs are distinct acquisitions with different illumination, water conditions, and transient content, although the same fixed revetment structures are present in both acquisitions. The maximum homography-projected minimum-polygon overlap score is 0.6474; because it extrapolates a locally fitted transform across the image footprints, it is a screening indicator rather than a verified fraction of identical pixels. Thus, splitting at the source-image level prevents direct reuse of tiles derived from the same source image across subsets, but it does not eliminate all spatial

dependence between neighboring views. The test results are interpreted as held-out views within the same corridor rather than geographically independent reach-level estimates.

2.2.4. Evaluation Metrics

Region-level segmentation performance is measured using class-wise intersection over union (IoU) and mean intersection over union (mIoU) [35]. For class c, the calculation process of IoU is as follows:

$$\mathrm{IoU}(c) = \frac{\mathrm{TP}(c)}{\mathrm{TP}(c) + \mathrm{FP}(c) + \mathrm{FN}(c)} \tag{2}$$

where $\mathrm{TP}(c)$ is the number of correctly predicted pixels of class c, $\mathrm{FP}(c)$ is the number of pixels incorrectly assigned to class c, $\mathrm{FN}(c)$ is the number of class-c pixels assigned to another class. Each class IoU is computed from the test-set confusion matrix, and mIoU is the unweighted arithmetic mean over the C evaluated classes. Higher IoU and mIoU indicate better region overlap.

Boundary-sensitive measures are computed for the revetment class only. Let P and G denote the predicted and reference binary revetment masks. Their one-pixel contours are expanded by a tolerance $\delta = 3$ pixels to form the boundary bands $\mathrm{B}_\delta(\mathrm{P})$ and $\mathrm{B}\delta(\mathrm{G})$. Boundary IoU (BIoU) [36] is then computed from the intersection and union of these two boundary bands:

$$\mathrm{BIoU} = \frac{|\mathrm{B}_\delta(\mathrm{P}) \cap \mathrm{B}_\delta(\mathrm{G})|}{|\mathrm{B}_\delta(\mathrm{P}) \cup \mathrm{B}_\delta(\mathrm{G})|} \tag{3}$$

A higher BIoU indicates closer overlap between the predicted and reference boundary bands.

Boundary F-score (BFscore) uses the same three-pixel matching tolerance [35,36]. Boundary precision and boundary recall are defined as:

$$\mathrm{P}_b = \frac{\left|\left\{x \in \mathrm{B}_{pred} \mid d(x, \mathrm{B}_{gt}) \le \sigma\right\}\right|}{\left|\mathrm{B}_{pred}\right|} \tag{4}$$

$$\mathrm{R}_b = \frac{\left|\left\{y \in \mathrm{B}_{gt} \mid d\left(y, \mathrm{B}_{pred}\right) \le \sigma\right\}\right|}{\left|\mathrm{B}_{gt}\right|} \tag{5}$$

where $d(\cdot, \mathrm{B})$ is the minimum Euclidean distance from a boundary pixel to boundary set B. The boundary F-score is:

$$\mathrm{BF} = \frac{2\mathrm{P}_b\mathrm{R}_b}{\mathrm{P}_b + \mathrm{R}_b} \tag{6}$$

A higher value indicates that the predicted boundary is both accurately located and sufficiently complete.

Chamfer distance (CD) measures the symmetric mean nearest-neighbor distance between the two contours [37]:

$$\mathrm{CD} = \frac{1}{\left|\mathrm{B}_{pred}\right|} \sum_{x \in \mathrm{B}_{pred}} d(x, \mathrm{B}_{gt}) + \frac{1}{\left|\mathrm{B}_{gt}\right|} \sum_{y \in \mathrm{B}_{gt}} d\left(y, \mathrm{B}_{pred}\right) \tag{7}$$

CD is reported in pixels, and a lower value indicates that the predicted contour lies closer to the reference contour.

Computational efficiency is characterized by parameter count (Params), floating-point operations (FLOPs), mean inference time per image (Time), and throughput in frames per second (FPS). Params reflects model size, FLOPs reflects the computation required for

one forward pass at the stated input size, and Time is measured in milliseconds under the common hardware setting in Section 2.5. FPS is calculated as 1000/Time when Time is expressed in milliseconds. Lower Params, FLOPs, and Time, and higher FPS indicate greater computational efficiency.

For cross-dataset evaluation, the absolute mIoU drop and the relative drop are utilized. For object-prior refinement, ΔIoU is reported in percentage points. Positive ΔIoU indicates improvement after SAM-based fusion, whereas a negative value indicates degradation.

2.3. Object-Prior-Guided Revetment Extraction

2.3.1. Overall Pipeline

The downstream extraction pipeline contains three stages (Figure 8): semantic segmentation supplies an initial revetment mask; SAM produces object-level candidate masks; and overlap and spatial constraints select candidates for fusion with the semantic seed. The experiment evaluates the effect of object priors on revetment region and boundary accuracy.

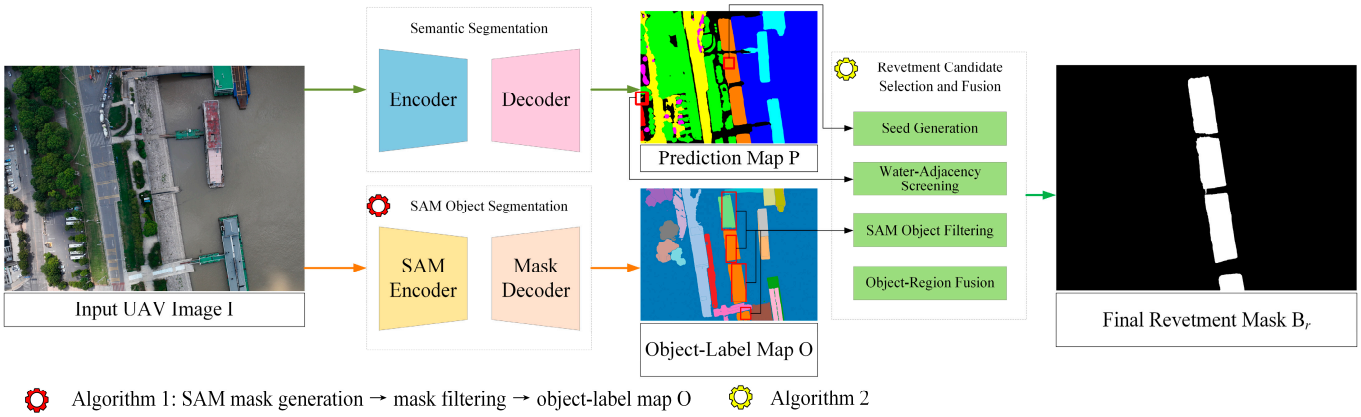


Figure 8. Workflow of SAM-based object-prior refinement for revetment extraction.

2.3.2. Initial Semantic Prediction

The refinement procedure can accept predictions from any semantic segmentation model. GHJNet predictions are used as semantic seeds in the main case study because GHJNet obtains the highest revetment IoU in the in-domain benchmark. This setting evaluates whether SAM object masks can further improve the strongest initial revetment prediction.

For an input image, the trained segmentation model outputs a multiclass prediction map. The binary seed source is obtained by selecting pixels assigned to the revetment class.

Let $\mathbf{P}$ denote the multiclass prediction and c_{rev} the revetment label. The initial binary mask is extracted as follows:

$$\mathbf{B}_{rev}(x,y) = \begin{cases} 1, & \mathbf{P}(x,y) = c_{rev} \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

where a value of one denotes a pixel predicted as revetment.

2.3.3. Revetment Seed Generation

The initial mask $\mathbf{B}_{rev}$ is first filtered by removing 8-connected components smaller than 80 pixels, yielding $\mathbf{B}_{clean}$. Morphological closing of $\mathbf{B}_{clean}$ produces the semantic seed $\mathbf{B}_{seed}$. A separate dilated support $\mathbf{B}_{adj}$ is constructed for the seed-contact test:

$$\mathbf{B}_{seed} = (\mathbf{B}_{clean} \oplus \mathbf{K}) \ominus \mathbf{K}, \mathbf{B}_{adj} = \mathbf{B}_{seed} \oplus \mathbf{K}_{dil} \tag{9}$$

where $\oplus$ and $\ominus$ denote morphological dilation and erosion, respectively; $\mathbf{K}$ is a 3×3 square kernel for closing; and $\mathbf{K}_{dil}$ is a 5×5 kernel for seed-support dilation. $\mathbf{B}_{seed}$ is used for candidate overlap and fusion, whereas $\mathbf{B}_{adj}$ is used only for the seed-contact test.

