Journal of
Marine Science and Engineering

MDPI

Article

Camera-LiDAR Cross-Modality Fusion Water Segmentation for Unmanned Surface Vehicles

Jiantao Gao, Jingting Zhang, Chang Liu, Xiaomao Li * and Yan Peng

Research Institute of USV Engineering, School of Mechatronic Engineering and Automation, Shanghai University, Shanghai 200444, China; summersunday@shu.edu.cn (J.G.); zjt322@shu.edu.cn (J.Z.); liuchang123@shu.edu.cn (C.L.); pengyan@shu.edu.cn (Y.P.)
* Correspondence: lixiaomaosia@163.com

Abstract: Water segmentation is essential for the autonomous driving system of unmanned surface vehicles (USVs), which provides reliable navigation for making safety decisions. However, existing methods have only used monocular images as input, which often suffer from the changes in illumination and weather. Compared with monocular images, LiDAR point clouds can be collected independently of ambient light and provide sufficient 3D information but lack the color and texture that images own. Thus, in this paper, we propose a novel camera-LiDAR cross-modality fusion water segmentation method, which combines the data characteristics of the 2D image and 3D LiDAR point cloud in water segmentation for the first time. Specifically, the 3D point clouds are first supplemented with 2D color and texture information from the images and then distinguished into water surface points and non-water points by the early 3D cross-modality segmentation module. Subsequently, the 3D segmentation results and features are fed into the late 2D cross-modality segmentation module to perform 2D water segmentation. Finally, the 2D and 3D water segmentation results are fused for the refinement by an uncertainty-aware cross-modality fusion module. We further collect, annotate and present a novel Cross-modality Water Segmentation (CMWS) dataset to validate our proposed method. To the best of our knowledge, this is the first water segmentation dataset for USVs in inland waterways consisting of images and corresponding point clouds. Extensive experiments on the CMWS dataset demonstrate that our proposed method can significantly improve image-only-based methods, achieving improvements in accuracy and MaxF of approximately 2% for all the image-only-based methods.

Keywords: water segmentation; semantic segmentation; image segmentation; LiDAR point cloud; deep learning; unmanned surface vessel

check for updates

Citation: Gao, J.; Zhang, J.; Liu, C.; Li, X.; Peng, Y. Camera-LiDAR Cross-Modality Fusion Water Segmentation for Unmanned Surface Vehicles. *J. Mar. Sci. Eng.* **2022**, *10*, 744. https://doi.org/10.3390/jmse10060744

Academic Editors: Mai The Vu, Hyeung-Sik Choi

Received: 2 May 2022
Accepted: 26 May 2022
Published: 28 May 2022

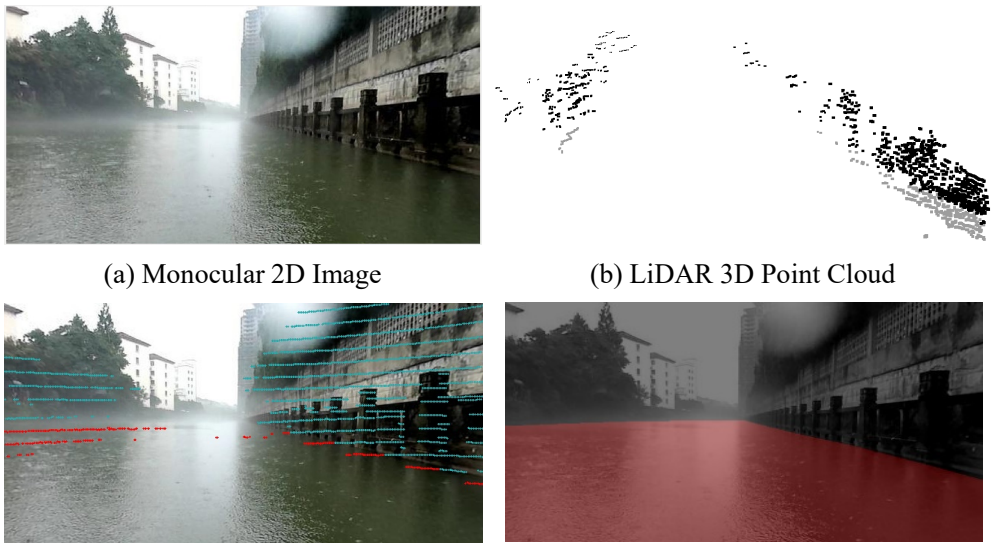
1. Introduction

Unmanned surface vehicles (USVs) are boats or ships operating autonomously on the water's surface without a crew. They have been widely used in recent years to perform various laborious and dangerous offshore operations, such as search and rescue operations [1], hydrographic surveying and charting [2], water quality monitoring [3] and other tasks. In particular, the application of USVs in inland waterways is closely related to human life and has great potential value, such as the construction of an autonomous transportation system for inland waterways [4]. In its autonomous driving system, stable and accurate water segmentation plays a crucial role, which provides reliable navigation for making safety decisions [5].

Over the years, water segmentation based on monocular images [6–13] has made significant progress. It only takes monocular images as input and classifies the images into the water surface and non-water region at the pixel level. Despite substantial progress, this method is adversely affected by changes in illumination and weather because the image quality is influenced by ambient light. When visual noise occurs, such as variable lighting, overexposure and blurring, these image-only-based methods perform poorly.

Unlike monocular images, LiDAR point clouds can be collected under various ambient light conditions and provide sufficient 3D information. However, these data have limitations, i.e., point clouds are sparse and lack the color and texture information of images. Therefore, the fusion of the dense RGB semantic information in images with the sparse 3D information from LIDAR point clouds to improve perception capability is an important topic in academia and industry [14,15].

Different from the applications in unmanned ground vehicle (UGV) tasks [14–17], LiDAR has unique advantages in the water segmentation task for USVs in inland waterways. As shown in Figure 1, the LiDAR pulsed laser light will not be reflected on most water surfaces, only on non-water surfaces and very few water surfaces. By exploiting this inherent property of LiDAR, the false positive water segmentation results of image-based methods can be reduced with the incorporation of the LiDAR point clouds. Based on this idea, we proposed a novel Camera-LiDAR cross-modality fusion water segmentation method for USVs in inland waterways, which cleverly combines the data characteristics of the 2D image and 3D LiDAR point cloud in water segmentation for the first time. Specifically, we enhance the 3D point cloud by using 2D color and texture information and distinguish point clouds into water surface points and non-water points by the early 3D cross-modality segmentation module. Subsequently, the 3D segmentation results and features are fed into the late 2D cross-modality segmentation module to perform image-based water segmentation. This method minimizes false positive water segmentation results. Finally, the 2D and 3D water segmentation results are fused for further refinement by an uncertainty-aware cross-modality fusion module.



(a) Monocular 2D Image (b) LiDAR 3D Point Cloud

(c) Camera-LiDAR Fused Data (d) Our Segmentation Results

Figure 1. Monocular 2D image (**a**) and the corresponding LiDAR 3D Point Cloud (**b**). The projection of the point cloud onto the 2D image (**c**) shows that most points (blue points) are reflected by non-water surfaces, whereas few points (red points) are reflected by the water surface due to near-shore reflection. This inherent property of LiDAR is exploited in our proposed method to reduce the false positive results in the image-only-based method and achieve high-accuracy water segmentation (**d**).

Publicly available datasets are essential for validating deep learning methods and conducting a fair comparison between different algorithms. However, existing water segmentation datasets [9,13,18] only have annotated image data but no corresponding LiDAR point cloud data, making it impossible to validate our method and compare it with others. Thus, we collect, annotate and present a novel Cross-Modality Water Segmentation dataset (CMWS dataset) in this paper. To the best of our knowledge, this is the first water segmentation dataset for USVs in inland waterways that has both images and correspond-

ing point clouds. We conduct extensive experiments on this CMWS dataset. The results demonstrate that the proposed camera-LiDAR cross-modality fusion water segmentation method can significantly improve image-only-based methods. It is worth noting that all image-only-based methods achieve stable improvements in accuracy and MaxF.

