

WaterScenes: A Multi-Task 4D Radar-Camera Fusion Dataset and Benchmarks for Autonomous Driving on Water Surfaces

Shanliang Yao^{1,*}, Runwei Guan^{1,*}, Zhaodong Wu², Yi Ni², Zile Huang², Ryan Wen Liu³, Yong Yue², Weiping Ding⁴, *Senior Member, IEEE*, Eng Gee Lim², *Senior Member, IEEE*, Hyungjoon Seo¹, Ka Lok Man², Jieming Ma², Xiaohui Zhu^{2,†}, Yutao Yue^{5,†}

Abstract—Autonomous driving on water surfaces plays an essential role in executing hazardous and time-consuming missions, such as maritime surveillance, survivor rescue, environmental monitoring, hydrography mapping and waste cleaning. This work presents WaterScenes, the first multi-task 4D radar-camera fusion dataset for autonomous driving on water surfaces. Equipped with a 4D radar and a monocular camera, our Unmanned Surface Vehicle (USV) proffers all-weather solutions for discerning object-related information, including color, shape, texture, range, velocity, azimuth, and elevation. Focusing on typical static and dynamic objects on water surfaces, we label the camera images and radar point clouds at pixel-level and point-level, respectively. In addition to basic perception tasks, such as object detection, instance segmentation and semantic segmentation, we also provide annotations for free-space segmentation and waterline segmentation. Leveraging the multi-task and multi-modal data, we conduct benchmark experiments on the uni-modality of radar and camera, as well as the fused modalities. Experimental results demonstrate that 4D radar-camera fusion can considerably improve the accuracy and robustness of perception on water surfaces, especially in adverse lighting and weather conditions. WaterScenes dataset is public on <https://waterscenes.github.io>.

Index Terms—Autonomous driving, multi-task, 4D radar-camera fusion, unmanned surface vehicle.

I. INTRODUCTION

AUTONOMOUS driving techniques are developing rapidly in recent years, achieving safer, more efficient, and more sustainable transportation across roads, skies,

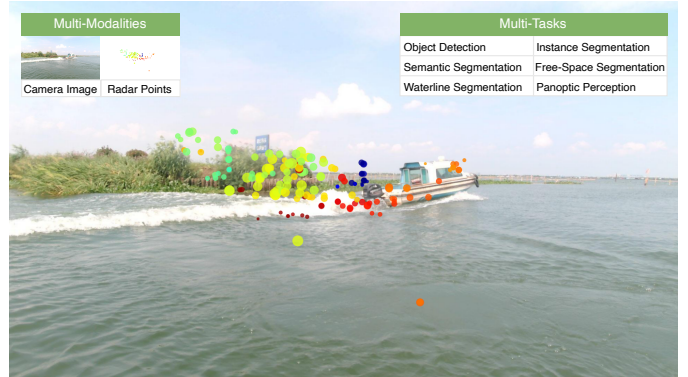


Fig. 1. Example scenario from our WaterScenes dataset. For each radar point on the image, the color denotes the range, and the size represents reflected power from the target.

and water surfaces [1]–[3]. Different scenarios offer unique prospects and challenges for autonomous driving vehicles. Unmanned Surface Vehicles (USVs) that navigate on water surfaces offer a versatile and cost-effective solution for various tasks, including coastal surveillance, environmental monitoring, river modeling, underwater detection, river rescue, and waste cleaning [4]–[7].

Compared to autonomous driving on road surfaces, perception challenges encountered on water surfaces are more daunting and unpredictable. Wind and waves significantly influence the stability of USVs, making it challenging for them to maintain desired heading and trajectory. The vibrations produced by USVs have a deleterious effect on sensor output, resulting in blurred transitions from the water to the sky or even object missing in the field of view [8]–[10]. Cameras may be disturbed by water splashes during navigation or water vapor formed due to temperature differences, leading to blurred or obscured images [11], [12]. To further complicate matters, floating debris (e.g., fallen leaves, water plants) along with the rippling caused by waterdrops on water surfaces are distractions to objects of interest. Mirror-like reflections of water surfaces are challenging to discern between virtual and actual objects. Adverse lighting and weather conditions significantly impact visibility, further diminishing the clarity of the images [13]–[15]. These manifold factors present a series of challenges to the camera sensor, making it difficult to detect and track objects in their surroundings. Although LiDARs

¹ Shanliang Yao, Runwei Guan and Hyungjoon Seo are with Faculty of Science and Engineering, University of Liverpool, Liverpool, UK. (email: {shanliang.yao, runwei.guan, hyungjoon.seo}@liverpool.ac.uk).

² Zhaodong Wu, Yi Ni, Zile Huang, Yong Yue, Eng Gee Lim, Ka Lok Man, Jieming Ma and Xiaohui Zhu are with School of Advanced Technology, Xi'an Jiaotong-Liverpool University, Suzhou, China. (email: {zhaodong.wu20, yi.ni21, zile.huang21}@student.xjtlu.edu.cn. {yong.yue, enggee.lim, ka.man, jieming.ma, xiaohui.zhu}@xjtlu.edu.cn).

³ Ryan Wen Liu is with School of Navigation, Wuhan University of Technology, Wuhan 430063, China, and also with the State Key Laboratory of Maritime Technology and Safety, Wuhan 430063, China (email: wenliu@whut.edu.cn).

⁴ Weiping Ding is with School of Information Science and Technology, Nantong University, Nantong 226019, China. (email: dwp9988@163.com).

⁵ Yutao Yue is with Thrust of Artificial Intelligence and Thrust of Intelligent Transportation, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou 511400, China. (email: yutaoyue@hkust-gz.edu.cn).

* Equal contribution

† Corresponding author: xiaohui.zhu@xjtlu.edu.cn, yutaoyue@hkust-gz.edu.cn

can assist in detection accuracy, they are also susceptible to adverse weather conditions [16], [17]. Moreover, LiDARs are limited by high waves and water reflectivity when applied to water environments [18], [19].

Unlike camera and LiDAR sensors, radar sensors emit radio waves that bounce off objects and return to the sensor, providing information about the object’s range, velocity, and azimuth angle. The ability of radar waves to penetrate severe weather conditions with minimal attenuation enables radar sensors to detect objects through rain, fog, and snow [20]–[22]. Moreover, the longer wavelength of radar signals makes them less susceptible to interference from adverse lighting conditions, including strong sunlight and darkness [23]. Furthermore, radar sensors can detect objects at long distances and even obstacles behind walls, providing the vehicle with early warning of potential obstacles or hazards [24]. All these advantages make radar sensors a reliable and robust component in autonomous driving vehicles, equally suitable for overcoming challenges on water surfaces. Nevertheless, conventional radars possess low resolution and lack semantic information about the detected objects [25], [26]. When applied on water surfaces, they produce weak echoes from non-metallic targets, along with clutter returned from the water environments [5], [27].

Therefore, a multi-modal sensor fusion approach that combines the strengths of radar and camera sensors is a potential solution to overcome the challenges and provide a comprehensive understanding of water surface perception. Numerous studies have demonstrated that multi-sensor fusion can overcome the shortcomings of each individual sensor, improving the overall scene understanding for intelligent transportation systems [28]–[30]. Radar-camera fusion, a typical representative in multi-sensor fusion, has also received considerable attention, demonstrating improved accuracy and robustness of models for autonomous driving vehicles on roads [31]–[34]. However, few works focus on radar-camera fusion on water surfaces, primarily due to the lack of available multi-modal datasets. To the best of our knowledge, FloW [27] is the only dataset that contains both radar and camera data for USVs. As there is only one category named “bottle” in FloW dataset, it is unsuitable for the complex water surface environment in real scenarios.

In recent years, 4D radar has shown its advantages in denser radar point clouds and higher angle resolution, providing richer information about the target. Thus, it is a potential perception sensor on USVs, tackling the unique challenges on water surfaces, such as surface reflections, adverse lighting and weather conditions. An increasing number of 4D radar-camera fusion datasets (e.g., Astyx [35], K-Radar [36], VoD [37] and TJ4DRadSet [38]) have emerged for autonomous driving on roads and proved to be effective in improving the accuracy of detection [39], [40]. However, there is no public 4D radar dataset for water surfaces till now, let alone a fused 4D radar and camera dataset. As shown in Fig. 1, our proposed dataset fills this gap with the following contributions:

- We present WaterScenes, the first multi-task 4D radar-camera fusion dataset on water surfaces, which offers data from multiple sensors, including a 4D radar, monocular

camera, GPS, and IMU. It can be applied in six perception tasks, including object detection, instance segmentation, semantic segmentation, free-space segmentation, waterline segmentation, and panoptic perception.

- Our dataset covers diverse time conditions (daytime, nightfall, night), lighting conditions (normal, dim, strong), weather conditions (sunny, overcast, rainy, snowy) and waterway conditions (river, lake, canal, moat). An information list is also offered for retrieving specific data for experiments under different conditions.
- We provide 2D box-level and pixel-level annotations for camera images, and 3D point-level annotations for radar point clouds. We also offer a toolkit¹ for WaterScenes that includes pre-processing, labeling, projection and visualization, assisting researchers in processing and analyzing our dataset.
- We build corresponding benchmarks and evaluate popular algorithms for object detection, point cloud segmentation, image segmentation, and panoptic perception. Experiments demonstrate the advantages of radar perception on water surfaces, particularly in adverse lighting and weather conditions.