2.3.4. SAM Object-Mask Generation

SAM with the Huge variant of Vision Transformer (ViT-H) image encoder [38] is applied without fine-tuning. Recent remote-sensing studies have also adapted prompt learning for instance segmentation and evaluated SAM under zero- and one-shot settings [39,40], supporting its use here as an object-prior generator rather than a replacement for the semantic benchmark. The automatic mask generator uses a grid of point prompts to produce object-level candidates from the same UAV image. At the mask-decoder stage, prompt features and image features are fused through cross-attention

$$\mathbf{F}_{fused} = \mathrm{CrossAttn}\left(\mathbf{F}_{prompt}, \mathbf{F}_{img}, \mathbf{F}_{img}\right) \tag{10}$$

and each mask is then obtained as:

$$\mathbf{M} = \sigma\left(\mathrm{MLP}\left(\mathbf{F}_{fused}\right)\right) \in \left[0,1\right]^{H \times W} \tag{11}$$

where $\sigma(\cdot)$ is the sigmoid activation. The SAM automatic mask generator also uses a minimum mask-region area of 100 pixels. The generated masks are sorted by area in descending order; the first $\mathrm{N}_{mask} = 20$ masks are considered, and masks smaller than $\mathrm{A}_{min} = 100$ pixels are removed. The retained masks are written sequentially to the object-label map O; consequently, a later, smaller mask can replace the label of a larger mask at overlapping pixels. Algorithm 1 formalizes this conversion. The retained candidates provide object boundaries but not waterway-specific semantics; they are therefore screened using their overlap with semantic revetment seeds and their spatial relationship to water.

Algorithm 1. SAM candidate-mask filtering and object-label-map construction.

Input: UAV image $\mathrm{I} \in \mathbb{R}^{H \times W \times 3}$; SAM automatic mask generator; $\mathrm{N}_{mask} = 20$; $\mathrm{A}_{min} = 100$ pixels
Output: Object-label map $\mathrm{O} \in \mathbb{N}^{H \times W}$
1: Generate the candidate-mask set $\mathbf{M} \leftarrow$ SAM(I)
2: Sort $\mathbf{M}$ by mask area in descending order to obtain M_{sorted}
3: Retain $\mathrm{M}_{top} \leftarrow \{\mathrm{M}_1, ..., \mathrm{M}_{min(\mathrm{N}mask,\ |\mathrm{M}sorted|)}\}$, where $|\mathrm{M}_{sorted}|$ denotes the number of candidate masks
4: Initialize $\mathrm{O}(x, y) \leftarrow 0$ for all pixels and object index $i \leftarrow 1$
5: For each $\mathrm{M}_k \in \mathrm{M}_{top}$ in the sorted order:
6: If $|\mathrm{M}_k| \geq \mathrm{A}_{min}$, assign $\mathrm{O}[\mathrm{M}_k] \leftarrow i$ and update $i \leftarrow i + 1$, where $|\mathrm{M}_k|$ denotes the pixel area of mask M_k
7: Return O

2.3.5. Overlap-Based Fusion with Spatial Constraints

For each SAM object region $Q_i = \{(x, y) \mid \mathrm{O}(x, y) = i\}$, we compute its overlap ratio with the closed, undilated semantic seed defined in Equation (9):

$$R_i = \frac{|\ Q_i \cap \mathbf{B}_{seed}\ |}{|\ Q_i\ |} \tag{12}$$

where $|\cdot|$ denotes the pixel count. Based on R_i, we apply a three-tier screening strategy:

1. High-overlap retention: a candidate is retained when its overlap ratio with the seed is at least 0.6.
2. Conditional retention: candidates with overlap ratios from 0.2 to 0.6 are retained only when both spatial conditions below are satisfied.
   - Water adjacency: the candidate contacts a dilated water mask derived from the semantic prediction, reflecting the typical location of revetments along the land–water interface.
   - Seed proximity: the candidate intersects the dilated support $\mathbf{B}_{adj}$. This contact test is retained as a consistency check. It does not independently reject candidates with qualifying positive overlap because the dilated support contains the semantic seed.
3. Low-overlap rejection: candidates with overlap ratios below 0.2 are discarded.

The retained object set is therefore:

$$\mathcal{Q}_{keep} = \left\{ Q_i \;\middle|\; R_i \geq r_{high} \text{ or } \left(r_{low} \leq R_i < r_{high} \text{ and } Q_i \cap \mathbf{B}_{water}^{dilate} \neq \emptyset \text{ and } Q_i \cap \mathbf{B}_{adj} \neq \emptyset \right) \right\} \tag{13}$$

The fused revetment mask, before final cleanup, is obtained as:

$$\mathbf{B}_r = \mathbf{B}_{seed} \cup \left(\bigcup_{Q_i \in \mathcal{Q}_{keep}} Q_i \right) \tag{14}$$

Algorithm 2 summarizes the overlap- and spatial-constraint screening and object-region fusion. The accepted SAM regions are combined with the closed, undilated semantic seed. The dilated seed support and water mask are used only for contact tests. After the final cleanup, pixels within the processed revetment mask are assigned the revetment label in the multiclass prediction, while pixels outside this mask retain their original labels.

Algorithm 2. Overlap- and spatial-constraint screening for revetment-mask fusion.

Input: Multiclass prediction $\mathbf{P} \in \mathbb{Z}^{H \times W}$; object-label map $\mathrm{O} \in \mathbb{N}^{H \times W}$; labels c_{rev} and c_{water}; kernels $\mathbf{K}(3 \times 3)$, $\mathbf{K}_{dil}(5 \times 5)$, and $\mathbf{K}_{water}(7 \times 7)$; $\mathrm{A}_{min} = 100$ pixels; $r_{low} = 0.2$; $r_{high} = 0.6$
Output: Refined binary revetment mask $\mathbf{B}_r \in \{0,1\}^{H \times W}$ and corrected prediction $\mathbf{P}_{refined} \in \mathbb{Z}^{H \times W}$
1: Extract $\mathbf{B}_{rev} \leftarrow 1[\mathbf{P} = c_{rev}]$; remove 8-connected components smaller than 80 pixels to obtain $\mathbf{B}_{clean}$
2: Compute $\mathbf{B}_{seed} \leftarrow (\mathbf{B}_{clean} \oplus \mathbf{K}) \ominus \mathbf{K}$
3: Compute $\mathbf{B}_{adj} \leftarrow \mathbf{B}_{seed} \oplus \mathbf{K}_{dil}$ and $\mathbf{B}_{water}^{dilate} \leftarrow 1[\mathbf{P} = c_{water}] \oplus \mathbf{K}_{water}$
4: Set $\mathbf{B}_{sam} \leftarrow 0$
5: For each nonzero object identifier i in O:
6: Set $Q_i \leftarrow \{(\mathrm{x}, \mathrm{y}) \mid \mathrm{O}(x, y) = i\}$; if $|Q_i| < \mathrm{A}_{min}$, continue
7: Compute $R_i \leftarrow \frac{|Q_i \cap \mathbf{B}_{seed}|}{|Q_i|}$
8: If $R_i \geq r_{high}$, set $\mathbf{B}_{sam} \leftarrow \mathbf{B}_{sam} \cup Q_i$
9: Else if $r_{low} \leq R_i < r_{high}$ and $Q_i \cap \mathbf{B}_{water}^{dilate} \neq \emptyset$ and $Q_i \cap \mathbf{B}_{adj} \neq \emptyset$, set $\mathbf{B}_{sam} \leftarrow \mathbf{B}_{sam} \cup Q_i$
10: Fuse $\mathbf{B}_r \leftarrow \mathbf{B}_{seed} \cup \mathbf{B}_{sam}$; then apply the final cleanup
11: Set $\mathbf{P}_{refined} \leftarrow \mathbf{P}$ and assign $\mathbf{P}_{refined}[\mathbf{B}_r] \leftarrow c_{rev}$
12: Return $\mathbf{B}_r$ and $\mathbf{P}_{refined}$

The parameters have distinct operational roles. Retaining the 20 largest candidate masks limits the object pool before semantic screening, while the 100-pixel minimum removes very small fragments. The 3×3 closing kernel fills narrow gaps within the semantic seed, and the 5×5 dilation kernel creates a local tolerance region for contact with object candidates. The 0.6 threshold defines direct high-overlap retention, whereas the interval from 0.2 to 0.6 invokes the additional water-adjacency and seed-proximity conditions; candidates below 0.2 are rejected.