The main contributions of this paper are as follows:

1. We propose a novel camera-LiDAR cross-modality fusion water segmentation method to combine the 2D image and 3D point cloud characteristics. To the best of our knowledge, we are the first to combine 3D LiDAR point cloud and image data to improve the water segmentation performance;
2. A novel CMWS dataset is proposed to validate the proposed approach. This is the first water segmentation dataset with images and corresponding point clouds for USVs in inland waterways;
3. Extensive experiments are conducted on the CMWS datasets. The results show that the proposed camera-LiDAR cross-modality fusion water segmentation method significantly improves image-only-based methods, achieving improvements in accuracy and MaxF of approximately 2% for all the image-only-based methods.

The rest of this paper is organized as follows. Section 2 reviews studies on water segmentation. Section 3 provides the details of the proposed camera-LiDAR cross-modality fusion water segmentation method. In Section 4, the CMWS dataset is described, and the experimental results on the CMWS dataset are present. The conclusions are provided in Section 5.

2. Related Work

2.1. Water Segmentation Methods for USVs

Water segmentation is a fundamental problem in the USV perception system. The aim is to classify the images into the water surface and non-water regions at the pixel level. Previous methods mainly relied on hand-craft features, such as decision forests [6], support vector machines [7] and expectation-maximization algorithms [8]. Recently, semantic segmentation based on deep learning has been developed rapidly. Wang et al. [19] proposed a novel high-resolution segmentation network (HRNet) to maintain high-resolution representations throughout the whole process for better discrimination. Cheng et al. [20] formulated semantic segmentation as a mask classification problem and combined semantic-level segmentation with instance-level segmentation to achieve better segmentation results. Mohan et al. [21] proposed an efficient panoptic segmentation network based on the semantic segmentation task. In addition, the Transformer in Natural Language Processing (NLP) was applied to the semantic segmentation task by Chen et al. [22]. Since deep learning has achieved excellent performance on the semantic segmentation task, some studies attempted to use deep learning for water segmentation in recent years. Lopez-Fuentes et al. [23] were the first to apply deep learning to water segmentation for USV perception, using fully convolutional networks (FCNs) [11] for semantic segmentation. Taipalmaa et al. [9] adapted the deep learning segmentation algorithm KittiSeg [10] to water segmentation. Moreover, a lightweight FCN [11] architecture was proposed and evaluated in [9] for near real-time applications. Wenqiang et al. [12] proposed a semi-supervised deep learning method for water segmentation for USVs in a dynamic navigation environment. Gui et al. [24] proposed an enhanced UNet [25] by model channel pruning for real-time water segmentation for USVs. Yuwei et al. [13] applied Deeplab V3+ [26] for water segmentation in their proposed water segmentation benchmark.

Nevertheless, existing methods have only used monocular images as input, which are adversely affected by changes in illumination and weather. In contrast to monocular cameras, LiDAR is an active sensor that works independently of ambient light and captures accurate 3D information. Therefore, we propose a novel camera-LiDAR cross-modality fusion deep learning method for water segmentation.

2.2. Water Segmentation Datasets for USVs

Publicly available datasets are crucial to validate deep learning methods and for training deep convolutional neural networks (CNNs). They provide a fair comparison between different algorithms and promote breakthroughs. Various datasets for UGVs have been released (e.g., KITTI [27], Cityscapes [28], Nuscenes [29] and Waymo [30]), and substantial progress has been made using these datasets. Furthermore, many datasets for Unmanned Aerial Vehicles (UAVs) (e.g., Stanford Drone Dataset [31], Okutama-Action dataset [32], UAVDT [33] and MOR-UAV [34]) have been released in the past few years, contributing to the development of UAVs. In contrast, there are few datasets for USVs, especially for water segmentation. The Tampere-WaterSeg dataset [9] consists of 600 manually labeled images selected from videos recorded by a camera mounted on a USV. The MaSTr1325 dataset [18] contains 1325 diverse images captured over a two-year period with a real USV, covering a range of realistic conditions encountered in coastal surveillance tasks. USVInland dataset [13] released a more challenging water segmentation benchmark containing 700 images captured under different weather or lighting conditions in inland waterways.

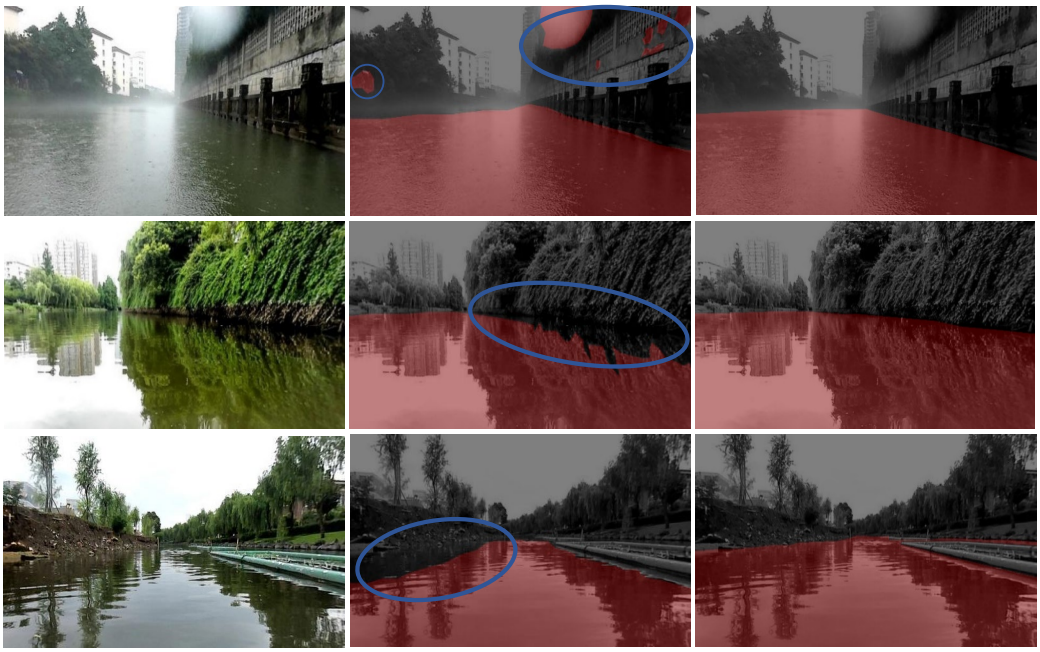
More importantly, existing water segmentation datasets [9,13,18] only have annotated image data, but no corresponding LiDAR point cloud data, making it impossible to validate our method and compare it with others. Thus, we collect, annotate and present a novel Cross-modality Water Segmentation (CMWS) dataset in this paper. To the best of our knowledge, this is the first water segmentation dataset for USVs in inland waterways that has both images and corresponding point clouds. The CMWS dataset contains a total of 3018 images and point cloud data, which is the largest water segmentation dataset. We used this CMWS dataset to conduct extensive experiments and perform fair comparisons of our proposed method and other image-only-based methods.

3. Method

3.1. Limitations of the Current Approaches

Recently, water segmentation for USVs has achieved promising performance. Existing methods [6–13] only use monocular images as input, so they heavily depend on the image quality of the input images. However, monocular images are passively collected relying on ambient light and therefore are adversely affected by changes in illumination and weather. As a result, these image-only-based methods perform poorly in the presence of visual noise, such as variable lighting, overexposure and blurring, as shown in Figure 2. Moreover, due to the water reflections, the visual discrimination between the water surface and the shore is also a challenge for image-based methods.

In contrast to the passive acquisition of images, LiDAR point clouds are actively acquired by emitting pulsed laser light and are not affected by lighting conditions. However, these data also have limitations, i.e., point clouds are sparse and lack the color and texture information of images. Therefore, in this paper, we proposed a novel camera-LiDAR cross-modality fusion water segmentation method, which combines the data characteristics of the 2D image and 3D LiDAR point cloud in water segmentation for the improvement of performance. Specifically, as shown in Figure 3, the proposed camera-LiDAR cross-modality fusion method consists of three parts: an early 3D cross-modality segmentation module, a late 2D cross-modality fusion module and an uncertainty-aware cross-modality prediction fusion module. We use the early 3D cross-modality segmentation module to enhance the 3D point cloud with 2D texture information and segment them into the surface and non-surface points. Subsequently, the late 2D cross-modality fusion module is utilized to integrate the 3D segmentation information and 2D image data for precise water segmentation. Finally, the 2D and 3D water segmentation results are fused for further refinement by the uncertainty-aware cross-modality fusion module.