The rest of our study is organized as follows: Section II reviews the related datasets on water surfaces, highlighting the significance of our WaterScenes. Section III offers detailed insights into the proposed dataset, including USV setup, data collection, data processing, and dataset analysis. Section IV and Section V present benchmark experiments to evaluate the dataset, along with discussions on challenges and potential research directions. Lastly, in Section VI, we summarize our study and provide an outlook for future works.

II. RELATED DATASETS

Table I gives an overview of public datasets related to water surfaces. MODD [41] dataset specifically focuses on obstacle detection in marine environments. It contains 12 marine video sequences, each manually labeled with water edges and obstacles. The specific obstacle classification does not refine the objects in each category, but only classifies them into large and small obstacles. Objects that straddle the water edge are marked as large obstacles, while those entirely located below the water edge are marked as small obstacles. MODD2 [42], an extended version of MODD, provides synchronized IMU data to assist obstacle detection. Additionally, this dataset includes stereo images, which can be used for stereo verification to further enhance the detection performance. SMD [43] contains more specific obstacle categories, including ferry, ship, vessel, speed boat, and sail boat, acquired from both shore and boat. Besides, some data are captured from a near-infrared camera, which can provide image data in low light or even dark conditions. MaSTr1325 [4] is a marine semantic segmentation dataset, consisting of 1,325 samples and four pixel-level categories, namely obstacle, water, sky and ignore region. Moreover, MODS [3] dataset provides annotations for both detection and segmentation tasks. In this dataset, dynamic obstacles (vessel, person and other) are

¹<https://github.com/WaterScenes/WaterScenes>

TABLE I

OVERVIEW OF PUBLIC DATASETS ON WATER SURFACES. (\dagger) DENOTES THE NUMBER OF CLASSES IN THE DETECTION TASK. (-) INDICATES THAT NO INFORMATION IS PROVIDED IN THE DATASET. OD: OBJECT DETECTION, LS: WATERLINE SEGMENTATION, OT: OBJECT TRACKING, SS: SEMANTIC SEGMENTATION, FS: FREE-SPACE SEGMENTATION, PS: PANOPTIC SEGMENTATION, IS: INSTANCE SEGMENTATION, PP: PANOPTIC PERCEPTION.

Name	Year	Camera	Radar	GPS, IMU	Tasks	Annotations	Classes \dagger	Annotated Frames	Adverse Lighting	Adverse Weather
MODD [41]	2015	Mono	-	-	OD, LS	2D Box, 2D Line	2	4,454	✓	-
MODD2 [42]	2018	Stereo	-	GPS, IMU	OD, LS	2D Box, 2D Line	2	11,675	✓	✓
SMD [43]	2019	Mono	-	-	OD, OT	2D Box	10	31,653	✓	-
MaStr1325 [4]	2019	Mono	-	IMU	SS	2D Pixel	4	1,325	✓	✓
MODS [3]	2021	Stereo	-	IMU	OD, SS	2D Box, 2D Line	3	24,090	✓	✓
MID [44]	2021	Mono	-	-	OD	2D Box	2	2,655	✓	✓
USVInland [19]	2021	Stereo	-	GPS, IMU	SS, FS	2D Line	1	700	✓	✓
FloW [27]	2021	Mono	3D	-	OD	2D Box	1	2,000	✓	-
LaRS [45]	2023	Mono	-	-	SS, PS	2D Line	11	4,006	-	-
MVDD13 [46]	2024	Mono	-	-	OD	2D Box	13	35,474	✓	✓
WaterScenes (Ours)	2023	Mono	4D	GPS, IMU	OD, IS, SS, FS, LS, PP	2D Box, 2D Pixel, 2D Line, 3D Point	7	54,120	✓	✓

annotated with bounding boxes, while static obstacles (shore and pier) are annotated by water-obstacle boundaries. MID [44] serves as a complementary dataset to the MODD [41] by capturing data in different severe weather conditions that coastal USVs may encounter. MVDD13 [46] dataset contains 13 categories, covering various types of marine vessels in both military and civilian fields. Realistic situations such as class proportions, image diversity, sample independence, and background clutter are considered in MVDD13, thus providing in-depth information for training and testing robust detectors.

The aforementioned datasets are tailored toward the marine environment, which predominantly features vast expanses of water. Conversely, inland rivers are characterized by their narrow and complex shapes, as well as diverse objects present on their surfaces. Introducing a dataset specifically geared towards inland USVs, USVInland [19] dataset serves as a resource for multiple tasks, including SLAM, stereo matching, and water segmentation. Unlike prior datasets such as MODD [41] and MODD2 [42], which solely traced the periphery of the water, USVInland provides comprehensive annotation of the entire water area via polygons. LaRS [45] is a large maritime panoptic obstacle segmentation dataset, capturing data from lakes, rivers and seas. Its excellence lies in the diversity of recording locations, scene types, obstacle categories, and acquisition conditions.

To draw attention to floating debris cleaning in inland waterways, FloW [27] dataset is proposed for floating waste detection using both camera and radar sensors. The benchmark in this dataset demonstrated the effectiveness of radar sensors in detecting small objects and their potential for application on water surfaces. However, this dataset has only one category (bottle), and the detection range for the radar sensor is limited to 14.5 meters.

III. WATERSCENES DATASET

As can be intuitively seen from Fig. 2, our WaterScenes provides multi-modal and multi-task data for autonomous driving on various water surface scenarios. Information about the WaterScenes is summarized in Table I, including equipped sensors, perception tasks, annotation types and collection

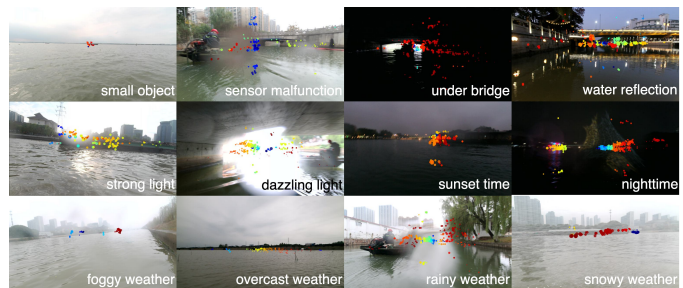


Fig. 2. Samples in WaterScenes. Radar points are projected onto the image plane as colored dots.

conditions. In this section, we present the process of creating this dataset and provide a statistical analysis of its contents.

A. USV Setup

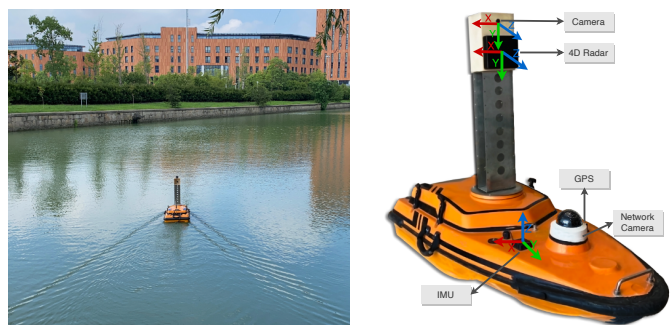


Fig. 3. Sensor suite for our USV and coordinate system of each sensor.

Our USV for data collection is equipped with various sensors, including a 4D radar for capturing radar point clouds, a monocular camera for gathering image information, a network camera for 360-degree observation, a GPS for geographical location information, and an IMU for tracking posture and motion information. The arrangement of these sensors on the USV is illustrated in Fig. 3, with each sensor's origin and direction denoted by different colors in the coordinate systems.

The detailed specifications of each sensor mounted on our USV are outlined in Table II.

TABLE II
SPECIFICATIONS OF SENSORS EQUIPPED ON OUR USV.

Sensor	Details
Radar	Oculii EAGLE 77GHz Point Cloud Radar, Medium Range Mode: 200 m detection range, 0.43 m range resolution, 0.27 m/s velocity resolution, $< 1^\circ$ azimuth/elevation angle resolution, 110° HFOV, 45° VFOV, 15Hz capture frequency
Camera	SONY IMX317 CMOS sensor, RGB color, 1920×1080 resolution, 100° HFOV, 60° VFOV, 30Hz capture frequency
GPS	latitude, longitude and altitude coordinates, < 2.5 m position accuracy, < 0.1 m/s velocity accuracy, 10Hz update rate
IMU	10-axis inertial navigation ARHS (3-axis gyroscope, 3-axis accelerometer, 3-axis magnetometer and a barometer), 0.5° heading accuracy, 0.1° roll/pitch accuracy, 50Hz update rate

B. Data Collection

Our dataset is collected from June to December 2022 in Suzhou, China. As the objects and surrounding environments vary across different water conditions, we select various waterways for data collection, such as small and large rivers, lakes, canals and moats. In order to capture high-quality data, we employ two distinct control methods during the data collection process. The first method utilizes our custom-designed software to create a precise navigation path, allowing the USV to travel to a specific location while recording data without human intervention. The second method involves remote control, which is used to acquire data for specific objects from multiple viewpoints. We focus on common objects of interest on water surfaces, including static objects such as piers and buoys, and dynamic objects such as ships, boats, vessels, kayaks, and sailors aboard these surface vehicles.

Meanwhile, to ensure the diversity and comprehensiveness of the dataset, we collect data across different waterways under different time conditions (e.g., daytime, nightfall and night), diverse lighting (e.g., normal, dim and strong) and weather conditions (e.g., sunny, overcast, rainy and snowy). We also document scenarios of sensor malfunction, including instances where waterdrops adhere to the camera lens, resulting in obscured images, as well as situations where radar connectivity is lost, rendering it impossible to capture point cloud data. These records are significant as they reflect real-world challenges that are likely to arise during autonomous driving.