2.4. Visual-Completeness Screening

2.4.1. Connectivity and Geometric Features

The refined binary revetment mask is converted from a pixel map into object-level descriptors. Connected components are obtained by border following [41], giving the number of components N, total revetment area A_{total}, largest-component area A_{max}, and second-largest area A_2. Connectivity is summarized by the connectivity index $CI = \frac{A_{max}}{A_{total}}$ and $R_{top2} = \frac{A_{max}+A_2}{A_{total}}$. Values close to one indicate that the predicted revetment is concentrated in one or two dominant components rather than many fragments.

For the dominant component, a minimum bounding rectangle with long and short sides L and W is used to calculate aspect ratio $AR = \frac{L}{W}$ and rectangularity $Extent = \frac{A}{LW}$. Shape compactness is $SC = \frac{4\pi A}{P^2}$, where A and P are component area and perimeter, and boundary roughness is $BR = \frac{P}{2(L+W)}$. These measures describe visible continuity, elongation, filling, and boundary irregularity of the extracted mask; they do not measure material strength or subsurface condition.

2.4.2. Rule Categories and Reference Interpretation

The visual-completeness case comprises 42 full-resolution source images: the 10 images in the RLGD test subset and 32 images selected in a second screening of the original archive of more than 1000 UAV images. This screening prioritizes views with substantial visible revetment coverage and seeks to reduce repeated depictions of the same photographed bank section. Thresholds are selected post hoc from the geometric-feature distributions of these 42 images and routine visual-interpretation criteria. The rules in Table 6 define three visible-morphology categories and a no-revetment category. The intermediate category is a residual review category containing detected-revetment images that satisfy neither the continuous nor the fragmented rule.

Table 6. Rule-based categories for visual-completeness screening.

Visual Category	Rule Applied to the Refined Revetment Mask
No revetment	No revetment pixel is detected
Continuous	$N \leq 2$, $CI \geq 0.95$, $SC \geq 0.25$, $Extent \geq 0.80$, $BR < 6$, and $AR < 4$
Fragmented	$CI \geq 0.95$ and at least one of $SC < 0.15$, $Extent < 0.60$, $BR \geq 10$, or $AR \geq 8$
Intermediate	All remaining images containing a detected revetment; manual review is retained for this transition category

Reference interpretation is performed by three domain experts with waterway-inspection and remote-sensing experience using the original high-resolution images and terminology informed by the Construction Code of Channel Regulation Works (JTS 224-2016) [27]. Each image is assessed for visible continuity, apparent completeness, and consistency of the revetment along the bank, and the final image-level reference category is recorded by consensus. These references express expert visual-screening experience, which is not pixel-level masks or field-verified structural-condition labels.

Overall agreement is the fraction of the 42 rule-based labels that match the expert-reference category. To reduce sensitivity to the imbalanced category counts, we additionally report the complete four-class confusion matrix, macro-F1, and balanced accuracy. These statistics describe agreement with the expert image-level interpretation.

2.5. Implementation Details

All segmentation models are implemented with MMSegmentation 0.30.0 [42] and PyTorch 1.12.1. Each 3000 × 4000 source image is divided into twelve 1024 × 1024 patches using the crop-first strategy, and each patch is resized to 512 × 512 before network input. The full-image reconstruction uses 32-pixel horizontal and 36-pixel vertical overlap, with uniform averaging in overlapping regions. Compared with scaling the complete source image directly, this strategy preserves more local detail for revetments, vehicles, and ships while reducing whole-scene context. Training uses batch size 2 for 40,000 iterations, AdamW [43] with the initial learning rate 6×10^{-5} and weight decay 0.01, validation every 2000 iterations, random flipping, rotation, scale variation, and color perturbation. RLGD samples are repeated three times during data loading. The six public baselines use their standard cross-entropy objectives, whereas GHJNet uses the online hard example mining (OHEM) semantic loss and class-balanced boundary loss. The Time and FPS values are measured for semantic-network inference on 512 × 512 inputs with batch size 1 on an NVIDIA GeForce RTX 3060 (NVIDIA Corporation, Santa Clara, CA, USA). They exclude full-image reconstruction and SAM-based processing. In addition, the median across the ten 3000 × 4000 test images is reported; per-image runtime is the sum of the measured semantic prediction, logit stitching, SAM generation, candidate screening, and mask fusion times. Peak allocated GPU memory is recorded across the complete per-image chain with both networks resident, and the maximum is reported.

GHJNet is trained with separate semantic and boundary supervision. The final multi-class prediction uses OHEM cross-entropy, with a true-class probability threshold of 0.90 and a batch-level lowest-confidence valid pixels setting of 131,072. For each batch, valid pixels are ranked by the predicted probability of their ground-truth class, and cross-entropy is averaged over the selected low-confidence pixels. The detail branch produces a single-channel boundary logit supervised by class-balanced binary cross-entropy. Boundary targets are generated from the ground-truth semantic mask by comparing valid neighboring labels in the horizontal, vertical, and two diagonal directions. The resulting edge map is dilated to form a boundary-supervision band. Boundary and non-boundary pixels receive weights of $w_1 = \frac{N_{non-boundary}}{N_{boundary}+N_{non-boundary}}$ and $w_0 = \frac{N_{boundary}}{N_{boundary}+N_{non-boundary}}$, respectively. The semantic and boundary losses are combined with weights of 1 and 20, respectively.

SAM with the ViT-H image encoder is used without fine-tuning. The automatic mask generator uses a 16 × 16 point-prompt grid with a point batch size of 16. The predicted-IoU and stability-score thresholds are both 0.90, the box- and crop-level non-maximum suppression (NMS) thresholds are both 0.50, multi-layer cropping is disabled, and the minimum mask-region area is 100 pixels. During object-label-map construction, candidates smaller than 100 pixels are excluded, and at most 20 masks are retained after area ranking. The subsequent refinement stage applies a separate 100-pixel object-area cutoff, 3 × 3 seed closing, 5 × 5 seed-support dilation, 7 × 7 water dilation, 3 × 3 final closing, and overlap thresholds of 0.20 and 0.60. For the local sensitivity analysis, one parameter family is varied at a time. Other settings remain at their defaults.

The default SAM and fusion settings are selected in small preliminary experiments on validation images; the held-out test images are not used for parameter selection. Before seed closing, revetment components smaller than 80 pixels are removed. After candidate fusion, cleanup is applied. The components smaller than 500 pixels are removed, enclosed

holes of at most 100 pixels are filled, and 3×3 closing is applied. In the component ablation, one processing stage is omitted at a time while the remaining configured stages are kept unchanged.

3. Results

3.1. Experimental Setup

The in-domain experiments use 840/240/120 tiles derived, respectively, from the fixed 70/20/10 source-image split (Table 7). The public release provides only RLGD masks. For the VDD-to-RLGD direct transfer experiment, the internal RLGD annotations retain separate wall and roof labels. Ship and revetment are excluded, and the seven labels shared with VDD are evaluated. No RLGD fine-tuning is performed in this transfer setting.

Table 7. Summary statistics of the RLGD after image tiling.

Split	Images	Tiles	Tiles with Revetment	Revetment Ratio (%)	Average Classes per Tile
Train	70	840	276	32.86	2.97
Validation	20	240	80	33.33	3.20
Test	10	120	30	25.00	3.19

The principal in-domain benchmark reports all eight-class IoUs and mIoU (Table 8), together with revetment BIoU, BFscore, and CD (Table 9), as mean ± sample standard deviation over the five training results. The cross-dataset experiment reports source and target mIoU together with absolute and relative drops, while the object-prior experiment reports initial and refined revetment IoU and ΔIoU.

Table 8. Class-wise IoU and mIoU on the RLGD in-domain test split.

Model	Other	Building	Road	Vegetation	Vehicle	Water	Ship	Revetment	mIoU ↑
GHJNet	84.01 ± 0.38	51.28 ± 0.38	71.30 ± 2.07	93.18 ± 0.30	41.26 ± 2.47	98.19 ± 0.14	93.44 ± 1.23	83.51 ± 1.18	77.02 ± 0.41
SegFormer-B0	82.66 ± 0.99	51.26 ± 1.58	65.05 ± 2.75	92.93 ± 0.88	31.25 ± 1.26	97.90 ± 0.29	91.64 ± 1.61	82.51 ± 0.74	74.40 ± 0.52
SegNeXt	81.02 ± 1.13	44.88 ± 5.99	68.44 ± 3.08	91.06 ± 0.63	32.49 ± 11.95	97.72 ± 0.28	89.55 ± 3.08	81.64 ± 2.05	73.35 ± 1.66
PSPNet	74.95 ± 5.62	50.21 ± 1.89	63.24 ± 5.86	87.07 ± 3.65	42.29 ± 13.71	96.97 ± 0.80	86.96 ± 3.65	75.46 ± 4.24	72.15 ± 1.69
OCRNet	76.38 ± 1.94	47.32 ± 4.60	60.50 ± 6.61	89.47 ± 1.18	41.13 ± 8.70	97.60 ± 0.66	73.74 ± 12.60	71.86 ± 6.82	69.75 ± 1.93
UPerNet	74.14 ± 1.30	44.78 ± 4.48	56.09 ± 5.92	88.07 ± 1.57	42.96 ± 4.20	97.57 ± 0.20	85.98 ± 5.63	64.55 ± 3.26	69.27 ± 1.67
DeepLabV3+	74.41 ± 2.83	43.02 ± 11.17	65.09 ± 8.63	86.02 ± 3.00	40.32 ± 5.42	97.02 ± 1.55	80.56 ± 11.90	72.09 ± 3.36	69.82 ± 4.51

Table 9. Five-run mIoU and revetment boundary performance (mIoU, BIoU, and BFscore in %, CD in pixels).