(a) Monocular 2D images (b) Segmentation results of UNet (c) Ground-truths

Figure 2. The water segmentation results obtained from an image-only-based method (UNet) in various scenarios. This method typically performs poorly due to overexposure, reflective surfaces, other visual noise and near-shore.

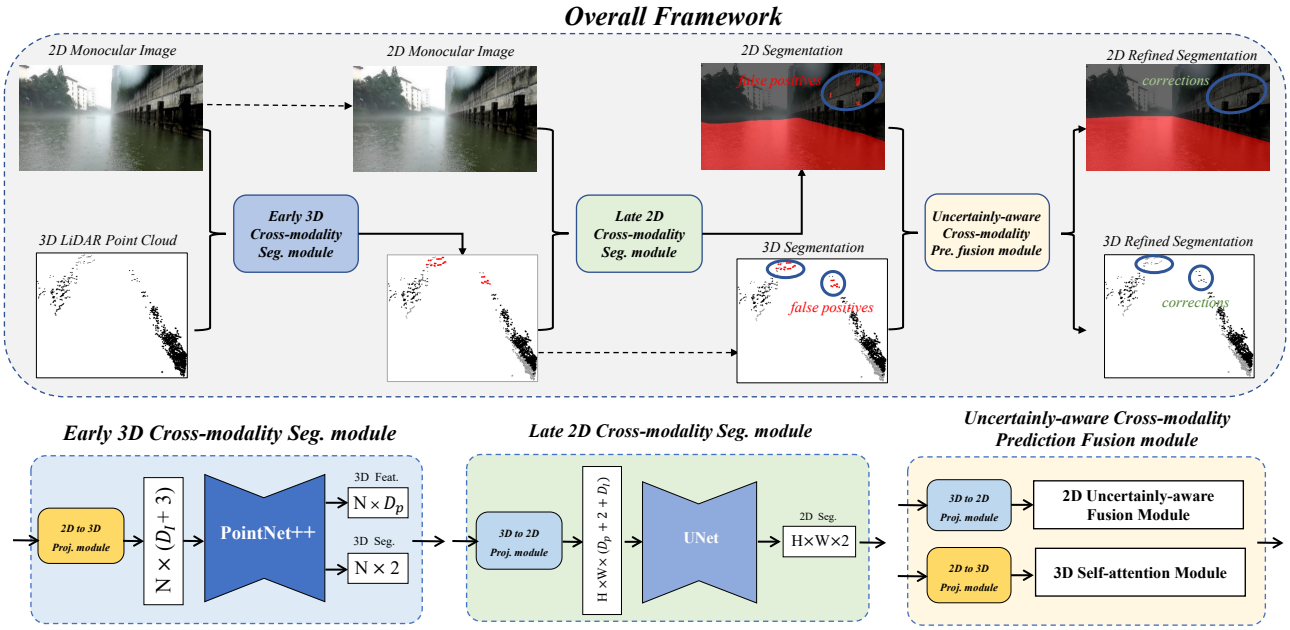


Figure 3. Overall framework of the proposed method. The 3D point clouds are first supplemented with 2D color and texture information from the images and then distinguished into water surface points and non-water points by the early 3D cross-modality segmentation module. Subsequently, the 3D segmentation results and features are fed into the late 2D cross-modality segmentation module to perform 2D water segmentation. Finally, the 2D and 3D water segmentation results are fused to refine their false positives (represented by the red area or points in the blue circles) by an uncertainty-aware cross-modality fusion module.

3.2. Early 3D Cross-Modality Segmentation Module

Although most points in point clouds are mostly reflected by non-water surfaces, some are reflected by the water surface due to near-shore reflections. Therefore, we first perform 3D semantic segmentation to classify the point clouds into water surface points and non-water surface points by using 3D geometric information. This aims at providing more accurate 3D priors for subsequent 2D water segmentation. However, although LiDAR point clouds can be collected independently of ambient light and provide sufficient 3D information, they have limitations, i.e., they are typically sparse and lack the rich color and texture information of images. Thus, we first enhance the 3D point cloud with 2D color and texture information using a 2D-to-3D projection module. We then feed the enhanced point cloud into the 3D segmentation network to distinguish between surface and non-surface points.

3.2.1. 2D-to-3D Projection Module

This module aims at supplementing the 3D point cloud with 2D color and texture information. The details are presented in Figure 4. Given a monocular image $I \in R^{H\times W\times 3}$ and its corresponding LiDAR point cloud $P = \{p_i\}_{i=1}^{N} \in R^{N\times 3}$, the monocular image I is used as input in a 2D feature extraction network. Image-wise 2D features $F^{2D} \in R^{H\times W\times D_I}$ with color and texture information are extracted in the hidden layer. Next, based on the camera's intrinsic $K \in R^{3\times 4}$ and extrinsic matrices $T \in R^{4\times 4}$, the projection of each 3D point $p_i = (\boldsymbol{x_i}, \boldsymbol{y_i}, \boldsymbol{z_i}) \in R^3$ to a pixel $\hat{p}_i(\boldsymbol{i_x}, \boldsymbol{i_y}) \in R^2$ on the image plane is defined as:

$$[i_x, i_y, 1]^T = \frac{1}{z_i} \times K \times T \times [x_i, y_i, z_i, 1]^T, \tag{1}$$

The 2D–3D mapping is represented as:

$$M^{img} = \{(\lfloor i_x \rfloor, \lfloor i_y \rfloor)\}_{i=1}^{N} \in \mathbb{R}^{N\times 2}, \tag{2}$$

where $\lfloor \cdot \rfloor$ is the floor operation.

According to the 2D–3D mapping, we extract a point-wise 2D feature $\hat{F}^{2D} \in R^{N\times D_I}$ from the original feature map F^{2D} if the pixel in the feature map is included in M^{img}. Next, we fuse the projected point-wise 2D feature $\hat{F}^{2D}$ with the 3D point cloud coordinate information P by concatenation to obtain the enhanced 3D point cloud.

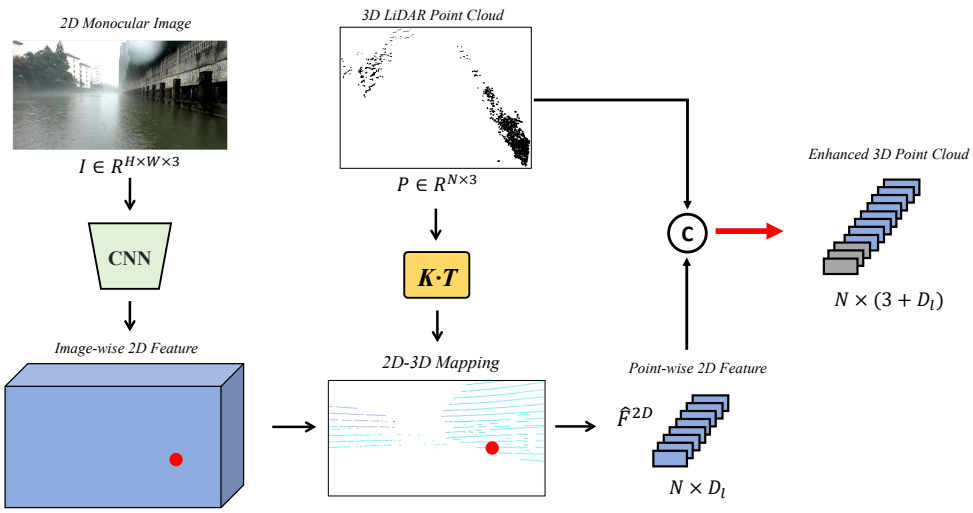


Figure 4. The details of 2D-to-3D projection module.

3.2.2. 3D Segmentation Network

After acquiring the enhanced 3D point cloud, a 3D point cloud segmentation network is employed to extract the point-wise 3D feature $F^{3D} \in R^{N \times D_p}$ and segment the points into water and non-water points $O^{3D} \in R^{N \times 2}$. In this paper, we adopt the efficient and robust 3D segmentation network PointNet++ [35] for 3D semantic segmentation. As shown in Figure 5, PointNet++ is composed of two parts: an encoder with set abstraction modules and a decoder with feature propagation modules. The set abstraction modules sample the farthest point in the given point cloud sets. These points are regarded as the center of a sphere, and their neighbors are determined by the K-nearest neighbor (KNN) [36] algorithm. Subsequently, PointNet [37] is used in the local regions to aggregate features. The feature propagation modules interpolate the subsampled points and concatenate these points with the point features from the set abstraction modules. These features are fed into the PointNet network with a multilayer perceptron (MLP) for updating. Finally, an MLP layer is adopted to classify the points into water and non-water points.