C. Processing and Annotation

Following the processing approach from the nuScenes dataset [47], we extract image keyframes at a rate of 2Hz. The radar, GPS, and IMU data are then synchronized with the image keyframes based on the closest timestamp, with a maximum tolerated time difference of 0.05 seconds [37]. Each image in the dataset is manually annotated by human annotators and is further validated by domain experts. In the object detection task, seven categories (pier, buoy, sailor, ship, boat, vessel and kayak) are enclosed in each image by 2D

bounding boxes. For the instance segmentation task, an additional category named free-space is labeled using polygonal masks, which indicates drivable areas for USVs. Annotations for semantic segmentation and free-space segmentation are later generated using the instance segmentation labels. To facilitate the waterline segmentation task, we draw lines that mark the boundary between water and land. In addition to annotating the class for each object, we also label the attributes (such as waterways, time conditions, lighting conditions, and weather conditions) for each frame in an information list, which facilitates the retrieval of specific data and the selection of desired data for experiments.

Annotation process for radar point clouds is extremely complicated and tedious, while annotation precision is essential to model training. Each point within the radar point clouds comprises various attributes, including range, Doppler velocity, azimuth angle, elevation angle, and reflected power. To establish the relationship between radar point clouds and camera images, we convert radar point clouds from Polar coordinates onto the image plane utilizing the extrinsic matrix between the radar sensor and camera sensor as well as the intrinsic matrix of the camera sensor [48]. With coordinates of radar point clouds corresponding to the image plane, we annotate point clouds within the image bounding box as the same category as the box. However, it should be noted that while this approach can provide some initial annotations for the radar point clouds, these annotations may not always be accurate due to the nature of radar detection. Radar point clouds may not consistently map onto objects and may detect targets behind them [49], [50]. Therefore, annotations are refined by domain experts based on projections derived from front and bird’s eye views, along with attributes (reflected power and Doppler velocity) of each point. Finally, every point within the radar point clouds is assigned a class label and an instance identification. Furthermore, we include radar data from three and five consecutive frames in WaterScenes, providing valuable resources for analyzing multi-frame accumulation techniques.

D. Dataset Statistics

WaterScenes dataset includes 54,120 sets of RGB images, radar point clouds, GPS and IMU data, covering over 200,000 objects. The specific number of frames and objects for each class is shown in Table III. Additionally, as an essential part of this dataset, images captured under unfavorable lighting and weather conditions are enumerated in the table. All images are in 1920×1080 pixels, containing a diverse range of objects, including piers, buoys, sailors, ships, boats, vessels and kayaks. Among the categories, buoys and piers are noticeable obstacles on water that USVs should avoid while driving, whereas ships, boats, vessels and kayaks represent common watercraft encountered on water surfaces. The term “sailor” specifically refers to the individuals on these watercraft.

We classify objects based on their size as follows: those with an area greater than 192×192 pixels are considered large, those with an area less than 32×32 pixels are deemed tiny, objects with an area between 32^2 and 64^2 pixels are referred to

TABLE III

DATASET STATISTICS. NUMBER OF ANNOTATED FRAMES (TOP), NUMBER OF OBJECTS (MIDDLE), AND PERCENTAGE OF OBJECTS BELONGING TO EACH CLASS COMPARED TO THE TOTAL NUMBER OF OBJECTS (BOTTOM). (\dagger) FREE-SPACE CLASS IS INCLUDED IN INSTANCE SEGMENTATION, SEMANTIC SEGMENTATION AND PANOPTIC PERCEPTION TASKS. ($\dagger\dagger$) WATERLINE ANNOTATIONS ARE IN WATERLINE SEGMENTATION AND PANOPTIC PERCEPTION TASKS.

	Total	Pier	Buoy	Sailor	Ship	Boat	Vessel	Kayak	Free-Space \dagger	Waterline $\dagger\dagger$	Adverse Lighting	Adverse Weather
Frames	54,120	25,787	3,769	3,613	19,776	9,106	9,362	366	54,057	53,926	5,604	10,729
Objects	202,807	121,827	16,538	8,036	34,121	10,819	11,092	374	54,057	159,901	30,517	46,784
Percentage		(60.07%)	(8.15%)	(3.96%)	(16.82%)	(5.33%)	(5.47%)	(0.18%)			(15.05%)	(23.07%)

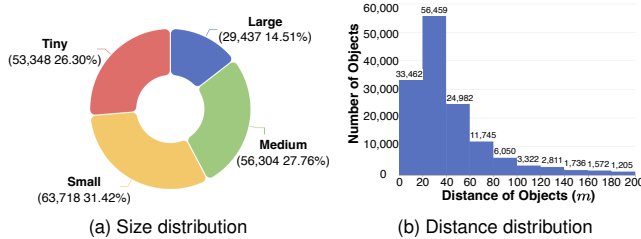


Fig. 4. Statistics of objects in WaterScenes. (a) Wide range of object size. (b) Wide distribution of object distance.

as small, and those between 64^2 and 192^2 pixels are classified as medium. The size distribution depicted in Fig. 4(a) reveals a wide range of object sizes, consistent with the diverse sizes of objects typically observed on water surfaces. We also analyze the distance distribution using the range attribute in radar point clouds. Fig. 4(b) demonstrates the relationship between the number of objects and distance at intervals of every 20 meters.

TABLE IV
POINT CLOUD ATTRIBUTES FOR EACH CATEGORY.

Attribute	Pier	Buoy	Sailor	Ship	Boat	Vessel	Kayak
Points	8.45	14.53	4.75	81.23	38.51	80.32	6.72
Power (dB)	13.68	17.88	12.15	14.40	14.14	13.52	10.12
Velocity (m/s)	0.08	0.09	0.79	1.08	0.40	2.21	0.88

Furthermore, we conduct a comprehensive analysis of the radar point clouds by calculating the average values of attributes for each specific class. As illustrated in Table IV, the number of points is highly correlated with object size. In particular, ships and vessels, being large objects, have the highest number of points, while sailors and kayaks, being small objects, have few points. Reflected power is similar for piers, ships, boats and vessels as they are primarily composed of cement. Buoys have higher power values as they are made of metal materials, while kayaks are composed of plastic materials with low power values. Velocity information is also instrumental in distinguishing between different types of objects. For example, stationary targets such as piers and buoys exhibit minimal velocity, while ships and vessels have relatively higher velocities. Above all, each attribute represents distinct target characteristics and is crucial for point cloud classification.

IV. BENCHMARKS

In this section, our WaterScenes serves as benchmarks for evaluating the performance on multiple tasks on water surfaces, including object detection, radar point cloud segmentation, camera image segmentation and panoptic perception. By analyzing the experimental results, we highlight the value, challenges and potential research directions posed by WaterScenes for further research.

A. Experimental Settings

After data processing and annotation, we divide the proposed dataset into three parts: a training set, a validation set, and a test set in the ratio of 7:2:1. All experiments are performed on two RTX 3090 GPUs with the training mode of data distributed parallel. All images in WaterScenes are resized to 640×640 pixels during the training phase. Results are evaluated on the test set in WaterScenes with Frames Per Second (FPS) assessed on a single RTX 3090 GPU.

Object Detection. We select five models for camera-based object detection with diverse paradigms (e.g., two-stage/one-stage, anchor-based/anchor-free, CNN-based/Transformer-based): CenterNet (ResNet-50) [51], Deformable DETR (ResNet-50) [52], Faster R-CNN (ResNet-50) [53], YOLOX-M [54] and YOLOv8-M [55]. We train these models from scratch with an initial learning rate of $1e-2$, accompanied by a cosine learning rate scheduler. We set the batch size to 32 and choose Adam as the optimizer with weight decay of $5e-4$. We also adopt Exponential Moving Average (EMA) to smooth model weights and mixed precision to speed up the training and reduce the CUDA memory.

For fusion-based object detection, we propose a generalized lightweight early fusion method for YOLOX-M and YOLOv8-M without altering their basic architectures. As depicted in Fig. 5, the detection process incorporates two input modalities: camera RGB images $C \in \mathbb{R}^{3 \times H \times W}$ and radar REVP maps $R \in \mathbb{R}^{4 \times H \times W}$. Specifically, as described in Algorithm 1, REVP maps capture the combined features of Range (R), Elevation (E), Velocity (V) and reflected Power (P) of the detected object from the radar point clouds matched to the image frame. The coordinate transformation process utilizes an extrinsic matrix accounting for the relative position and orientation of the radar and camera sensors. Subsequently, 3D coordinates in the camera frame are projected onto the 2D image plane using the camera's intrinsic matrix, yielding image plane coordinates (u, v) .