Model	mIoU ↑	BIoU ↑	BFscore ↑	CD ↓	Params (M) ↓	FLOPs (G) ↓	Time (ms) ↓	FPS ↑
GHJNet	77.02 ± 0.41	70.61 ± 1.10	79.53 ± 1.00	12.07 ± 3.48	4.40	19.19	29.25	34.18
SegFormer-B0	74.40 ± 0.52	69.22 ± 1.68	75.71 ± 1.68	14.84 ± 3.50	3.72	7.91	22.30	44.84
SegNeXt	73.35 ± 1.66	71.38 ± 1.68	77.58 ± 1.30	16.99 ± 3.14	4.23	6.30	28.73	34.81
PSPNet	72.15 ± 1.69	64.08 ± 3.79	70.52 ± 4.84	19.48 ± 4.99	48.97	178.64	47.01	21.27
OCRNet	69.75 ± 1.93	62.22 ± 4.00	67.57 ± 3.48	20.33 ± 4.42	12.07	53.24	42.62	23.46
UPerNet	69.27 ± 1.67	54.88 ± 4.34	59.63 ± 3.79	35.02 ± 13.33	20.38	62.36	19.32	51.77
DeepLabV3+	69.82 ± 4.51	59.84 ± 4.63	65.69 ± 3.45	23.31 ± 9.14	43.58	176.53	52.24	19.14

The evaluated models are DeepLabV3+ [29], PSPNet [30], UPerNet [31], OCRNet [32], SegFormer [33], SegNeXt [34], and GHJNet. SegFormer-B1 is additionally included in the transfer experiment.

For the independent 12-image annotation audit, the two masks agree on 96.68% of pixels and obtain 85.16% eight-class mIoU. Class-wise IoUs are 74.79% for other, 81.74% for building, 87.69% for road, 95.29% for vegetation, 56.98% for vehicle, 99.03% for water, 92.53% for ship, and 93.28% for revetment. The revetment boundary F-score is 25.37%, 42.83%, 55.90%, 71.52%, and 82.69% at tolerances of 1, 3, 5, 10, and 20 pixels, respectively. Thus, these annotators agree strongly on the revetment region and on water, vegetation,

and ships, while the exact contour position and small vehicles remain the main sources of annotation uncertainty.

3.2. In-Domain Semantic Segmentation Benchmark

Table 8 displays the complete eight-class RLGD benchmark and reports the mean ± sample standard deviation for each class and mIoU over five training results. GHJNet has the highest descriptive mean mIoU (77.02 ± 0.41%), followed by SegFormer-B0 (74.40 ± 0.52%) and SegNeXt (73.35 ± 1.66%). Water remains comparatively stable across architectures, whereas vehicle, building, and some road/ship results vary more with the model and initialization.

The revetment column shows that GHJNet obtains the highest five-run mean IoU (83.51 ± 1.18%), followed by SegFormer-B0 (82.51 ± 0.74%) and SegNeXt (81.64 ± 2.05%). Conversely, no one architecture dominates every category: the vehicle mean ranges from 31.25 ± 1.26% for SegFormer-B0 to 42.96 ± 4.20% for UPerNet, and several models show much greater variation for this sparse class than for water. The results therefore support the task-specific revetment comparison while retaining the less favorable and less stable class-level outcomes for inspection.

The five-run boundary summary in Table 9 qualifies the region-level ranking. GHJNet yields the highest mean revetment BFscore (79.53 ± 1.00%) and the lowest mean Chamfer distance (12.07 ± 3.48 pixels), whereas SegNeXt has the highest mean BIoU (71.38 ± 1.68%); GHJNet is close on BIoU (70.61 ± 1.10%). Thus, GHJNet does not lead every boundary criterion, but it combines the strongest aggregate mIoU and revetment IoU with competitive boundary localization. As for time and FPS, it can be seen that SegFormer-B0 remains the smallest model (3.72 M parameters), SegNeXt has the lowest reported FLOPs, and UPerNet has the highest reported network-only FPS.

The complete eight-class results in Table 8 make the sparse-category concern directly inspectable. To test one potential mitigation, we conduct a controlled three-seed loss-function audit with SegFormer-B0. Standard cross-entropy obtains 74.75 ± 0.21% mIoU, 31.98 ± 1.07% vehicle IoU, 92.52 ± 0.46% ship IoU, 82.19 ± 0.63% revetment IoU, and 68.63 ± 1.66% revetment Boundary IoU. Class-balanced cross-entropy obtains 74.85 ± 0.55%, 36.52 ± 2.22%, 91.43 ± 3.20%, 81.39 ± 2.53%, and 70.83 ± 1.09%, respectively; softmax focal loss with $\gamma = 2$ obtains 73.55 ± 1.76%, 34.30 ± 4.89%, 90.81 ± 0.56%, 81.55 ± 1.71%, and 70.73 ± 1.27%.

The class-balanced objective therefore improves mean vehicle IoU by 4.54% without a material change in overall mIoU, but neither alternative improves ship or revetment IoU, and both introduce trade-offs. This result supports loss reweighting as a useful sparse-class option while showing that it is not a uniform solution for all infrequent or boundary-sensitive classes.

3.3. Cross-Dataset Transfer and the Waterway-Domain Gap

To test whether performance on a varied general UAV dataset transfers directly to waterway imagery, models trained on VDD are evaluated on the seven-label mapping used in Table 10 without target-domain fine-tuning. This protocol isolates the unadapted domain gap. A separate SegFormer-B0 audit then evaluates few-shot target adaptation on the six classes that can be mapped exactly between the two datasets (other, building, road, vegetation, vehicle, and water).

All models lose accuracy under direct transfer (Table 10). OCRNet obtains the highest RLGD mIoU (46.23%), followed by SegFormer-B0 (45.17%). Relative mIoU reductions range from 11.29% for DeepLabV3+ to 49.81% for SegFormer-B1. The absolute values and

ranking changes show that strong in-domain VDD performance does not guarantee robust waterway transfer.

Table 10. Cross-dataset transfer from VDD to the seven-class RLGD common-label subset.

Model	VDD Test mIoU	RLGD mIoU	Absolute Drop	Relative Drop (%)
SegFormer-B0	70.05	45.17	24.88	35.52
PSPNet	50.21	35.09	15.12	30.11
SegNeXt	61.83	41.88	19.95	32.27
OCRNet	65.23	46.23	19.00	29.13
UPerNet	49.96	41.98	7.98	15.97
DeepLabV3+	40.13	35.60	4.53	11.29
SegFormer-B1	43.97	22.07	21.90	49.81

Under the strict six-class audit, direct transfer obtains 37.35% shared-class mIoU. Fine-tuning with 4, 7, and 14 RLGD source images (approximately 5.7%, 10%, and 20% of the 70-image training set) yields 38.83%, 46.23%, and 60.02%, respectively, compared with 71.99% for the in-domain reference after the same remapping. These settings close 4.3%, 25.6%, and 65.5% of the measured direct-to-in-domain gap. The 20% result demonstrates substantial adaptation benefit, while the remaining gap shows that limited target labels do not remove the domain shift.

The transfer degradation supports the need for RLGD. Waterway images contain large water and vegetation regions, narrow engineered bank structures, oblique views of riverside buildings, and land–water transitions that differ from general urban UAV imagery.

Figure 9 presents predictions from the validation-selected checkpoints of all seven segmentation models on the same five representative RLGD images. It can be seen that all seven models recover the dominant regions in the five selected scenes but differ at narrow boundaries and small objects. GHJNet generally follows the reference revetment shape closely, particularly at interfaces with water and vegetation. SegNeXt and SegFormer-B0 also produce relatively coherent regions. DeepLabV3+, OCRNet, and PSPNet show local omissions or class leakage along these interfaces. UPerNet has conspicuous spurious predictions in the first revetment scene and around the curved road in the second scene. In the vehicle-rich scene, predictions vary across all models, with both missed and excessive vehicle regions. These examples illustrate the difficulty of delineating class boundaries and sparse objects despite broadly accurate predictions of large regions.