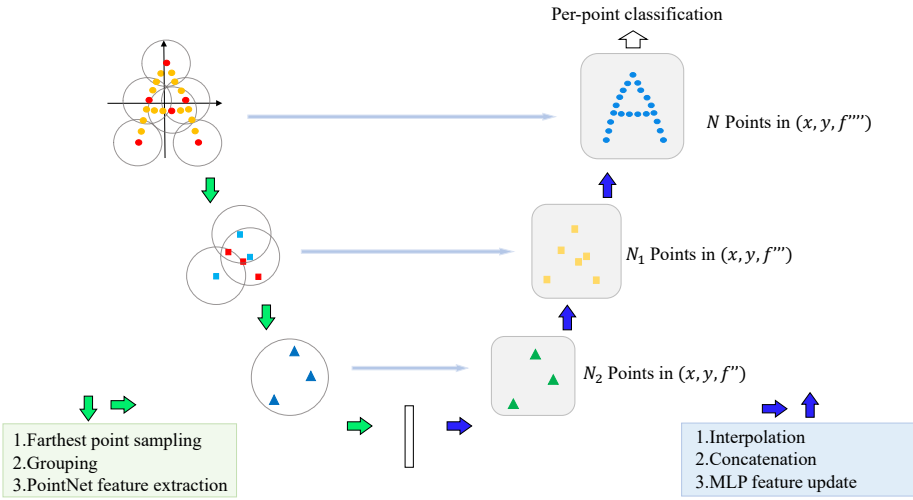


Figure 5. The architecture of the PointNet++.

3.3. Late 2D Cross-Modality Segmentation Module

The early 3D cross-modality segmentation module provides the 3D feature and segmentation result. We first project these priors into a 2D image plane and use them to enhance the 2D image features by the 3D-to-2D projection module. Afterward, the enhanced image features are fed into a 2D segmentation network to conduct 2D water segmentation.

3.3.1. 3D-to-2D Projection Module

As described in Section 3.2.1, the projection of each 3D point to the pixel image plane is calculated based on the camera's intrinsic and extrinsic matrices (i.e., the inverse operation of 2D–3D mapping). We first transfer the 3D feature $F^{3D} \in R^{N \times D_p}$ and segmentation result $O^{3D} \in R^{N \times 2}$ into 2D presentation $\hat{F}^{3D} \in R^{H \times W \times (D_p + 2)}$ and use them to enhance the 2D image features by a Channel-Exchanging-Network (CEN) [19], as shown in Figure 6. The CEN is a general multimodal fusion framework that does not require additional parameters. It achieves multimodal fusion by dynamically exchanging feature channels of different modalities. Specifically, the magnitude of the batch normalization (BN) scaling factor is utilized to measure the importance of each channel and guide the channel exchange when it is greater than a threshold value.

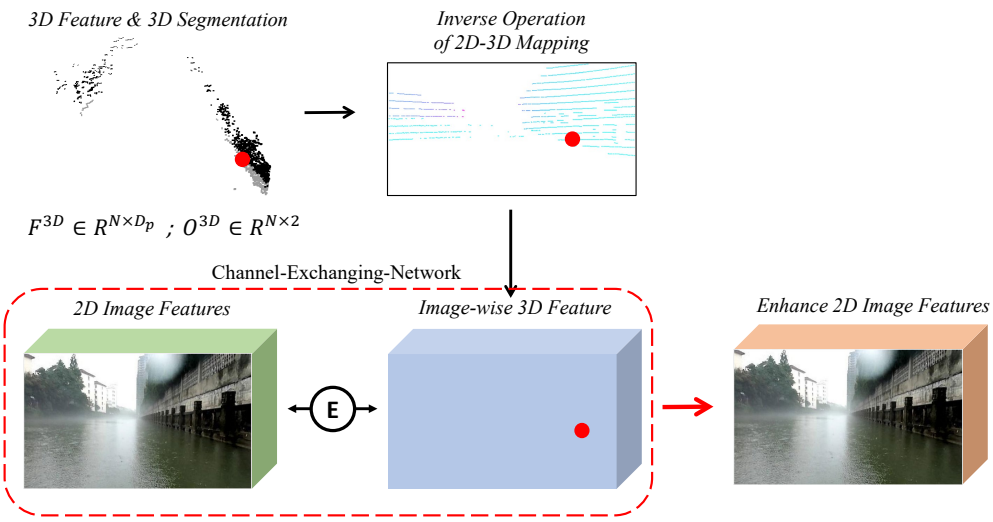


Figure 6. The details of 3D-to-2D projection module.

3.3.2. 2D Segmentation Network

After using the 3D priors to enhance the 2D image features, we feed the enhanced image features into the 2D segmentation network to conduct water segmentation. We adopt the fast and classical UNet [25] as the 2D image segmentation network. As shown in Figure 7, UNet is composed of two parts: a top-down encoder and a bottom-up decoder. The encoder achieves further feature aggregation by reusing down-sampling blocks and residual blocks, while the decoder implements segmentation prediction by reusing up-sampling blocks and residual blocks. The down-sampling and up-sampling blocks are implemented by a 2×2 convolution with stride 2. The output channels of the down-sampling block are doubled, whereas those of the up-sampling block are reduced by half. The residual blocks consist of multiple 3×3 convolutions with identity shortcut connections [38]. Each convolution layer in the network is followed by BN and a rectified linear unit (ReLU).

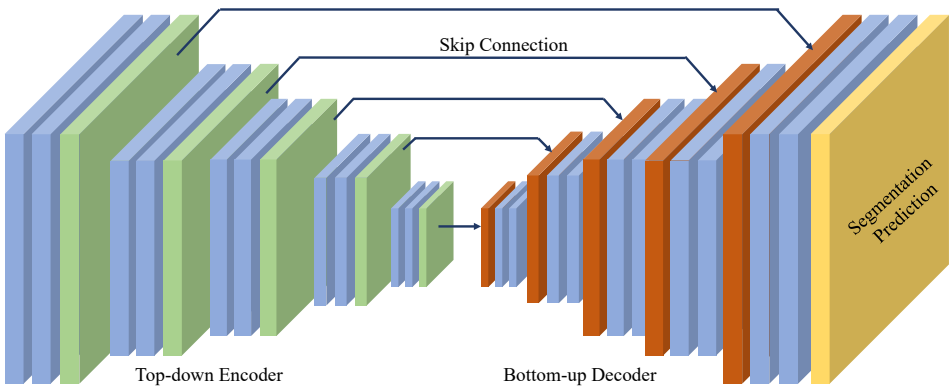


Figure 7. The architecture of the UNet.

It is worth noting that other 2D segmentation networks, such as FCN [11] and Deeplab V3+ [26], can also be used and achieve stable improvements. Thus, our proposed method is a plug-and-play framework that can be used to enhance most existing image-only-based methods. More details are provided in Section 4.

3.4. Uncertainty-Aware Cross-Modality Prediction Fusion Module

Since it is difficult to perform water segmentation in 2D in some scenes but very easy in 3D, a reasonable fusion of the 2D and 3D segmentation results can optimize the segmentation performance. Therefore, we propose an uncertainty-aware cross-modality

prediction fusion module to further refine the 2D and 3D water segmentation predictions to exploit the complementary properties of 2D images and 3D point clouds. In this module, the 2D and 3D water segmentation predictions are transformed into 3D and 2D dimensions, respectively, and fused based on the uncertainties in the different dimensions.