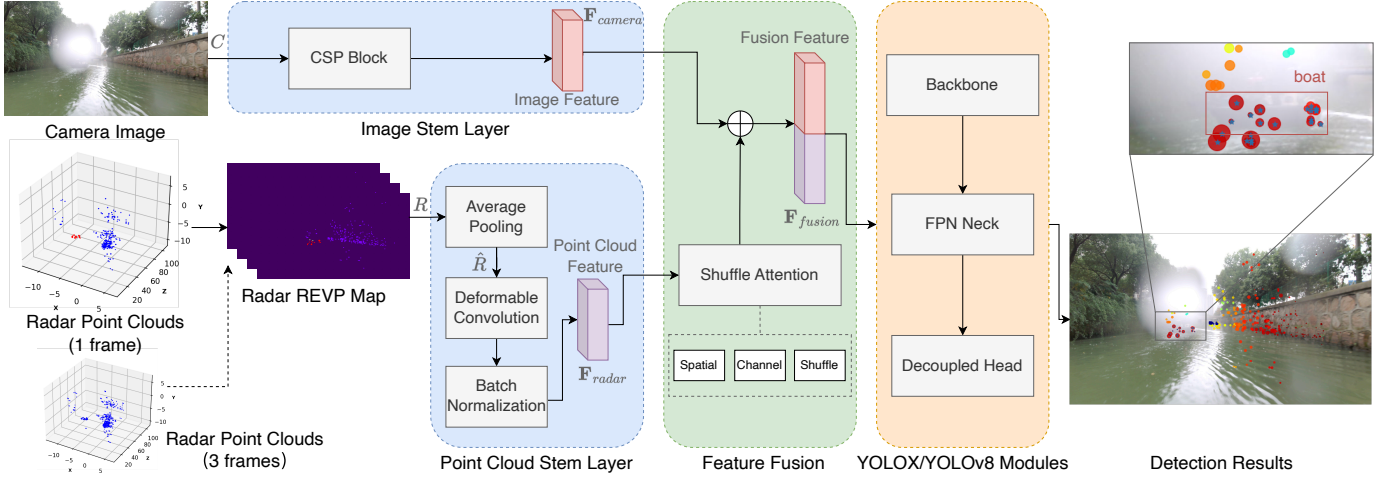


Fig. 5. Radar-camera fusion network for the detection benchmark on WaterScenes. Camera images and radar point clouds are fed into the stem layers for feature extraction. Subsequently, the extracted features are processed by the attention mechanism and added along the channel dimension before forwarding into YOLOX-M and YOLOv8-M modules. As a result, the fusion-based network successfully detects boats even when cameras are occluded by waterdrops.

In stem layers of the camera input, we follow the default settings of YOLOX [54] and YOLOv8 [55] to conduct the stem step for initial feature downsampling and channel expansion. We then obtain the shallow feature of image $\mathbf{F}_{camera} \in \mathbb{R}^{C \times \frac{H}{2} \times \frac{W}{2}}$, as is shown in Equation 1.

$$\mathbf{F}_{camera} = W_{stem}C, \quad (1)$$

where W_{stem} is the learnable weight of stem layer.

In the case of radar input, we first employ Average Pooling with a window size of 3×3 and padding value of 1 to rapidly aggregate sparse neighborhood point clouds \hat{R} (Equation 2). Subsequently, we utilize the Deformable Convolution [52] to extract the irregular radar point cloud features (Equation 3). Following this, we apply a Batch Normalization layer, resulting in the radar feature $\mathbf{F}_{radar} \in \mathbb{R}^{C \times \frac{H}{2} \times \frac{W}{2}}$.

$$\hat{R}_{i,j} = \frac{1}{9} \sum_{f=1}^3 \sum_{g=1}^3 R_{i+f-1,j+g-1}, \quad (2)$$

$$\mathbf{F}_{radar} = \sum_{k=1}^K w_k \cdot \hat{R}(p + p_k + \Delta p_k) \cdot \Delta m_k, \quad (3)$$

where K is the convolution kernel of the sampling location. In our experiments, we set $K = 9$ as we use a 3×3 kernel size. p is the pre-specified offset of feature map \hat{R} for K locations. Δp_k and Δm_k are the learnable offset and modulation scalar for k -th location, respectively.

$$\mathbf{F}_{fusion} = \mathbf{F}_{camera} + \alpha \cdot SA(\mathbf{F}_{radar}), \quad (4)$$

To mitigate the negative impacts of clutter in radar point clouds while focusing on object features, we apply shuffle attention [56] on \mathbf{F}_{radar} , which is a lightweight attention module combining spatial and channel attention. The shuffle operation enhances channel interaction and alleviates over-dependency between inter-layer channels. Moreover, considering that radar plays different roles in various scenarios,

Algorithm 1: Radar-Camera Fusion Algorithm

- 1: /* Prepare radar REVP maps */
 - Input:** Radar frames (1 frame or 3 frames), number of radar frames N_r ;
 - Output:** Radar REVP maps;
 - 2: $Features \leftarrow [Range, Elevation, Velocity, Power]$
 - 3: **for** $i \leftarrow 1$ **to** N_r **do**
 - 4: /* Project each radar point onto camera plane */
 - 5: $u, v \leftarrow$ coordinates for radar point on camera plane
 - 6: **for** $channel$ **in** $Features$ **do**
 - 7: $R_i[channel][u][v] \leftarrow$ radar point channel value
 - 8: **end for**
 - 9: **end for**
 - 10: /* Set the training stage */
 - Input:** Camera images with annotations, radar REVP maps;
 - Output:** Radar-camera fusion model;
 - 11: Number of epochs $N_e \leq 100$;
 - 12: **for** $i \leftarrow 1$ **to** N_e **do**
 - 13: /* Feature initialization of camera input C */
 - 14: Convolution: $\mathbf{F}_{camera} \leftarrow$ Equation 1
 - 15: /* Feature initialization of radar input R */
 - 16: Average Pooling: $\hat{R} \leftarrow$ Equation 2
 - 17: Deformable Convolution: $\mathbf{F}_{radar} \leftarrow$ Equation 3
 - 18: Batch Normalization: \mathbf{F}_{radar}
 - 19: /* Feature fusion upon camera and radar */
 - 20: Adaptive Feature Fusion: $\mathbf{F}_{fusion} \leftarrow$ Equation 4
 - 21: Run YOLOX/YOLOv8 modules
 - 22: **end for**
-

assisting the camera modality in some cases and struggling with noise interference in others, we introduce a learnable dynamic weight α to balance the importance of the current sample in the REVP map. The outputs from both branches are then element-wise added to generate fused features \mathbf{F}_{fusion} , as illustrated in Equation 4. After that, we follow the paradigms

of YOLOX and YOLOv8 for the backbone, neck and detection head. Overall, the pseudo-code of the proposed radar-camera fusion algorithm is presented in Algorithm 1.

Radar Point Cloud Segmentation. We select four point cloud processing models with various paradigms, including PointMLP [57], Point-NN [58], PointNet++ [59] and Point Transformer [60]. PointMLP [57], one of the State-Of-The-Art (SOTA) models in 3D point cloud processing, serves as the primary model for detailed analysis. We train all models from scratch with an initial rate of 5e-4, accompanied by a cosine learning rate scheduler. The batch size is 128 with AdamW as the optimizer and a weight decay of 5e-4. We employ the negative log-likelihood loss with focal [61] as the loss function.

Camera Image Segmentation. We select four classical models with various paradigms: DeepLabv3+ (atrous convolution with ASPP) [62], HRNet-W32 (multi-scale fusion with high-resolution features) [63], SegNeXt-B (convolution-attention-based) [64], SegFormer-B1 (self-attention-based) [65] and Mask2Former-R50 (transformer-based all-in-one model) [66] for image semantic segmentation; and another four models with different paradigms: YOLACT (two-stage of localization and segmentation) [67], SOLO (one-stage without localization) [68], YOLOv5-M (anchor-based) [69], YOLOv8-M (anchor-free) [55] and Mask2Former (transformer-based all-in-one model) [66] for image instance segmentation. We train these models from scratch with an initial learning rate of 9e-3, accompanied by a cosine learning rate scheduler. We adopt the dice loss for semantic segmentation and the focal loss for instance segmentation. We set the batch size to 32 and choose SGD as the optimizer with the weight decay of 1e-4 and momentum of 0.937. Moreover, we adopt mixed precision to accelerate the training process and reduce the CUDA memory.

Panoptic Perception. In our experiments, the panoptic perception includes tasks of object detection, free-space segmentation and waterline segmentation, covering an all-round perception of water surfaces. We evaluate the performance of panoptic perception on WaterScenes using two camera-based networks (YOLOP [70], HybridNets [71]) and one fusion-based network named Achelous [72]. YOLOP and HybridNets comprise one encoder for feature extraction and three decoders to handle the panoptic tasks. Achelous [72] is a lightweight panoptic perception framework dedicated to water surfaces. In Achelous, we select MobileViT [73] as the backbone and Ghost Dual-FPN (GDF) as the neck. Besides, we select Radar Convolution [72] to extract radar point cloud features. Furthermore, the homoscedastic-uncertainty-based learning strategy [74] is applied to assist multi-task learning. In the training stage, the detection head poses challenges in early convergence with an end-to-end strategy. Hence, following the approaches in [70] and [71], we first train the encoder and detection head for 100 epochs. Then, we freeze the encoder and detection head as well as train free-space and waterline segmentation heads for 50 epochs. Finally, the entire network is jointly trained for 50 epochs across all three tasks.

B. Metrics Settings

This section elaborates on the metrics utilized for evaluating WaterScenes across different tasks.