3.4. Downstream Revetment Extraction with Object Priors

Table 11 compares the initial prediction, morphology-only result, and complete SAM-based refinement for all seven segmentation backbones. Morphology alone changes revetment IoU by only 0.01–0.08%. After the complete refinement, SegNeXt, SegFormer-B0, and GHJNet gain 3.42%, 2.62%, and 3.15%, respectively, whereas DeepLabV3+, OCRNet, UPerNet, and PSPNet lose 8.46%, 0.83%, 3.66%, and 0.91%, respectively. GHJNet-SAM achieves the highest refined revetment IoU at 87.25%.

The morphology-to-fusion contrast is particularly clear for DeepLabV3+ and GHJNet: their revetment IoUs change from 69.33% to 60.82% and from 84.13% to 87.25%, respectively. The opposite responses arise after object-prior fusion, indicating that the suitability of accepted SAM candidates is more consequential than the small morphological adjustment. Figure 10 illustrates the refinement sequence for a representative GHJNet result, from the initial semantic prediction and SAM candidate objects to the extracted revetment seed and final fused mask.

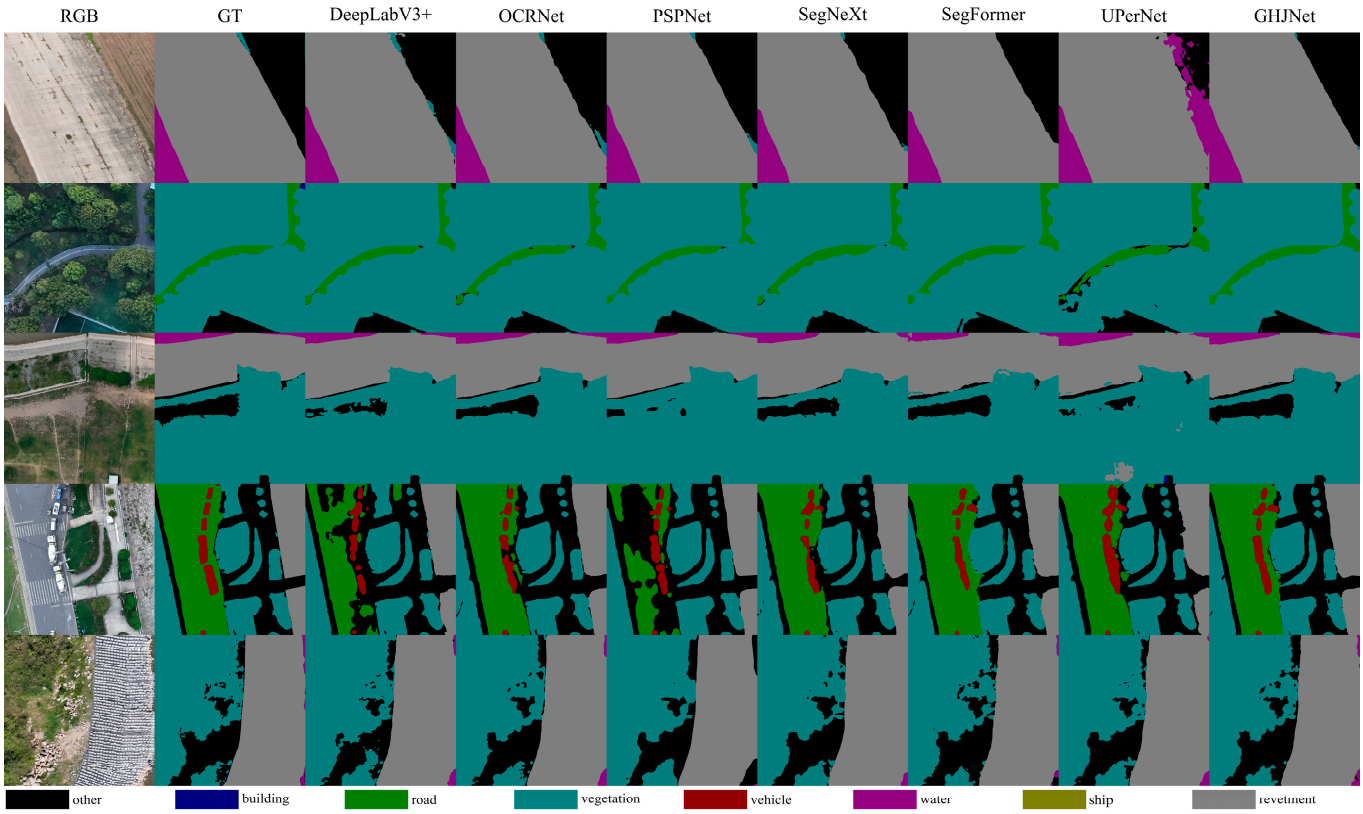


Figure 9. Qualitative comparison on five representative RLGD images. Columns show RGB, ground truth, DeepLabV3+, OCRNet, PSPNet, SegNeXt, SegFormer-B0, UPerNet, and GHJNet.

Table 11. Effect of SAM-based object-prior refinement on revetment extraction.

Model	Initial IoU	Morph. IoU	SAM IoU	ΔIoU	SAM BIoU	SAM BFscore	SAM CD ↓
DeepLabV3+-SAM	69.28	69.33	60.82	−8.46	63.91	73.74	20.09
OCRNet-SAM	61.48	61.52	60.65	−0.83	62.14	70.98	30.25
UPerNet-SAM	62.64	62.72	58.98	−3.66	59.56	68.03	36.50
PSPNet-SAM	75.16	75.19	74.25	−0.91	62.35	71.75	27.32
SegNeXt-SAM	82.59	82.60	86.01	+3.42	72.51	83.61	14.09
SegFormer-B0-SAM	82.43	82.47	85.05	+2.62	72.48	82.48	25.13
GHJNet-SAM	84.10	84.13	87.25	+3.15	73.52	84.16	10.95

A local sensitivity analysis further examines the GHJNet-SAM configuration. The default yields 77.63% mIoU, 87.25% revetment IoU, and 73.52% BIoU. Varying the retained-mask count from 10 to 40 yields 77.25–77.63% mIoU, 86.22–87.25% revetment IoU, and 73.50–73.52% BIoU. Varying the paired overlap thresholds yields 77.15–77.63%, 85.63–87.25%, and 72.64–73.52%, respectively. The tested downstream object-area cutoff and seed-closing, seed-dilation, and water-dilation kernels leave these three metrics unchanged. The final-closing kernel changes the revetment IoU only from 87.24% to 87.25%.

In the component ablation (Table 12), removing water adjacency causes the largest decreases (0.42%, 1.35%, and 1.72%, respectively). Removing morphological seed processing changes these metrics by 0.09%, 0.07%, and 0.05%. Removing the seed-proximity check does not change the values.

We further profile the complete semantic–SAM pipeline for the seven models on the ten source images. Median stage-summed runtime ranges from 50.03 to 54.94 s per 3000 × 4000 image; GHJNet requires 54.09 s. SAM mask generation alone takes a median of 33.96 s. The maximum observed peak allocated GPU memory ranges from 6989 to 7163 MB across models and is 6991 MB for GHJNet. These full-image measurements include semantic prediction, stitching, SAM generation, candidate screening, and fusion.

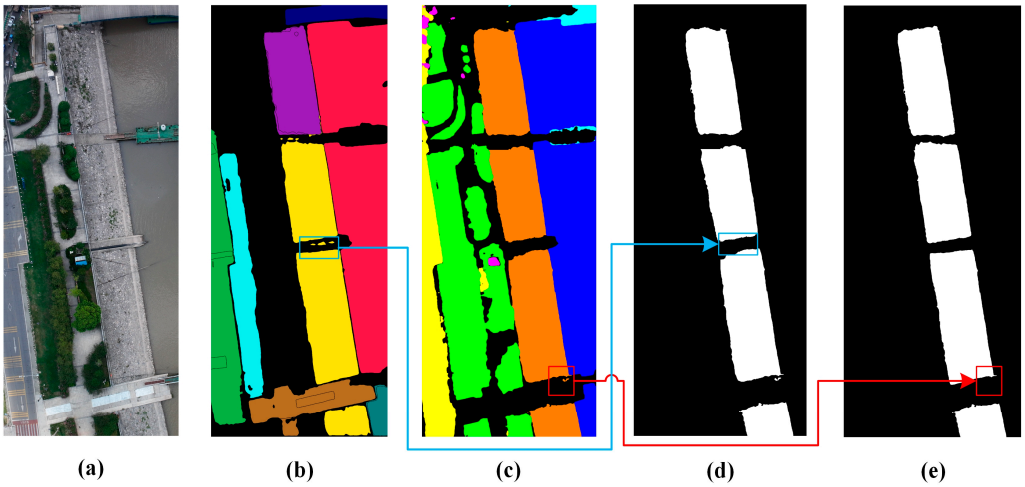


Figure 10. Stepwise revetment refinement for a representative RLGD image: (**a**) input image; (**b**) SAM candidate objects; (**c**) initial GHJNet prediction; (**d**) semantic seed; and (**e**) fused revetment mask.