3.4.1. 2D Water Segmentation Refinement

For the 2D image water segmentation, we project the 3D segmentation result $O^{3D} \in R^{N\times 2}$ into the 2D domain $\hat{O}^{3D} \in R^{H\times W\times 2}$ using the 3D-to-2D projection module proposed in Section 3.2.1, and we densify the image-wise 3D results by dilation. Subsequently, a 2D uncertainty-aware weight module is used to calculate the uncertainty weight of the image-wise 2D and 3D segmentation results. The 2D uncertainty-aware weight module is inspired by the convolutional block attention module (CBAM) [39]. It is composed of two parts: the domain weight and the spatial weight modules, as shown in Figure 8. The domain weight module generates an uncertainty confidence interval for the overall segmentation results in 2D and 3D. The spatial weight module generates an uncertainty confidence interval spatially for the 2D and 3D individual segmentation results. Specifically, as shown in Figure 8, given the 2D segmentation prediction O^{2D} and image-wise 3D segmentation prediction $\hat{O}^{3D}$, the domain weight module concatenates the predictions and uses them as input to infer a 1D domain weight map $W_d^{2D} \in R^2$. After applying the domain weight map to the image-wise 2D and 3D segmentation results, the spatial weight is utilized to produce a 2D spatial weight map $W_s^{2D} \in R^{H\times W\times 2}$. Then, the overall image-wise uncertainty confidence weight $W^{2D} \in R^{H\times W\times 2}$ is generated as:

$$W^{2D} = W_d^{2D} \otimes W_s^{2D}, \tag{3}$$

where $\otimes$ denotes element-wise multiplication. During multiplication, the weight values are broadcasted accordingly.

Finally, based on the overall image-wise uncertainty confidence weight, the 2D water segmentation result is refined as:

$$\tilde{O}^{2D} = W^{2D}\left(O^{2D}, \hat{O}^{3D}\right), \tag{4}$$

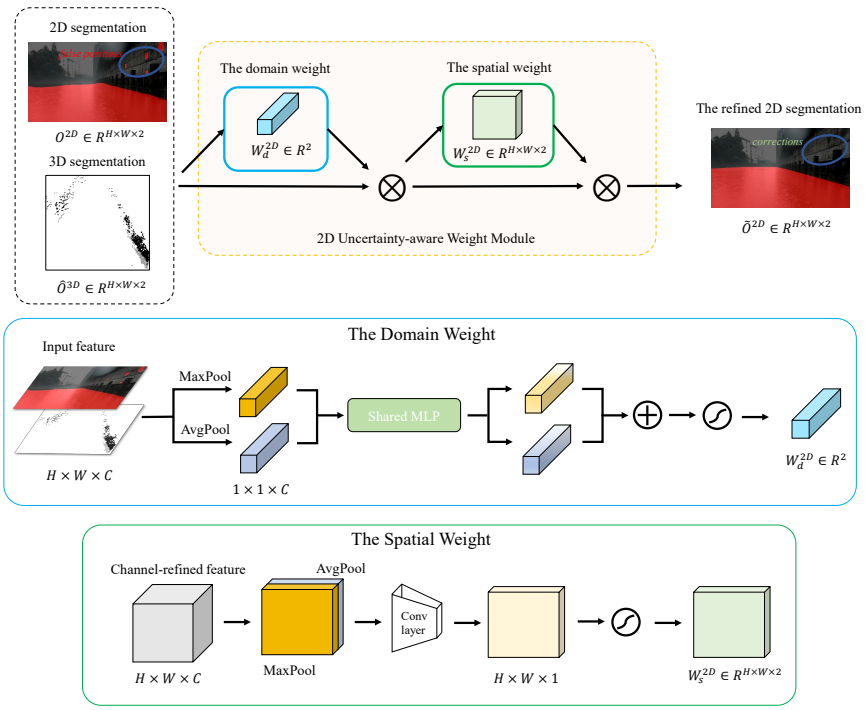


Figure 8. The details of the 2D uncertainty-aware cross-modality prediction fusion module. The false positives of 2D segmentation are represented by the red area in the blue circles.

3.4.2. 3D Water Segmentation Refinement

Water segmentation refinement is also performed in 3D, similar to the 2D water segmentation refinement. Specifically, the 2D segmentation prediction $O^{2D} \in R^{H \times W \times 2}$ is projected into the 3D domain $\hat{O}^{2D} \in R^{N \times 2}$ by the 2D-to-3D projection module described in Section 3.2.1. Then, a 3D self-attention module is utilized to calculate the uncertainty weight of the point-wise 2D and 3D segmentation results. As shown in Figure 9, the 3D self-attention module is implemented through an MLP consisting of five hidden layers with neuron sizes of 64, 128, 64 and 2. BN is used for all layers with a ReLU, and dropout layers are used for the last MLP. The point-wise 2D water segmentation prediction $\hat{O}^{2D}$ and 3D water segmentation prediction O^{3D} are passed through the 3D self-attention module to the obtain the overall point-wise uncertainty confidence weight $W^{3D} \in R^{N \times 2}$ as:

$$W^{3D} = \sigma\left(MLP\left(\hat{O}^{2D}, O^{3D}\right)\right), \tag{5}$$

where σ denotes the sigmoid function.

The 3D segmentation result is refined as:

$$\tilde{O}^{3D} = W^{2D}\left(\hat{O}^{2D}, O^{3D}\right), \tag{6}$$

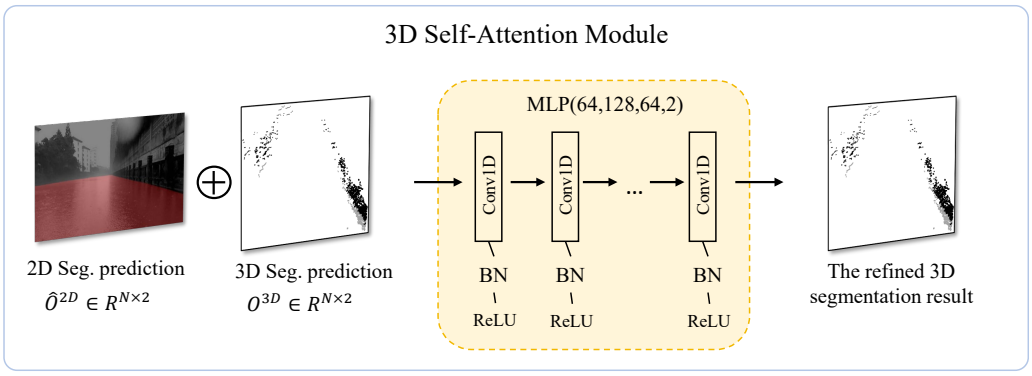


Figure 9. The details of the 3D self-attention module.

4. Experiments

We first present the novel Cross-modality Water Segmentation (CMWS) dataset —the first water segmentation dataset for USVs in inland waterways that has both images and corresponding point clouds. Then, extensive experiments are conducted on the CMWS dataset to validate the proposed method. The results indicate that our proposed method significantly improves those image-only-based methods. It is worth noting that all image-only-based methods show 2% improvements in accuracy and MaxF. Furthermore, our proposed method also results in accuracy improvements in 3D water segmentation.

4.1. Cross-Modality Water Segmentation Dataset

The CMWS dataset contains 3018 frames acquired under various weather and lighting conditions. Each frame has a 640 × 320 resolution image and its corresponding LiDAR point cloud. All frames were manually annotated and split into a training set, and a validation set, containing 1944 and 1074 pairs, respectively. As shown in Figure 10, this dataset is composed of two parts: the re-labeled data from the simultaneous localization and mapping (SLAM) task in USVInland [13] and the new labeled data acquired by the Jinghai USV [2] platform.

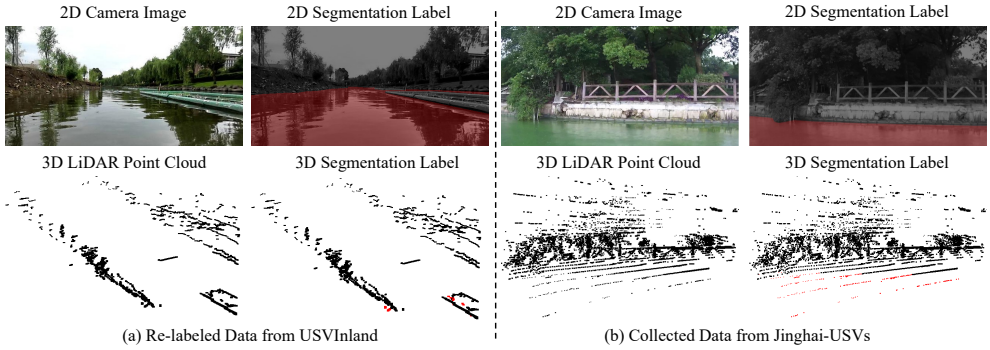


Figure 10. The data and labels in the CMWS dataset: the re-labeled data from the SLAM task in USVInland (**a**) and the new labeled data acquired by the Jinghai-USVs platform (**b**).