Object Detection. We adopt the mean Average Precision (mAP) with an Intersection-over-Union (IoU) threshold of 0.5, denoted as mAP₅₀, and the mAP with an IoU threshold range of 0.5 to 0.95, denoted as mAP₅₀₋₉₅. Mathematical formulations of these metrics are presented in Equation 8 and Equation 9, respectively, serving as quantitative indicators of a model’s effectiveness in detecting objects using bounding boxes.

$$P = \frac{TP}{TP + FP}, \quad (5)$$

$$R = \frac{TP}{TP + FN}, \quad (6)$$

$$AP = \int_0^1 P(r) dr, \quad (7)$$

$$\text{mAP}_{50} = \frac{1}{N} \sum_{i=1}^N \text{AP}_{50}^i, \quad (8)$$

$$\text{mAP}_{50-95} = \frac{1}{N} \sum_{i=1}^N \text{AP}_{50-95}^i. \quad (9)$$

P and R correspond to precision and recall as outlined in Equation 5 and Equation 6, respectively. Here, TP , FP and FN represent predicted samples of true positive, false positive, and false negative, respectively. AP symbolizes the average precision in Equation 7, where $P(r)$ denotes the precision on the recall-precision curve and r represents the recall. In the equation of mAP₅₀, AP_{50}^i stands for the AP value of class i targets with an IoU of 50% and above with ground truth in the predicted bounding boxes. mAP₅₀₋₉₅ denotes the average AP value of class i targets with an IoU ranging from 50% to 95% in the prediction box compared to the ground truth.

Radar Point Cloud Segmentation. We adopt Point Accuracy (PA) and mean Intersection-over-Union (mIoU) to evaluate the performances of radar point cloud semantic segmentation, as shown in Equation 10 and Equation 12.

$$\text{PA} = \frac{C}{T}, \quad (10)$$

$$\text{IoU} = \frac{I}{U}, \quad (11)$$

$$\text{mIoU} = \frac{1}{N} \sum_{i=1}^N \text{IoU}_i, \quad (12)$$

where C represents the number of correctly classified point clouds and T represents the total number of point clouds. In the equations for IoU and mIoU, I represents the number of intersection points, U represents the number of union points, and N represents the number of categories.

Camera Image Segmentation. For image semantic segmentation, Overall Accuracy (OA), Mean Pixel Accuracy

TABLE V

BENCHMARK RESULTS OF OBJECT DETECTION ON WATER SCENES. IN THE MODALITIES COLUMN, C DENOTES THE MODALITY FROM THE CAMERA SENSOR, R DENOTES THE MODALITY FROM THE 4D RADAR SENSOR, N-FRAME(S) DENOTES THE ACCUMULATION OF N-FRAME RADAR POINT CLOUDS. ADVERSE LIGHTING AND WEATHER CONDITIONS ARE EVALUATED USING MAP₅₀ METRIC.

Model	Modalities	mAP ₅₀₋₉₅	mAP ₅₀	FPS	Pier	Buoy	Sailor	Ship	Boat	Vessel	Kayak	Adverse lighting	Adverse weather
Faster R-CNN [53]	C	47.8	81.1	31.5	81.3	78.4	75.6	93.0	88.9	92.2	58.4	69.4	71.1
CenterNet [51]	C	54.7	82.9	117.4	83.0	80.1	79.3	92.7	89.5	93.1	62.9	72.2	73.7
Deformable DETR [52]	C	56.5	84.0	18.2	83.9	82.2	80.2	92.9	89.4	92.7	66.8	74.5	76.2
YOLOX-M [54]	C	57.8	85.1	54.7	85.1	81.1	80.5	91.4	89.5	92.1	76.1	77.4	78.9
YOLOv8-M [55]	C	59.2	84.4	58.8	80.6	84.3	82.1	93.7	90.8	95.8	62.5	74.8	79.5
YOLOX-M [54]	C + R ₁ -frame	59.5	86.1	51.2	85.5	82.2	81.3	92.9	91.3	92.5	77.1	79.8	82.5
YOLOX-M [54]	C + R ₃ -frames	60.3	87.4	51.2	87.1	84.1	86.5	93.7	91.8	91.2	77.7	81.5	83.5
YOLOv8-M [55]	C + R ₁ -frame	61.2	88.0	54.2	86.2	85.9	85.1	94.6	91.2	95.0	77.9	80.1	82.4
YOLOv8-M [55]	C + R ₃ -frames	62.5	88.8	54.2	84.5	87.2	87.1	94.1	93.2	96.3	79.5	82.1	84.2

(MPA) and mIoU are introduced as illustrated in Equation 13, Equation 14 and Equation 16, respectively. For image instance segmentation, mAP₅₀ and mAP₅₀₋₉₅ in box and mask are employed similarly to the object detection metrics as described in Equation 8 and Equation 9.

$$OA = \frac{TP + TN}{TP + TN + FP + FN}, \quad (13)$$

$$MPA = \frac{1}{C} \sum_{i=1}^C \frac{TP_i}{TP_i + FP_i}, \quad (14)$$

$$IoU_i = \frac{TP_i}{TP_i + FP_i + FN_i}, \quad (15)$$

$$mIoU = \frac{1}{C} \sum_{i=1}^C IoU_i, \quad (16)$$

where TP , TN , FP and FN denote predicted samples of true positive, true negative, false positive and false negative, respectively. OA represents the classification accuracy for the whole image and is the proportion of correctly classified pixels among all pixels. MPA is the average pixel accuracy across all classes, which is the proportion of correctly classified pixels to the total number of pixels in that class. Specifically, C represents the number of categories. TP_i , FP_i and FN_i are the true positive, false positive and false negative numbers for the i -th category, respectively. IoU_i denotes the IoU value for the i -th category.

Panoptic Perception. Panoptic perception includes three tasks: object detection, free-space segmentation and waterline segmentation. We adopt mAP₅₀ (Equation 8) and mAP₅₀₋₉₅ (Equation 9) to evaluate object detection performance. Additionally, we utilize OA (Equation 13) and $mIoU$ (Equation 16) to evaluate the free-space segmentation and waterline segmentation tasks.

C. Object Detection

Baseline. Table V categorizes the object detection baselines into two sections: camera-based detection and fusion-based detection. For camera-based detection, YOLOv8-M achieves the highest mAP₅₀₋₉₅ of 59.2%, 1.4% higher than YOLOX-M and 2.7% higher than Deformable DETR. Besides, it is worth

noting that YOLOX-M gets 85.1% mAP₅₀, the highest among all detectors. CenterNet gets the fastest inference speed among all detectors, reaching an impressive 117.4 FPS. Furthermore, we evaluate the performance of the models on images captured in challenging lighting and weather conditions. Notably, the accuracy of all models decreases in this case, while YOLOX-M and YOLOv8-M still maintain the highest mAP₅₀.

Fusion-based YOLOX-M and YOLOv8-M both get higher mAP₅₀₋₉₅ and mAP₅₀ than camera-based YOLOX-M and YOLOv8-M. Specifically, the fusion-based YOLOv8-M achieves an increase in mAP₅₀ from 84.4% to 88.0% compared to the camera-based YOLOv8-M. In adverse lighting and weather conditions, fusion-based models also achieve accuracy improvements. For example, in challenging lighting conditions, the fusion-based YOLOv8-M exhibits remarkable improvement, with the mAP₅₀ increasing from 74.8% to 80.1%, resulting in a noteworthy improvement of 5.3% mAP₅₀. Moreover, to enhance the density of radar point clouds, we perform experiments on accumulated 3-frame radar point clouds. It is explicit that denser radar point clouds are conducive to improving the mAP of object detection, both under normal conditions and adverse lighting and weather conditions. Despite our fusion network relying on basic operations derived from the camera model, the radar-camera fusion approaches still exhibit notable performance improvements. The highest observed improvement in performance amounts to 7.3% mAP₅₀ for challenging lighting conditions.

Discussion. Fig. 6 shows the representative outcomes obtained from both camera-based and fusion-based detection models. Obviously, 4D radar enriches features to improve the recall of distant small objects (Fig. 6(a) and Fig. 6(d)), as well as objects located in dark environments (Fig. 6(b) and Fig. 6(e)). Additionally, due to the inherent unreliability of cameras, particularly with lens failure, as presented in Fig. 6(c), camera-based YOLOX-M fails to detect sailors on the boat. Fusion-based YOLOX-M successfully identifies the sailors, as shown in Fig. 6(f), thus improving the robustness of water surface perception. Although fusion-based models perform better than camera-based models, the confidence score is relatively low, and one sailor remains undetected.

Designing efficient fusion methods based on the characteristics of different modalities is still a considerable challenge

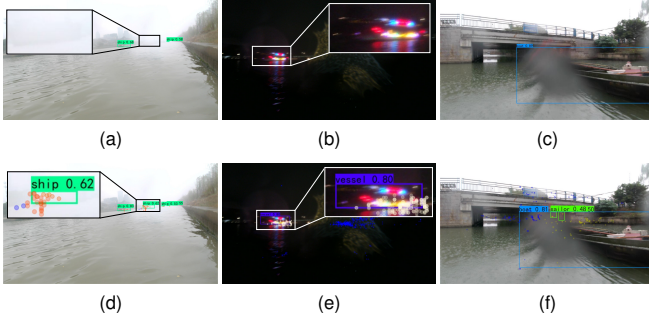


Fig. 6. Visualization of object detection on WaterScenes under foggy weather (a, d), nighttime lighting (b, e) and partial sensor failure (c, f) conditions. The first row presents the detection results by the camera-based YOLOX-M. The second row presents the detection results by fusion-based YOLOX-M with input from camera and radar modalities.

for water surfaces. On the one hand, attention mechanisms for multi-modal fusion can be applied to the water surface domain. For example, the cross-attention modules in TransFusion [75] enable adaptive determination of what and where information should be taken from the camera and LiDAR data, leading to a robust and effective fusion strategy. On the other hand, it is essential to address challenges specific to water surfaces. By leveraging techniques such as low-light enhancement [76], waterdrop removal [77], rain and fog removal [78], data quality from different modalities can be enhanced and contribute to more accurate fusion results.