Table 12. Ablation study of GHJNet-SAM refinement on RLGD (%).

Setting	mIoU (%)	Revetment IoU (%)	BIoU (%)
GHJNet-SAM	77.63	87.25	73.52
Without morphological seed processing	77.54	87.18	73.47
Without water adjacency	77.21	85.90	71.80
Without seed proximity	77.63	87.25	73.52

Separately from SAM refinement, we compared three stitching rules using GHJNet predictions. Hard ownership, uniform averaging, and center-weighted averaging yielded mIoUs of 76.86%, 77.02%, and 76.93%, respectively; revetment IoUs of 84.10%, 84.10%, and 84.11%; and seam-area revetment IoUs of 85.92%, 86.05%, and 86.27%. Center weighting produced the highest seam-area revetment IoU, whereas uniform averaging produced the highest mIoU. Whole revetment IoU differed by no more than 0.02% among the methods.

3.5. Visual-Completeness Screening Results

The rule system is applied to 42 images and compared with the three-expert consensus interpretation (Table 13). The complete confusion matrix contains 34 diagonal cases, giving 80.95% overall agreement, 83.93% macro-F1, and 81.25% balanced accuracy. All eight disagreements are conservative assignments to the intermediate category: seven expert-reference continuous images and one fragmented image are assigned as intermediate. Thus, the rule system never confuses a detected-revetment case with no revetment in this collection, but recall for the continuous category is 50.0%.

Table 13. Confusion matrix between three-expert image-level reference interpretation (rows) and rule-based visual-completeness screening (columns).

Reference\Rule	Continuous	Intermediate	Fragmented	No Revetment	Row Total
Continuous	7	7	0	0	14
Intermediate	0	20	0	0	20
Fragmented	0	1	3	0	4
No revetment	0	0	0	4	4
Column total	7	28	3	4	42

Figure 11 shows representative examples. The continuous mask is compact and fills most of its minimum bounding rectangle; the intermediate example remains connected but has a less regular boundary; and the fragmented example is elongated and rough.

These labels describe the visible geometry of the extracted mask rather than confirmed physical deterioration.

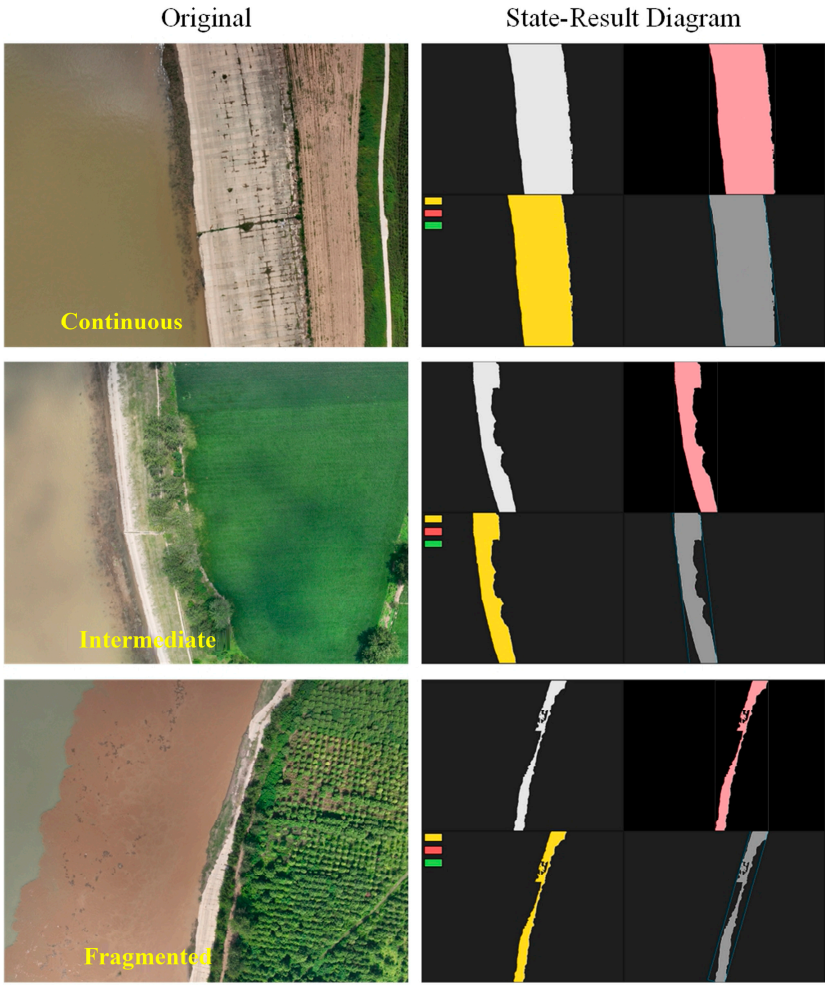


Figure 11. Representative visual-completeness screening examples. From top to bottom, the image displays revetments categorized as continuous, intermediate, and fragmented. The right-hand 2 × 2 blocks display the refined binary mask and auxiliary connected-component, contour, and minimum-bounding-rectangle renderings. White, pink, yellow, and gray/cyan are visualization aids for these derived mask and geometry views; they are not semantic classes or indicators of structural damage.

4. Discussion

4.1. Dataset Characteristics, Benchmark Behavior, and Domain Transfer

The contribution of RLGD lies in its task definition and annotation density rather than image count alone. The 100 images contain 1.2 billion labeled pixels at 3000 × 4000 pixels per image. More importantly, the released RLGD explicitly defines revetment while retaining the land–water context needed to distinguish it from roads, vegetation, exposed surfaces, and riverside structures. The annotation workflow initially distinguishes wall and roof, but RLGD consolidates them as building so that the public benchmark emphasizes waterway functional semantics rather than viewpoint-driven wall/roof decomposition. RLGD adapts complementary design principles from VDD, AFID, and HRSID. Annotation density and independent scene diversity remain distinct: the pixel count describes labeling volume, whereas the benchmark contains 100 source images from one corridor.

The dataset statistics explain why model performance must be interpreted beyond one run and one loss. Vegetation and water occupy 70.36% of all labeled pixels, whereas vehicle, ship, and building pixels are comparatively sparse; revetment accounts for 8.52%

and has low Bhattacharyya distances to the other and road classes. Across the seven models, five-run mean mIoU ranges from 69.27% to 77.02% and mean revetment IoU from 64.55% to 83.51%. The class-wise Table 8 shows that water IoU is high and relatively stable, whereas vehicle and some other sparse-category results vary substantially. In the controlled three-seed SegFormer-B0 audit, class-balanced cross-entropy raises vehicle IoU from 31.98% to 36.52% with similar overall mIoU (74.75% versus 74.85%), but ship and revetment IoU do not improve. Sparse-class optimization is therefore useful but class-specific, and no single loss removes the long-tail problem.

The five-run boundary results further clarify model selection. GHJNet combines the highest mean mIoU and revetment IoU with the highest mean BFscore and lowest mean CD, but SegNeXt yields the highest mean BIoU. In terms of computational efficiency, SegFormer-B0 has the smallest parameter count, SegNeXt has the lowest reported FLOPs, and UPerNet has the highest throughput under the batch-size-1 protocol. The preferred model therefore depends on whether the application prioritizes scene-wide and revetment-region overlap, contour agreement, run-to-run stability, or computational cost.

Direct VDD-to-RLGD evaluation documents a substantial source-to-target gap under the seven-label transfer protocol: every model loses accuracy, with relative mIoU reductions of 11.29–49.81%. The separate strict six-class SegFormer-B0 audit shows that adaptation with 4, 7, and 14 target source images closes 4.3%, 25.6%, and 65.5% of the direct-to-in-domain gap, respectively. This evidence supports both conclusions: broad UAV pretraining alone does not remove the waterway-domain gap, but a limited amount of target supervision can recover a substantial share of performance.