The USVInland dataset is the first multi-sensor USV dataset for inland waterways. This dataset was collected under various weather conditions for real driving scenarios on inland waterways. In this dataset, benchmarks for SLAM, stereo matching and water segmentation tasks were proposed. However, its water segmentation benchmark only has annotated image data but no corresponding LiDAR point cloud data. Thus, we focused on a SLAM task that uses images and corresponding LiDAR point cloud data. We re-annotated the data in the USVInland SLAM task to make them applicable to the water segmentation task.

Jinghai-USVs are a series of unmanned surface vehicles developed by Shanghai University, and a total of eight models have been developed so far. We collected the data with a Pandora sensor module and a GPS/IMU localization system on a Jinghai-USVs platform. The Pandora is an integrated sensor system consisting of a camera, LiDAR system and data processing system with advanced synchronization and calibration capabilities. The data were acquired under a diverse range of weather conditions, from sunny days to days with light rain. We carefully selected the representative sample data for manual annotation.

4.2. Experimental Settings

4.2.1. Parameter Settings

The experiments were implemented in PyTorch. The models were trained for 64 epochs with a batch size of 6 and the following training loss L:

$$L = L_{2D} + L_{2D}^{\text{refined}} + L_{3D} + L_{3D}^{\text{refined}}, \tag{7}$$

where L_{2D} and L_{3D} denote the weighted cross-entropy losses for the 2D and 3D segmentations before refinement, and L_{2D}^{refined} and L_{3D}^{refined} denote the weighted cross-entropy losses for the 2D and 3D segmentation after refinement, respectively.

The learning rate of the linear warm-up Radam [40] Lookahead [41] optimizer was initialized as 0.1 with a decay of 0.9 for every 10 epochs. The same data split and training settings were applied to all the following experiments. All experiments were implemented using a single NVIDIA 1080Ti GPU.

4.2.2. Evaluation Metrics

Similar to [13,42], we used the following eight metrics for performance evaluation: ***Accuracy*** (**ACC**), ***Max F-measure*** (**MaxF**), ***Average Precision*** (**AvgPrec**), ***Precision*** (**PRE**), ***Recall*** (**REC**), ***False Positive Ratio*** (**FPR**) and ***False Negative Ratio*** (**FNR**). These evaluation metrics can be calculated as follows:

$$Acc = \frac{TP + TN}{TP + FP + TN + FN}, \tag{8}$$

$$PRE = \frac{TP}{TP + FP}, \tag{9}$$

$$REC = \frac{TP}{TP + FN}, \tag{10}$$

$$\text{MaxF} = \text{argmax}_{\tau}\left[\frac{(1+\beta^2)PRE * REC}{\beta^2(PRE + REC)}\right], \tag{11}$$

$$\text{AvgPrec} = \frac{1}{11}\sum_{r\in[0,0.1,\ldots,1]}\max_{\tilde{r}:\tilde{r}>1} PRE(\tilde{r}), \tag{12}$$

$$\text{FPR} = \frac{FP}{FP + TN}, \tag{13}$$

$$\text{FNR} = \frac{FN}{FN + TP}, \tag{14}$$

where *TP* (true positive) represents the number of water surface pixels classified correctly; *TN* (true negative) represents the number of non-water surface pixels classified correctly; *FN* (false negative) indicates the number of water surface pixels wrongly classified as non-water surface; *FP* (false positive) denotes the number of non-water surface pixels wrongly classified as the water surface.

ACC represents the ratio of the number of correct predictions to the total number of samples. **PRE** is the rate of correct predictions among the samples predicted to be positive, and **REC** is the rate of correct predictions among all positive samples in the dataset. **MaxF** is a single score that balances the **PRE** and **REC** values. **AvgPrec** is the average of all precision values. **FPR** is the ratio of falsely predicted positive samples to all predicted positive samples, and **FNR** is the ratio of falsely predicted negative samples to all predicted negative samples. The higher the value of the **ACC**, **MaxF**, **AvgPrec**, **PRE** and **REC**, the better the performance of the algorithm is. The opposite is true for **FNR** and **FNR**.

4.3. Experimental Results

We compared the proposed method with the image-only-based methods, UNet [25], FCN [11] and DeeplabV3+ [26]. As described in Section 3.3.2, the proposed method is a plug-and-play framework enabling the use of any 2D segmentation network. Therefore, we used the proposed method to improve UNet, FCN and Deeplabv3+, respectively. As Table 1 shows, the incorporation of the 3D LiDAR point cloud information results in significant and stable improvements of all image-only-based methods, achieving about 2% improvements in accuracy and MaxF. Note that the more powerful FCN only achieves a 1.07% improvement in accuracy over UNet, and DeeplabV3+ only achieves a 0.95% improvement in accuracy over FCN. Compared with them, the proposed method brings twice as much improvement. In addition to the accuracy and MaxF metrics, the proposed method improves the PRE and REC metrics and reduces the FPR and FNR metrics for all image-only-based methods. This further illustrates that the proposed method reduces the misjudgment of water surface and non-water surface well by utilizing the 3D LiDAR point cloud information. In addition, the steady improvement in the AvgPrec metric for all image-only-based methods is also a good illustration of the stability of the improvement brought by the proposed method. Remarkably, the lightweight model UNet even outperforms the more powerful FCN and DeeplabV3+ models after being improved by the proposed method. The stronger the 2D segmentation network, the higher the performance of the proposed method is. The highest performance is achieved with the DeeplabV3+ model improved by the proposed method.

Table 1. Performance comparison of the proposed camera-LiDAR method and image-only-based methods on the CMWS dataset."I" denotes "images"; "L" denotes "LIDAR"; "CM-" denotes the enhancement by our proposed method. **Bold** denotes the best performance, and underline in red shows the improvement achieved by our proposed method.

Methods	Input	Acc(%)↑	MaxF(%)↑	AvgPrec(%)↑	PRE(%)↑	REC(%)↑	FPR(%)↓	FNR(%)↓
UNet [25]	I	94.31	94.85	92.03	93.59	96.14	6.41	3.86
FCN [11]	I	95.38	96.38	95.69	97.13	95.64	2.87	4.36
DeepLabV3+ [26]	I	96.33	96.69	95.50	96.54	96.83	3.46	3.17
CM-UNet (Ours)	L+I	96.77 *+2.46*	97.34 *+2.49*	95.86 *+3.83*	96.94 *+3.35*	97.75 *+1.61*	3.06 *−3.35*	2.25 *−1.61*
CM-FCN (Ours)	L+I	97.64 *+2.26*	97.83 *+1.45*	96.28 *+0.59*	97.78 *+0.65*	97.97 *+2.33*	2.22 *−0.65*	2.03 *−2.33*
CM-DeepLabV3+ (Ours)	L+I	**98.08** *+1.75*	**98.24** *+1.55*	**96.77** *+1.27*	**97.49** *+0.95*	**98.00** *+1.17*	**2.51** *−0.95*	**2.00** *−1.17*

Furthermore, we visualize the qualitative results of the proposed method and the image-only-based methods in Figure 11 (More qualitative results can be seen in Figures S1 and S2). Although the image-based methods can segment most of the water surface, they typically perform poorly for segmenting near-shore or reflective water surfaces. In contrast, our proposed method segments most of the water surface well and performs accurate segmentation in difficult cases, such as near-shore or reflective water.