D. Radar Point Cloud Segmentation

Baseline. We implement the semantic segmentation of radar point clouds based on different radar features. Table VI indicates that PointMLP achieves the lowest PA and mIoU with only location features x , y and z . By incorporating the physical features of the target, the combination of reflected power (p), compensated Doppler velocity (v), and elevation angle (e) achieves the highest accuracy, with 89.7% PA and 55.7% mIoU. Through ablation experiments, we discover that p , v , and e all exhibit the potential to improve the semantic segmentation of radar point clouds. Specifically, p proves to be more effective in semantic segmentation than v and e , as it serves as the reflected power indicating the materials of the target.

Discussion. Fig. 7 presents the visualization of 4D radar point cloud semantic segmentation in diverse environments, including normal weather, foggy weather and dark night. Radar point clouds demonstrate the capability to distinguish between targets and exhibit excellent robustness. However, it is essential to note that unlike dense point clouds produced by LiDARs, radar point clouds are sparse and lack inherent semantic characteristics. Therefore, semantic segmentation of radar point clouds relies heavily on the physical attributes of the detected targets. Moreover, applying radar sensors on water surfaces may result in water clutter, thereby reducing the accuracy of point cloud segmentation. Thus, it is necessary to consider clutter removal methods such as those proposed in [79], [80] to enhance the segmentation accuracy of 4D radar point clouds on water surfaces.

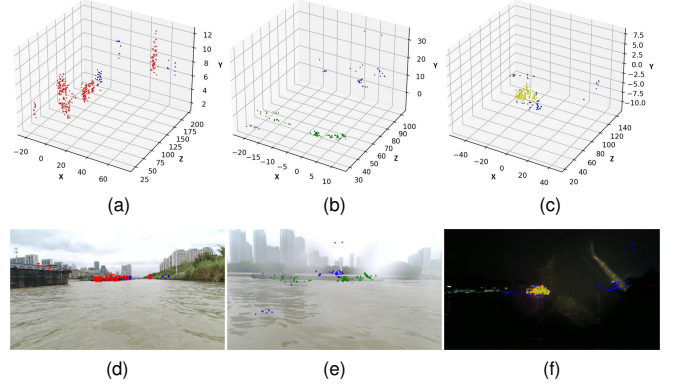


Fig. 7. Visualization of radar point cloud semantic segmentation on WaterScenes. The first row is the semantic segmentation results of 3D radar point clouds in the world coordinates. The second row shows the radar point clouds projected onto the image plane. Blue point clouds indicate clutter while point clouds of other colors represent different kinds of objects.

E. Camera Image Segmentation

Baseline of Semantic Segmentation. Table VII presents that DeepLabv3+ obtains the highest FPS among the four models. Meanwhile, HRNet, using HRNetV1-W32 as its backbone, gets 83.1% mIoU, 91.7% MPA and 95.3% OA. The above two models are based on pure-convolution networks. SegNeXt integrates the convolutional attention and employs MSCAN-B as the backbone, resulting in 95.4% OA. SegFormer adopts multi-head self-attention at the last stage of its backbone and uses a naive MLP decoder, achieving the second highest 85.7% mIoU among all models. As a transformer-based all-in-one segmentation model, Mask2Former achieves SOTA performance, exceeding SegFormer by 0.9% mIoU.

Baseline of Instance Segmentation. Experiments show that YOLOv8-M outperforms all other box-based models with 85.9% mAP₅₀, 58.2% mAP₅₀₋₉₅, and 54.6 FPS in Table VIII. Transformer-based Mask2Former achieves the highest mask mAP₅₀ of 80.7% and mAP₅₀₋₉₅ of 45.8%. For CNN-based networks, SOLO obtains the highest mask mAP₅₀ of 79.5%. Overall, YOLOv8-M offers an excellent trade-off between accuracy and inference speed.

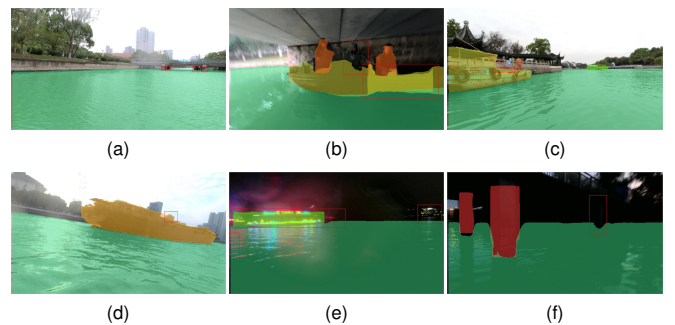


Fig. 8. Visualization of semantic segmentation on WaterScenes. (a) Blurred segmentation of distant piers. (b) Fuzzy sailor-boat boundaries. (c) Ambiguous complex boat segmentation. (d) Buildings misclassified as ship parts. (e) Indistinct ship edges in low light. (f) Piers excluded from segmentation.

Discussion. As illustrated in Fig. 8, our WaterScenes presents considerable challenges. First of all, Fig. 8(a) and

TABLE VI

BENCHMARK RESULTS OF SEMANTIC SEGMENTATION ON RADAR POINT CLOUDS, INCLUDING THE POINT ACCURACY (PA) AND mIoU OF ALL CLASSES. IN THE FEATURES COLUMN, x, y, z DENOTE THE COORDINATES IN THE CARTESIAN SYSTEM. p, v, e DENOTE REFLECTED POWER, COMPENSATED DOPPLER VELOCITY AND ELEVATION ANGLE OF THE TARGET, RESPECTIVELY.

Model	Features	PA	mIoU	Pier	Buoy	Sailor	Ship	Boat	Vessel	Kayak	Clutter
PointMLP [57]	x, y, z	81.1	38.7	36.5	11.5	2.7	85.5	26.4	53.7	9.7	83.2
PointMLP [57]	x, y, z, p	86.3	51.7	61.2	18.3	2.9	87.9	50.7	55.2	6.8	86.8
PointMLP [57]	x, y, z, v	83.0	45.3	45.4	37.6	3.2	90.6	41.2	51.7	4.8	87.5
PointMLP [57]	x, y, z, e	84.1	46.9	44.8	33.5	0.7	86.9	41.4	59.0	24.2	84.5
PointMLP [57]	x, y, z, p, v	86.9	53.0	50.4	48.3	1.1	92.1	54.3	81.8	7.8	88.0
PointMLP [57]	x, y, z, p, e	87.1	53.5	56.8	51.9	1.1	90.5	59.5	80.1	0.6	87.3
PointMLP [57]	x, y, z, v, e	84.7	49.7	48.7	32.3	1.3	87.0	41.1	60.1	43.2	84.2
PointMLP [57]	x, y, z, p, v, e	89.7	55.7	45.7	39.8	8.3	93.2	57.8	88.6	21.1	90.7
Point-NN [58]	x, y, z, p, v, e	82.1	47.9	41.6	33.4	2.1	85.6	43.8	78.7	15.6	82.4
PointNet++ [59]	x, y, z, p, v, e	86.6	53.2	45.3	40.1	5.2	90.1	53.6	82.9	22.6	85.7
Point Transformer [60]	x, y, z, p, v, e	87.9	54.4	42.0	37.8	8.0	92.1	58.1	87.6	20.7	88.9

TABLE VII

BENCHMARK RESULTS OF SEMANTIC SEGMENTATION ON WATERSCENES.

Model	mIoU	MPA	OA	FPS
DeepLabv3+ [62]	82.6	89.9	95.2	63.7
HRNet [63]	83.1	91.7	95.3	21.5
SegNeXt [64]	85.3	92.8	95.4	24.2
SegFormer [65]	85.7	93.1	95.4	59.5
Mask2Former [66]	86.6	93.9	96.2	6.8

TABLE VIII

BENCHMARK RESULTS OF INSTANCE SEGMENTATION ON WATERSCENES.

Model	Box		Mask		FPS
	mAP ₅₀	mAP ₅₀₋₉₅	mAP ₅₀	mAP ₅₀₋₉₅	
YOACT [67]	75.7	51.2	74.9	37.3	46.3
SOLO [68]	-	-	79.5	41.3	16.2
YOLOv5-M [55]	80.2	55.1	79.3	40.1	48.1
YOLOv8-M [55]	85.9	58.2	79.2	44.8	54.6
Mask2Former [66]	-	-	80.7	45.8	5.9

Fig. 8(b) demonstrate that models are not good at segmenting small objects (e.g., piers) and objects that are in close contact, such as sailors and boats. Additionally, Fig. 8(c) suggests that models struggle with boats that have complex structures (e.g., a boat with a roof supported by poles), especially when sailors are present on the boat. Furthermore, background buildings are sometimes misidentified as part of the same object as the ship with the steel structure, as shown in Fig. 8(d). In the case of dim lighting conditions, the segmentation results become quite rough or even completely missing, as illustrated in Fig. 8(e) and Fig. 8(f). Inaccurate segmentation of objects poses a significant challenge to the autonomous driving of USVs. Consequently, in addition to specific network design for camera modality on water surfaces, leveraging radar to assist camera image segmentation is a valuable research direction.

F. Panoptic Perception

Baseline. As can be seen from Table IX, benchmark results indicate both the feasibility of our dataset for panoptic perception and the challenges associated with multi-task perception on water surfaces. In general, fusion-based Achelous exhibits superior performance compared to camera-based YOLOP and HybridNets among object detection, free-space segmentation and waterline segmentation tasks. In terms of object detection, Achelous outperforms HybridNets by 15.7% mAP₅₀, demonstrating the effectiveness of radar-camera fusion on water surfaces. However, it still has a lower detection mAP than the YOLOv8-M model with radar-camera fusion in Table V, which is specifically designed for the object detection task.