Recent task-specific remote-sensing studies further contextualize these challenges. Bakirci [44] shows how small maritime targets and atmospheric or preprocessing effects complicate high-resolution satellite-image detection. Zhao et al. [45] combine super-resolution reconstruction with UAV semantic segmentation for crop monitoring, highlighting the importance of fine-detail preservation and task-specific scene variation. Wang et al. [46] demonstrate that high waterbody-segmentation scores do not necessarily imply accurate shoreline positioning, reinforcing the need for boundary-sensitive evaluation at land–water interfaces. In contrast to these detection, agricultural, and shoreline-focused studies, RLGD provides a dedicated multiclass waterway taxonomy with an explicit revetment label and evaluates both region and boundary accuracy.

4.2. Conditional Effectiveness of Semantic–SAM Fusion and Visual Screening

The contrasting outcomes follow from the different roles of the semantic prediction and SAM. The semantic model identifies the revetment, whereas SAM supplies class-agnostic object masks [38]. A candidate is accepted according to its overlap with the predicted revetment seed and, for intermediate overlap, its relationship with the predicted water and expanded seed regions. When the seed lies on the true revetment, an accepted object can recover omitted segments and extend the mask along their boundaries. When the seed overlaps a neighboring road, exposed bank, or vegetation region, fusion can instead extend that error across the candidate object and add false positives.

This mechanism is consistent with the pattern in Table 11. The three models with the highest initial revetment IoUs—GHJNet (84.10%), SegNeXt (82.59%), and SegFormer-B0 (82.43%)—improve after fusion, whereas the other four decline. DeepLabV3+ shows the largest reduction (−8.46%), a pattern expected when added non-revetment area is large relative to recovered revetment area. The fusion rule thus makes spatial alignment of the seed central to the outcome.

The change in IoU reflects the balance between recovered revetment pixels and non-revetment pixels added by accepted objects. Table 11 shows that morphology alone has a

small effect, whereas candidate fusion can help or hurt according to the quality of the semantic seed. The local sensitivity analysis shows that the retained-mask count and overlap thresholds influence candidate selection. In the GHJNet component ablation (Table 12), water adjacency has the clearest measured contribution, consistent with restricting candidates to the waterway interface. Seed morphology makes smaller changes. The seed-proximity contact check is retained as an explicit implementation safeguard, but it does not yield an independent gain under the present rule because a candidate with qualifying positive overlap already intersects the seed and its dilated support.

The visual-completeness experiment extends the refined masks to an image-level screening task. Agreement with three-expert consensus is 80.95%, macro-F1 is 83.93%, and balanced accuracy is 81.25%. All disagreements are intermediate assignments rather than extreme continuous-versus-absent confusions. These exploratory results suggest that connectivity and geometry of the predicted two-dimensional mask may help prioritize images for human review. Assessing material condition or structural safety requires separate field or complementary sensing evidence.

4.3. Limitations and Future Directions

RLGD currently represents one Yangtze River corridor (the Jingjiang Section), and all presently available images were acquired during the dry season (low-water period). The benchmark therefore does not cover seasonal hydrological variation, high-water conditions, or external river reaches. The benchmark uses 840/240/120 tiles derived from a 70/20/10 source-image split rather than a random tile-level split, and all tiles from the same source image remain in the same subset. Nevertheless, visual review of the source-image audit identifies one short-interval similar-sortie train–test pair with repeated coverage and two reacquisition pairs of the same site collected five days apart. The latter are distinct acquisitions, but the presence of the same fixed bank structures in both acquisitions indicates that the test set should not be interpreted as geographically independent. Future datasets should preserve grouping metadata, reserve complete river reaches for testing, and expand across river systems, seasons, water levels, weather conditions, sensors, and revetment forms.

The crop-first strategy preserves local detail but reduces whole-scene context. In the pre-SAM stitching comparison, center-weighted averaging increased seam-area revetment IoU relative to hard ownership (86.27% versus 85.92%), while whole-image revetment IoU remained essentially unchanged (84.11% versus 84.10%). Although center-weighted averaging produced the highest seam-area IoU, we retained uniform logit averaging for full-image reconstruction because it yielded the highest mIoU (77.02%) with comparable whole-image revetment IoU (84.10%). Input strategies with broader scene context remain future work.

Annotation undergoes manual correction and second-person review, and a representative 12-image subset is independently reannotated for audit. The results support high region-level inter-annotator agreement, with 96.68% pixel agreement, 85.16% mIoU, and 93.28% revetment IoU, while revealing lower agreement for vehicles (56.98% IoU) and residual variability in precise revetment boundary placement. The revetment boundary F-score increases from 42.83% at a 3-pixel tolerance to 82.69% at 20 pixels, indicating that the two annotations agree well at the region level but differ more in exact contour placement.

The SAM-ViT-H stage is the main computational bottleneck of the downstream workflow. On the RTX 3060, SAM mask generation takes a median of 33.96 s per 3000 × 4000 source image, while the complete pipeline takes 50.03–54.94 s per image across the seven models and reaches 6989–7163 MB in peak allocated GPU memory. The current workflow is therefore intended for offline analysis rather than real-time inspection. Lighter foundation

models, prompt-efficient variants, multispectral observations, and temporal consistency remain important directions for reducing computational cost and improving recovery under occlusion.

5. Conclusions

This study introduces RLGD, a high-resolution UAV dataset and benchmark for semantic segmentation and revetment extraction in waterway scenes. RLGD contains 100 images at 3000 × 4000 pixels and provides an eight-class taxonomy with an explicit revetment category and surrounding land–water context. Under a unified evaluation protocol, seven segmentation models achieve five-run mean mIoU values of 69.27–77.02%, with GHJNet obtaining the highest mean mIoU (77.02 ± 0.41%) and revetment IoU (83.51 ± 1.18%). An independent reannotation audit further supports the region-level reliability of the dataset while revealing greater variability in precise boundary placement and sparse classes.

Cross-dataset evaluation shows a substantial domain gap between general UAV imagery and waterway scenes, whereas limited target-domain adaptation recovers a considerable portion of the lost performance. The downstream semantic–SAM experiments further show that object-prior refinement can improve revetment extraction in some settings, but its benefit depends on seed-guided candidate selection. Geometry-based screening also demonstrates potential for image-level inspection prioritization.

RLGD should be regarded as a task-specific benchmark for dry-season, in-corridor waterway analysis rather than a geographically independent or universally representative dataset. Future work should extend the benchmark to additional river reaches, seasons, water levels, sensors, and revetment forms, with geographically grouped evaluation and independent field validation.

Author Contributions: Conceptualization, A.S., R.X., S.Y., Z.Z., Z.C., and N.C.; methodology, A.S., R.X., R.Z., S.Y., Z.C., and N.C.; software, A.S., R.X., R.Z., S.Y., and T.X.; validation, S.Y., Z.Z., D.L., and Y.J.; formal analysis, A.S., R.X., S.Y., and N.C.; investigation, A.S., R.X., and R.Z.; resources, A.S., Z.Z., D.L., and N.C.; data curation, R.X., Z.Z., D.L., and Y.J.; writing—original draft preparation, A.S., R.X., R.Z., and S.Y.; writing—review and editing, R.X., S.Y., and N.C.; visualization, A.S., R.Z., and T.X.; supervision, Z.C., and N.C.; project administration, N.C.; funding acquisition, N.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Key Science and Technology Project of the Changjiang River Administration of Navigational Affairs (MOT) (Grant No. 2025-020-6-Z-Y) and the Key Research and Development Project of Hubei province (Grant No. 2024BCB101).

Data Availability Statement: The RLGD data will be deposited in a versioned public repository. Repository URL: https://github.com/yang199809/RLGD (accessed on 12 October 2026).

Acknowledgments: The authors would like to thank the editors and anonymous reviewers for their valuable comments and suggestions to improve the quality of this paper.

Conflicts of Interest: Authors Rongjun Xiong, Ziwei Zhu, and Dongsheng Li were employed by the company Changhang Testing Technology (Wuhan) Co., Ltd., Wuhan, China. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Chen, T.; He, H.; Li, D.; An, P.; Hui, Z. Damage signature generation of revetment surface along urban rivers using UAV-based mapping. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 283. [CrossRef]
2. Sun, H.; Jiang, Z.; Ying, J.; Yu, Z.; Xiao, J. Temporal-spatial characteristics of technical status of channel regulation buildings in trunk stream of the Yangtze River. *Port Waterw. Eng.* **2024**, *6*, 128–134.