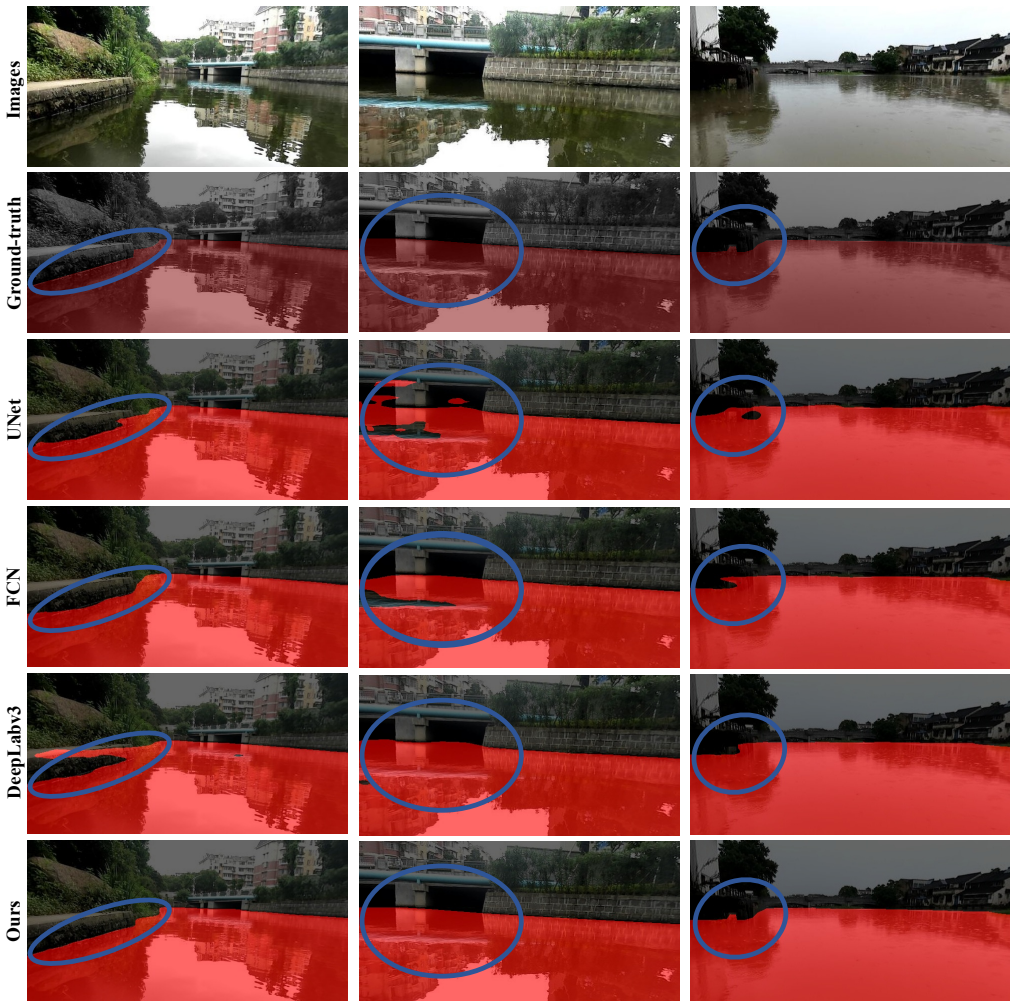


Figure 11. The qualitative results of our methods and those image-only-based methods.

4.4. Design Analysis

We conducted comprehensive ablation experiments on the CMWS dataset to validate the design of the proposed method. All ablation experiments were conducted on the model with the lightweight UNet as the 2D segmentation network.

4.4.1. Ablation Study

The ablation results for different modules are summarized in Table 2. In the baseline model (model A), training for water segmentation is conducted separately in 2D and 3D. The baseline achieves a MaxF of 78.91%, an AvgPrec of 79.38% and a PRE of 83.59% for 3D water segmentation, and it achieves a MaxF of 94.85%, an AvgPrec of 92.03% and a PRE of 93.59% for 2D water segmentation. The performance of 3D point cloud water segmentation (Model B) is greatly improved due to the incorporation of 2D image color and texture information in the early 3D cross-modality segmentation module. The MaxF is 86.13%, AvgPrec is 85.12% and PRE is 86.22%, showing a significant improvement. Incorporating the 3D point cloud segmentation results and geometric information into the 2D segmentation through the late 2D cross-modality segmentation module (Models C and D) significantly improves the performance of 2D water segmentation. Moreover, the higher the accuracy of the 3D point cloud segmentation, the greater the performance improvement of the 2D segmentation is. After implementing the uncertainty-aware cross-modality prediction fusion module, the proposed method achieves the best results for 2D and 3D water segmentation tasks.

Table 2. Results of different ablations on the CMWS dataset, in which "**3D CM**" denotes the "early 3D cross-modality segmentation module"; "**2D CM**" denotes the "late 2D cross-modality segmentation module"; "**Fusion**" denotes the "uncertainty-aware cross-modality prediction fusion module".

Model	3D CM	2D CM	Fusion	3D Segmentation			2D Segmentation		
				MaxF	AvgPrec	PRE	MaxF	AvgPrec	PRE
A				78.91	79.38	83.59	94.85	92.03	93.59
B	✓			86.13	85.12	86.22	-	-	-
C		✓		78.91	79.38	83.59	95.53	94.26	94.98
D	✓	✓		86.13	85.12	86.22	96.82	95.12	95.62
E	✓	✓	✓	**89.64**	**89.73**	**87.77**	**97.34**	**95.86**	**96.94**

4.4.2. Image-Wise Feature Fusion Strategy

Several feature fusion methods were used in the late 2D cross-modality segmentation module to fuse the 2D image with the projected image-wise 3D feature and segmentation results, such as Max fusion, Sum fusion, Concatenation fusion, and Conv fusion (i.e., CEN [19], CBAM [39]). Table 3 presents the results of the different fusion methods. Conv fusion provides the best results, and the Conv-CEN fusion outperforms the Conv-CBAM fusion. Thus, we choose the Conv-CEN fusion as the default fusion method.

Table 3. Results of different image-wise feature fusion strategies in the Late 2D Cross-modality Segmentation Module on the CMWS dataset.

Fusion Method	MaxF (%)	AvgPrec (%)	PRE (%)
Max	90.98	91.13	87.98
Sum	93.71	92.9	92.19
Concatenation	94.83	94.46	94.16
Conv-CBAM [39]	96.82	95.12	95.62
Conv-CEN [19]	97.34	95.86	96.94

4.4.3. Cross-Modality Prediction Fusion Strategy

We further evaluated different fusion methods for the 2D and 3D segmentation results. As Table 4 shows, the proposed uncertainty-aware fusion method significantly outperforms

all other fusion methods for the 2D and 3D segmentations. In the 3D water segmentation, the MaxF is 89.64%, AvgPrec is 89.73% and PRE is 87.77%.

Table 4. Results of different cross-modality prediction fusion strategies on the CMWS dataset.

Fusion Method	3D Segmentation			2D Segmentation		
	MaxF (%)	AvcgPrec (%)	PRE (%)	MaxF (%)	AvgPrec (%)	PRE (%)
Max	87.13	85.96	85.6	95.32	94.66	94.44
Sum	87.93	86.01	86.06	96.13	95.69	95.73
Mean	88.27	88.23	87.17	96.58	95.58	95.93
Uncertainty-aware	**89.64**	**89.73**	**87.77**	**97.34**	**95.86**	**96.94**

5. Conclusions

We proposed a novel camera-LiDAR cross-modal fusion water segmentation method to combine the 2D image and 3D point cloud characteristics to improve the accuracy of water segmentation. By exploiting the advantages of 2D images and 3D point clouds, the proposed method handles simple cases, such as large water surfaces well, and performs accurate segmentation in difficult cases, such as near-shore or reflective water. In addition, we collected, annotated and presented a novel CMWS dataset to validate the proposed method. To the best of our knowledge, this is the first water segmentation dataset for USVs that contains images and corresponding point clouds. Extensive experiments were conducted on the CMWS dataset, indicating that the proposed method is a plug-and-play framework that can significantly improve existing image-only-based methods. All image-only-based methods achieved improvements in accuracy and MaxF.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/jmse10060744/s1. Figure S1: More qualitative results of our methods and those image-only-based methods. Figure S2: More qualitative results of our methods.

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

Funding: This research was funded by the National Key Research and Development Program of China, grant number 2020YFC1521700; the Joint Founds of National Natural Science Foundation of China, grant number U1813217; and the National Natural Science Foundation of China, grant number 62073075.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Pastore, T.; Djapic, V. Improving autonomy and control of autonomous surface vehicles in port protection and mine countermeasure scenarios. *J. Field Robot.* **2010**, *27*, 903–914. [CrossRef]
2. Peng, Y.; Yang, Y.; Cui, J.; Li, X.; Pu, H.; Gu, J.; Xie, S.; Luo, J. Development of the USV 'JingHai-I'and sea trials in the Southern Yellow Sea. *Ocean Eng.* **2017**, *131*, 186–196. [CrossRef]
3. Madeo, D.; Pozzebon, A.; Mocenni, C.; Bertoni, D. A low-cost unmanned surface vehicle for pervasive water quality monitoring. *IEEE Trans. Instrum. Meas.* **2020**, *69*, 1433–1444. [CrossRef]
4. Wang, W.; Gheneti, B.; Mateos, L.A.; Duarte, F.; Ratti, C.; Rus, D. Roboat: An autonomous surface vehicle for urban waterways. In Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Macau, China, 3–8 November 2019; pp. 6340–6347.