Discussion. Unlike panoptic perception of object detection, drivable area and lane line segmentation for autonomous vehicles on roads, reflections on water surfaces and the unclear boundary line between water and shore make it challenging to segment free-space and waterlines. For example, as shown in Fig. 9(a) and 9(b), areas of bright spots caused by light and waves are incorrectly identified as free-space. The water surface mirrors buildings on the shore at night, further complicating the segmentation of the free-space area, as demonstrated in Fig. 9(c). Small objects tend to have less contact area with the water surface, making them easier to be missed, as is indicated in Fig. 9(d). Furthermore, as illustrated in Fig. 9(e), the boundary between the water surface and the shore is unclear, especially in low-light environments. Consequently, the model misidentifies the waterline as part of the water surface, presenting a potential risk of collision in real-world scenarios.

Multi-modal panoptic perception on water surfaces remains an unexplored and valuable research direction. In Multi-Task Learning (MTL) paradigm, multiple task-specific heads share the feature extraction process. The co-training strategy across tasks could leverage feature abstraction to save computation cost for onboard chips. Panoptic perception also serves for downstream tasks on water surfaces, such as path planning, obstacle avoidance and navigation control for USVs. Therefore, lightweight architectures that can handle multiple modalities and multiple tasks in real-time are highly desirable for edge devices on USVs.

TABLE IX

BENCHMARK RESULTS OF PANOPTIC PERCEPTION ON WATERSCENES. IN THE MODALITIES COLUMN, C DENOTES THE IMAGE MODALITY FROM THE CAMERA SENSOR, AND R DENOTES A SINGLE FRAME POINT CLOUD MODALITY FROM THE 4D RADAR SENSOR.

Model	Modalities	Params (M)	Object Detection		Free-Space Segmentation		Waterline Segmentation		FPS
			mAP ₅₀	mAP ₅₀₋₉₅	OA	mIoU	OA	mIoU	
YOLOP [70]	C	7.9	68.0	42.6	99.5	99.0	67.6	72.1	50.5
HybridNets [71]	C	12.8	69.8	49.5	97.2	98.0	65.3	69.8	45.8
Achelous-MV-GDF-S0 [72]	C + R	1.6	81.1	51.0	99.6	99.3	68.3	65.0	70.3
Achelous-MV-GDF-S1 [72]	C + R	2.8	83.5	54.1	99.6	99.4	69.5	68.7	69.6
Achelous-MV-GDF-S2 [72]	C + R	5.3	85.5	56.0	99.7	99.6	70.3	72.2	68.5

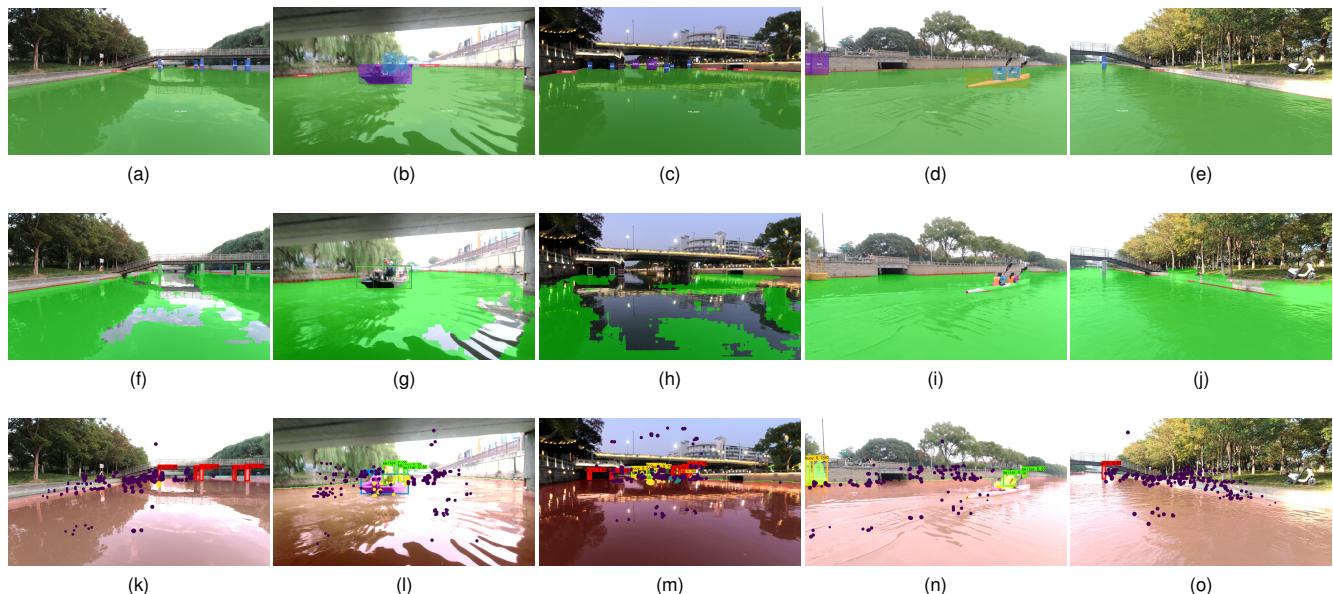


Fig. 9. Visualization of panoptic perception on WaterScenes. Images in the first row are the ground truth, in the second row are results of camera-based YOLOP, and in the third row are results of fusion-based Achelous. Panoptic perception includes object detection (boxes), free-space segmentation (masks) and waterline segmentation (lines).

V. DISCUSSIONS

A. Dataset Diversity

TABLE X

EXPERIMENTS OF DATASET DIVERSITY ON OBJECT DETECTION AND SEMANTIC SEGMENTATION TASKS.

Task	Pre-training Dataset	Evaluation Dataset	Result
Detection	-	WaterScenes	59.8
		WaterScenes	59.9 (0.1 \uparrow)
	WaterScenes	SMD	55.8
		SMD	61.5 (5.7 \uparrow)
Segmentation	-	WaterScenes	85.7
		WaterScenes	86.1 (0.4 \uparrow)
	WaterScenes	USVInland	92.5
		USVInland	98.3 (5.8 \uparrow)

To understand the superiority of our new datasets over existing datasets focused on water surfaces, we conduct experiments in two tasks: comparing WaterScenes and SMD [43] in object detection task and comparing WaterScenes

and USVInland [19] in semantic segmentation task. Specifically, in object detection experiments, we pre-train YOLOv8-N [55] on WaterScenes, followed by training and testing on SMD dataset. Table X shows a remarkable performance improvement of 5.7% mAP₅₀ using our WaterScenes as the pre-training dataset. In contrast, there is only 0.1% increase of mAP₅₀ when we use SMD as the pre-training dataset while training and testing on WaterScenes. This stark contrast highlights the superior generalization capabilities of a model trained on WaterScenes compared to scenarios in SMD dataset. Similarly to object detection experiments, we perform SegFormer-B0 [65] on WaterScenes and USVInland, achieving 5.8% mIoU improvement leveraging WaterScenes as the pre-training dataset. Experimental results from both object detection and semantic segmentation indicate the diversity compared to existing datasets as well as the inherent value derived from using WaterScenes as a pre-training resource.

B. Limitations

Although WaterScenes represents the first multi-task 4D radar-camera fusion dataset on water surfaces, offering valu-

able resources to this field, some limitations still exist in our work. Given that we aim to explore a low-cost and robust perception approach using radar and camera modalities, we excluded high-definition LiDAR. Thus, object detection is limited to 2D annotations, as sparse radar point clouds cannot replace LiDAR for 3D bounding box annotation. Instead, radar data serves as a feature pattern to assist the camera in fusion-based 2D object detection rather than independently completing reliable detection tasks. In addition, we mainly focused on providing a foundational baseline for radar-camera fusion on water surfaces using our newly introduced dataset. The accuracy improvement might seem insignificant due to the absence of advanced fusion techniques. Nevertheless, our baseline serves as an essential starting point, and more advanced fusion algorithms could yield significantly higher accuracy levels.

C. Future Works

As a relatively unexplored field, autonomous driving on water surfaces presents several potential research directions. Compared to autonomous driving on road surfaces, perception challenges encountered on water surfaces are more daunting and unpredictable, including water splashes, mirror-like reflections, adverse lighting and weather conditions. With our WaterScenes dataset containing diverse scenarios and environmental conditions, researchers can customize algorithms to address the challenges of camera-based perception on water surfaces. Additionally, current perception models for autonomous driving emphasize multi-modal fusion and multi-task learning trends [81], [82]. A high-generalization, reusable fusion approach can reduce operational costs and power consumption, thus improving the inference speed [83]. With our diverse collection of radar and camera data captured from real-world water environments, constructing a multi-task and multi-sensor robust perception model suitable for water surfaces is an interesting and potential research direction.