3. Su, H.; Ma, J.; Zhou, R.; Wen, Z. Detect and identify earth rock embankment leakage based on UAV visible and infrared images. *Infrared Phys. Technol.* **2022**, *122*, 104105. [CrossRef]
4. Wan, X.; Liu, H. Analysis on development process and trends of inland waterway navigation in China. *China Harb. Eng.* **2025**, *45*, 25–32.
5. Li, D.; Li, M. Research advance and application prospect of unmanned aerial vehicle remote sensing system. *Geomat. Inf. Sci. Wuhan Univ.* **2014**, *39*, 505–513.
6. Hemmelder, S.; Marra, W.; Markies, H.; de Jong, S.M. Monitoring river morphology and bank erosion using UAV imagery—A case study of the River Buëch, Hautes-Alpes, France. *Int. J. Appl. Earth Obs. Geoinf.* **2018**, *73*, 428–437. [CrossRef]
7. Lyu, Y.; Vosselman, G.; Xia, G.-S.; Yilmaz, A.; Yang, M.Y. UAVid: A semantic segmentation dataset for UAV imagery. *ISPRS J. Photogramm. Remote Sens.* **2020**, *165*, 108–119. [CrossRef]
8. Cai, W.; Jin, K.; Hou, J.; Guo, C.; Wu, L.; Yang, W. VDD: Varied drone dataset for semantic segmentation. *J. Vis. Commun. Image Represent.* **2025**, *109*, 104429. [CrossRef]
9. Wang, Z.; Mahmoudian, N. Aerial fluvial image dataset for deep semantic segmentation neural networks and its benchmarks. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 4755–4766. [CrossRef]
10. Wei, S.; Zeng, X.; Qu, Q.; Wang, M.; Su, H.; Shi, J. HRSID: A high-resolution SAR images dataset for ship detection and instance segmentation. *IEEE Access* **2020**, *8*, 120234–120254. [CrossRef]
11. Cheng, J.; Deng, C.; Su, Y.; An, Z.; Wang, Q. Methods and datasets on semantic segmentation for unmanned aerial vehicle remote sensing images: A review. *ISPRS J. Photogramm. Remote Sens.* **2024**, *211*, 1–34. [CrossRef]
12. Nigam, I.; Huang, C.; Ramanan, D. Ensemble knowledge transfer for semantic segmentation. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018; pp. 1499–1508.
13. Wang, J.; Zheng, Z.; Ma, A.; Lu, X.; Zhong, Y. LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*; Neural Information Processing Systems Foundation (NeurIPS): San Diego, CA, USA, 2021; Volume 1.
14. Xia, J.; Yokoya, N.; Adriano, B.; Broni-Bediako, C. OpenEarthMap: A benchmark dataset for global high-resolution land cover mapping. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 2–7 January 2023; pp. 6254–6264.
15. Garioud, A.; Gonthier, N.; Landrieu, L.; De Wit, A.; Valette, M.; Poupée, M.; Giordano, S.; Wattrelos, B. FLAIR: A country-scale land cover semantic segmentation dataset from multi-source optical imagery. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 16456–16482. [CrossRef]
16. Zamir, S.W.; Arora, A.; Gupta, A.; Khan, S.; Sun, G.; Khan, F.S.; Zhu, F.; Shao, L.; Xia, G.-S.; Bai, X. iSAID: A large-scale dataset for instance segmentation in aerial images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Long Beach, CA, USA, 16–20 June 2019; pp. 28–37.
17. Azimi, S.M.; Henry, C.; Sommer, L.; Schumann, A.; Vig, E. SkyScapes—Fine-grained semantic understanding of aerial scenes. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 7393–7403.
18. Rahnemoonfar, M.; Chowdhury, T.; Sarkar, A.; Varshney, D.; Yari, M.; Murphy, R.R. FloodNet: A high resolution aerial imagery dataset for post flood scene understanding. *IEEE Access* **2021**, *9*, 89644–89654. [CrossRef]
19. Boguszewski, A.; Batorski, D.; Ziemba-Jankowska, N.; Dziedzic, T.; Zambrzycka, A. LandCover.ai: Dataset for automatic mapping of buildings, woodlands, water and roads from aerial imagery. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Nashville, TN, USA, 19–25 June 2021; pp. 1102–1110.
20. Demir, I.; Koperski, K.; Lindenbaum, D.; Pang, G.; Huang, J.; Basu, S.; Hughes, F.; Tuia, D.; Raskar, R. DeepGlobe 2018: A challenge to parse the Earth through satellite images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 18–22 June 2018; pp. 172–181.
21. 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. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 8815–8827. [CrossRef]
22. Cheng, B.; Misra, I.; Schwing, A.G.; Kirillov, A.; Girdhar, R. Masked-attention mask transformer for universal image segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 1290–1299.
23. Krähenbühl, P.; Koltun, V. Efficient inference in fully connected CRFs with Gaussian edge potentials. *Adv. Neural Inf. Process. Syst.* **2011**, *24*, 109–117.
24. Borse, S.; Wang, Y.; Zhang, Y.; Porikli, F. InverseForm: A loss function for structured boundary-aware segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 5901–5911.
25. Kirillov, A.; Wu, Y.; He, K.; Girshick, R. PointRend: Image segmentation as rendering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 9799–9808.

26. Berman, M.; Triki, A.R.; Blaschko, M.B. The Lovász-Softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 4413–4421.
27. *JTS 224-2016*; Construction Code of Channel Regulation Works. China Communications Press: Beijing, China, 2016.
28. Kailath, T. The divergence and Bhattacharyya distance measures in signal selection. *IEEE Trans. Commun. Technol.* **1967**, *15*, 52–60. [CrossRef]
29. 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 European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
30. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
31. Xiao, T.; Liu, Y.; Zhou, B.; Jiang, Y.; Sun, J. Unified perceptual parsing for scene understanding. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 418–434.
32. Yuan, Y.; Chen, X.; Wang, J. Object-contextual representations for semantic segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; pp. 173–190.
33. 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.
34. Guo, M.-H.; Lu, C.-Z.; Hou, Q.; Liu, Z.; Cheng, M.-M.; Hu, S.-M. SegNeXt: Rethinking convolutional attention design for semantic segmentation. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 1140–1156. [CrossRef]
35. Csurka, G.; Larlus, D.; Perronnin, F. What is a good evaluation measure for semantic segmentation? In Proceedings of the British Machine Vision Conference (BMVC), Bristol, UK, 9–13 September 2013; pp. 1–11.
36. Cheng, B.; Girshick, R.; Dollár, P.; Berg, A.C.; Kirillov, A. Boundary IoU: Improving object-centric image segmentation evaluation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 15334–15342.
37. Barrow, H.G.; Tenenbaum, J.M.; Bolles, R.C.; Wolf, H.C. Parametric correspondence and chamfer matching: Two new techniques for image matching. In Proceedings of the 5th International Joint Conference on Artificial Intelligence (IJCAI), Cambridge, MA, USA, 22–25 August 1977; pp. 659–663.
38. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.-Y.; et al. Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 4015–4026.
39. Chen, K.; Liu, C.; Chen, H.; Zhang, Z.; Li, W.; Zou, Z.; Shi, Z. RSPrompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4701117. [CrossRef]
40. Osco, L.P.; Wu, Q.; de Lemos, E.L.; Gonçalves, W.N.; Ramos, A.P.M.; Li, J.; Marcato Junior, J. The Segment Anything Model (SAM) for remote sensing applications: From zero to one shot. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *124*, 103540. [CrossRef]
41. Suzuki, S.; Abe, K. Topological structural analysis of digitized binary images by border following. *Comput. Vis. Graph. Image Process.* **1985**, *30*, 32–46. [CrossRef]
42. MMSegmentation: OpenMMLab Semantic Segmentation Toolbox and Benchmark. Available online: https://github.com/open-mmlab/mmsegmentation (accessed on 23 March 2026).
43. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. In Proceedings of the International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019.
44. Bakirci, M. Advanced ship detection and ocean monitoring with satellite imagery and deep learning for marine science applications. *Reg. Stud. Mar. Sci.* **2025**, *81*, 103975. [CrossRef]
45. Zhao, F.; He, Y.; Song, J.; Wang, J.; Xi, D.; Shao, X.; Wu, Q.; Liu, Y.; Chen, Y.; Zhang, G.; et al. Smart UAV-assisted blueberry maturity monitoring with Mamba-based computer vision. *Precis. Agric.* **2025**, *26*, 56. [CrossRef]
46. Wang, W.; Lu, B.; Li, Y.; Ji, F. A quantitative assessment of the inconsistency between waterbody segmentation and shoreline positioning in deep learning models. *Geomatics* **2026**, *6*, 21. [CrossRef]

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.