5. Zhan, W.; Xiao, C.; Wen, Y.; Zhou, C.; Yuan, H.; Xiu, S.; Zou, X.; Xie, C.; Li, Q. Adaptive semantic segmentation for unmanned surface vehicle navigation. *Electronics* **2020**, *9*, 213. [CrossRef]
6. Mettes, P.; Tan, R.T.; Veltkamp, R. On the segmentation and classification of water in videos. In Proceedings of the 2014 International Conference on Computer Vision Theory and Applications (VISAPP), Lisbon, Portugal, 5–8 January 2014; Volume 1, pp. 283–292.
7. Achar, S.; Sankaran, B.; Nuske, S.; Scherer, S.; Singh, S. Self-supervised segmentation of river scenes. In Proceedings of the 2011 IEEE International Conference on Robotics and Automation, Shanghai, China, 9–13 May 2011; pp. 6227–6232.
8. Kristan, M.; Kenk, V.S.; Kovačič, S.; Perš, J. Fast image-based obstacle detection from unmanned surface vehicles. *IEEE Trans. Cybern.* **2015**, *46*, 641–654. [CrossRef] [PubMed]
9. Taipalmaa, J.; Passalis, N.; Zhang, H.; Gabbouj, M.; Raitoharju, J. High-resolution water segmentation for autonomous unmanned surface vehicles: A novel dataset and evaluation. In Proceedings of the 2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP), Pittsburgh, PA, USA, 13–16 October 2019; pp. 1–6.
10. Teichmann, M.; Weber, M.; Zoellner, M.; Cipolla, R.; Urtasun, R. Multinet: Real-time joint semantic reasoning for autonomous driving. In Proceedings of the 2018 IEEE Intelligent Vehicles Symposium (IV), Changshu, China, 26–30 June 2018; pp. 1013–1020.
11. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
12. Zhan, W.; Xiao, C.; Wen, Y.; Zhou, C.; Yuan, H.; Xiu, S.; Zhang, Y.; Zou, X.; Liu, X.; Li, Q. Autonomous visual perception for unmanned surface vehicle navigation in an unknown environment. *Sensors* **2019**, *19*, 2216. [CrossRef] [PubMed]
13. Cheng, Y.; Jiang, M.; Zhu, J.; Liu, Y. Are we ready for unmanned surface vehicles in inland waterways? The usvinland multisensor dataset and benchmark. *IEEE Robot. Autom. Lett.* **2021**, *6*, 3964–3970. [CrossRef]
14. Ma, F.; Cavalheiro, G.V.; Karaman, S. Self-supervised sparse-to-dense: Self-supervised depth completion from lidar and monocular camera. In Proceedings of the 2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–24 May 2019; pp. 3288–3295.
15. Ma, F.; Karaman, S. Sparse-to-dense: Depth prediction from sparse depth samples and a single image. In Proceedings of the 2018 IEEE international conference on robotics and automation (ICRA), Brisbane, Australia, 21–25 May 2018; pp. 4796–4803.
16. Xiao, L.; Wang, R.; Dai, B.; Fang, Y.; Liu, D.; Wu, T. Hybrid conditional random field based camera-LIDAR fusion for road detection. *Inf. Sci.* **2018**, *432*, 543–558. [CrossRef]
17. Caltagirone, L.; Bellone, M.; Svensson, L.; Wahde, M. LIDAR–camera fusion for road detection using fully convolutional neural networks. *Robot. Auton. Syst.* **2019**, *111*, 125–131. [CrossRef]
18. Bovcon, B.; Muhovič, J.; Perš, J.; Kristan, M. The mastr1325 dataset for training deep usv obstacle detection models. In Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Macau, China, 3–8 November 2019; pp. 3431–3438.
19. Wang, Y.; Huang, W.; Sun, F.; Xu, T.; Rong, Y.; Huang, J. Deep multimodal fusion by channel exchanging. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 4835–4845.
20. Cheng, B.; Schwing, A.; Kirillov, A. Per-pixel classification is not all you need for semantic segmentation. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 17864–17875.
21. Mohan, R.; Valada, A. Efficientps: Efficient panoptic segmentation. *Int. J. Comput. Vis.* **2021**, *129*, 1551–1579. [CrossRef]
22. Chen, Z.; Duan, Y.; Wang, W.; He, J.; Lu, T.; Dai, J.; Qiao, Y. Vision Transformer Adapter for Dense Predictions. *arXiv* **2022**, arXiv:2205.08534.
23. Lopez-Fuentes, L.; Rossi, C.; Skinnemoen, H. River segmentation for flood monitoring. In Proceedings of the 2017 IEEE International Conference on Big Data (Big Data), Boston, MA, USA, 11–14 December 2017; pp. 3746–3749.
24. Ling, G.; Suo, F.; Lin, Z.; Li, Y.; Xiang, J. Real-time Water Area Segmentation for USV using Enhanced U-Net. In Proceedings of the 2020 Chinese Automation Congress (CAC), Shanghai, China, 6–8 November 2020; pp. 2533–2538.
25. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
26. 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.
27. Geiger, A.; Lenz, P.; Urtasun, R. Are we ready for autonomous driving? the kitti vision benchmark suite. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 3354–3361.
28. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3213–3223.
29. Caesar, H.; Bankiti, V.; Lang, A.H.; Vora, S.; Liong, V.E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; Beijbom, O. nuscenes: A multimodal dataset for autonomous driving. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11621–11631.
30. Sun, P.; Kretzschmar, H.; Dotiwalla, X.; Chouard, A.; Patnaik, V.; Tsui, P.; Guo, J.; Zhou, Y.; Chai, Y.; Caine, B.; et al. Scalability in perception for autonomous driving: Waymo open dataset. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 2446–2454.

31. Robicquet, A.; Sadeghian, A.; Alahi, A.; Savarese, S. Learning social etiquette: Human trajectory understanding in crowded scenes. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 549–565.
32. Barekatain, M.; Martí, M.; Shih, H.F.; Murray, S.; Nakayama, K.; Matsuo, Y.; Prendinger, H. Okutama-action: An aerial view video dataset for concurrent human action detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 28–35.
33. Du, D.; Qi, Y.; Yu, H.; Yang, Y.; Duan, K.; Li, G.; Zhang, W.; Huang, Q.; Tian, Q. The unmanned aerial vehicle benchmark: Object detection and tracking. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 370–386.
34. Mandal, M.; Kumar, L.K.; Vipparthi, S.K. Mor-uav: A benchmark dataset and baselines for moving object recognition in uav videos. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle WA, USA, 12–16 October 2020; pp. 2626–2635.
35. Qi, C.R.; Yi, L.; Su, H.; Guibas, L.J. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Adv. Neural Inf. Process. Syst.* **2017**, *30*. [CrossRef]
36. Altman, N.S. An introduction to kernel and nearest-neighbor nonparametric regression. *Am. Stat.* **1992**, *46*, 175–185.
37. Qi, C.R.; Su, H.; Mo, K.; Guibas, L.J. Pointnet: Deep learning on point sets for 3d classification and segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 652–660.
38. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NA, USA, 27–30 June 2016; pp. 770–778.
39. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
40. Liu, L.; Jiang, H.; He, P.; Chen, W.; Liu, X.; Gao, J.; Han, J. On the variance of the adaptive learning rate and beyond. *arXiv* **2019**, arXiv:1908.03265.
41. Zhang, M.; Lucas, J.; Ba, J.; Hinton, G.E. Lookahead optimizer: K steps forward, 1 step back. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 9597–9608.
42. Fritsch, J.; Kuehnl, T.; Geiger, A. A new performance measure and evaluation benchmark for road detection algorithms. In Proceedings of the 16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013), The Hague, The Netherlands, 6–9 October 2013; pp. 1693–1700.