VI. CONCLUSION

This work presents a pioneering multi-modal and multi-task dataset that sheds light on previously unexplored 4D radar-camera fusion on water surfaces. Leveraging the complementary advantages of radar and camera sensors, our WaterScenes dataset enables multi-attribute and all-weather perception of the water environment. We evaluate SOTA algorithms on camera image modality, radar point cloud modality and radar-camera fusion modality on WaterScenes, generating insights into water surface perception that were previously unknown. Experimental results demonstrate the value of the dataset for further investigation and also indicate that the 4D radar-camera combination is a robust solution for USVs on water surfaces. Without optimization on popular models, radar-camera fusion can actually improve detection performance, especially in adverse lighting and weather conditions. Overall, the presented WaterScenes offers a valuable resource for researchers interested in autonomous driving on water surfaces and aims to motivate novel ideas and directions for the development of water surface perception algorithms.

ACKNOWLEDGMENTS

This research was funded by the Suzhou Municipal Key Laboratory for Intelligent Virtual Engineering (SZS2022004), the Suzhou Science and Technology Project (SYG202122), the Research Development Fund of XJTU (RDF-19-02-23), XJTU AI University Research Centre, Jiangsu Province Engineering Research Centre of Data Science and Cognitive Computation at XJTU and SIP AI innovation platform (YZCXPT2022103). This work received financial support from Jiangsu Industrial Technology Research Institute (JITRI) and Wuxi National Hi-Tech District (WND).

REFERENCES

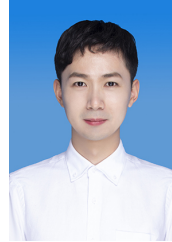
- [1] D. Feng, C. Haase-Schütz, L. Rosenbaum, H. Hertlein, C. Glaeser, F. Timm, W. Wiesbeck, and K. Dietmayer, "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 3, pp. 1341–1360, 2020.
- [2] X. Wu, W. Li, D. Hong, R. Tao, and Q. Du, "Deep learning for unmanned aerial vehicle-based object detection and tracking: A survey," *IEEE Geoscience and Remote Sensing Magazine*, vol. 10, no. 1, pp. 91–124, 2021.
- [3] B. Bovcon, J. Muhovič, D. Vranac, D. Mozetič, J. Perš, and M. Kristan, "Mods—a usv-oriented object detection and obstacle segmentation benchmark," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 8, pp. 13403–13418, 2021.
- [4] B. Bovcon, J. Muhovič, J. Perš, and M. Kristan, "The mastr1325 dataset for training deep usv obstacle detection models," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 3431–3438.
- [5] Y. Cheng, H. Xu, and Y. Liu, "Robust small object detection on the water surface through fusion of camera and millimeter wave radar," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 15263–15272.
- [6] J. Lin, P. Diekmann, C.-E. Framing, R. Zweigel, and D. Abel, "Maritime environment perception based on deep learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 9, pp. 15487–15497, 2022.
- [7] N. Wang, T. Chen, S. Liu, R. Wang, H. R. Karimi, and Y. Lin, "Deep learning-based visual detection of marine organisms: A survey," *Neurocomputing*, vol. 532, pp. 1–32, 2023.
- [8] N. Wang, H. He, Y. Hou, and B. Han, "Model-free visual servo swarming of manned-unmanned surface vehicles with visibility maintenance and collision avoidance," *IEEE Transactions on Intelligent Transportation Systems*, 2023.
- [9] H. He, N. Wang, D. Huang, and B. Han, "Active vision-based finite-time trajectory-tracking control of an unmanned surface vehicle without direct position measurements," *IEEE Transactions on Intelligent Transportation Systems*, 2024.
- [10] S. Fang, Z. Liu, X. Wang, Y. Cao, and Z. Yang, "Dynamic analysis of emergency evacuation in a rolling passenger ship using a two-layer social force model," *Expert Systems with Applications*, vol. 247, p. 123310, 2024.
- [11] N. Wang, Y. Wang, and M. J. Er, "Review on deep learning techniques for marine object recognition: Architectures and algorithms," *Control Engineering Practice*, vol. 118, p. 104458, 2022.
- [12] N. Wang, Y. Wang, Y. Feng, and Y. Wei, "Aodemar: Attention-aware occlusion detection of vessels for maritime autonomous surface ships," *IEEE Transactions on Intelligent Transportation Systems*, 2024.
- [13] R. W. Liu, Y. Lu, Y. Guo, W. Ren, F. Zhu, and Y. Lv, "Aioenet: All-in-one low-visibility enhancement to improve visual perception for intelligent marine vehicles under severe weather conditions," *IEEE Transactions on Intelligent Vehicles*, 2023.
- [14] J. Qu, R. W. Liu, Y. Gao, Y. Guo, F. Zhu, and F.-Y. Wang, "Double domain guided real-time low-light image enhancement for ultra-high-definition transportation surveillance," *IEEE Transactions on Intelligent Transportation Systems*, 2024.
- [15] N. Wang, Y. Wang, Y. Feng, and Y. Wei, "Mdd-shipnet: Math-data integrated defogging for fog-occlusion ship detection," *IEEE Transactions on Intelligent Transportation Systems*, 2024.

- [16] B. Xie, Z. Yang, L. Yang, A. Wei, X. Weng, and B. Li, "Ammf: Attention-based multi-phase multi-task fusion for small contour object 3d detection," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 2, pp. 1692–1701, 2022.
- [17] Y.-J. Li, J. Park, M. O'Toole, and K. Kitani, "Modality-agnostic learning for radar-lidar fusion in vehicle detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 918–927.
- [18] Z. Liu, Y. Zhang, X. Yu, and C. Yuan, "Unmanned surface vehicles: An overview of developments and challenges," *Annual Reviews in Control*, vol. 41, pp. 71–93, 2016.
- [19] Y. Cheng, M. Jiang, J. Zhu, and Y. Liu, "Are we ready for unmanned surface vehicles in inland waterways? the usinland multisensor dataset and benchmark," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 3964–3970, 2021.
- [20] A. Ouaknine, A. Newson, J. Rebut, F. Tupin, and P. Pérez, "Carrada dataset: Camera and automotive radar with range-angle-doppler annotations," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 5068–5075.
- [21] M. Bijelic, T. Gruber, F. Mannan, F. Kraus, W. Ritter, K. Dietmayer, and F. Heide, "Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 682–11 692.
- [22] S. Yao, R. Guan, Z. Peng, C. Xu, Y. Shi, Y. Yue, E. G. Lim, H. Seo, K. L. Man, X. Zhu *et al.*, "Radar perception in autonomous driving: Exploring different data representations," *arXiv preprint arXiv:2312.04861*, 2023.
- [23] I. Bilik, "Comparative analysis of radar and lidar technologies for automotive applications," *IEEE Intelligent Transportation Systems Magazine*, vol. 15, no. 1, pp. 244–269, 2022.
- [24] N. Scheiner, F. Kraus, F. Wei, B. Phan, F. Mannan, N. Appenrodt, W. Ritter, J. Dickmann, K. Dietmayer, B. Sick *et al.*, "Seeing around street corners: Non-line-of-sight detection and tracking in-the-wild using doppler radar," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2068–2077.
- [25] Y. Wang, Z. Jiang, Y. Li, J.-N. Hwang, G. Xing, and H. Liu, "Rodnet: A real-time radar object detection network cross-supervised by camera-radar fused object 3d localization," *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 4, pp. 954–967, 2021.
- [26] J. Liu, W. Xiong, L. Bai, Y. Xia, T. Huang, W. Ouyang, and B. Zhu, "Deep instance segmentation with automotive radar detection points," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 1, pp. 84–94, 2022.
- [27] Y. Cheng, J. Zhu, M. Jiang, J. Fu, C. Pang, P. Wang, K. Sankaran, O. Onabola, Y. Liu, D. Liu *et al.*, "Flow: A dataset and benchmark for floating waste detection in inland waters," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 953–10 962.
- [28] J. Zhang, H. Liu, K. Yang, X. Hu, R. Liu, and R. Stiefelhagen, "Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers," *IEEE Transactions on Intelligent Transportation Systems*, 2023.
- [29] J. S. Berrio, M. Shan, S. Worrall, and E. Nebot, "Camera-lidar integration: Probabilistic sensor fusion for semantic mapping," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 7, pp. 7637–7652, 2021.
- [30] Y. Guo, R. W. Liu, J. Qu, Y. Lu, F. Zhu, and Y. Lv, "Asynchronous trajectory matching-based multimodal maritime data fusion for vessel traffic surveillance in inland waterways," *IEEE Transactions on Intelligent Transportation Systems*, 2023.
- [31] S. Chadwick, W. Maddern, and P. Newman, "Distant vehicle detection using radar and vision," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 8311–8317.
- [32] R. Nabati and H. Qi, "Centerfusion: Center-based radar and camera fusion for 3d object detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1527–1536.
- [33] Y. Kim, S. Kim, J. W. Choi, and D. Kum, "Craft: Camera-radar 3d object detection with spatio-contextual fusion transformer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 1, 2023, pp. 1160–1168.
- [34] S. Yao, R. Guan, X. Huang, Z. Li, X. Sha, Y. Yue, E. G. Lim, H. Seo, K. L. Man, X. Zhu *et al.*, "Radar-camera fusion for object detection and semantic segmentation in autonomous driving: A comprehensive review," *IEEE Transactions on Intelligent Vehicles*, 2023.
- [35] M. Meyer and G. Kusch, "Automotive radar dataset for deep learning based 3d object detection," in *2019 16th european radar conference (EuRAD)*. IEEE, 2019, pp. 129–132.
- [36] D.-H. Paek, S.-H. Kong, and K. T. Wijaya, "K-radar: 4d radar object detection for autonomous driving in various weather conditions," *Advances in Neural Information Processing Systems*, vol. 35, pp. 3819–3829, 2022.
- [37] A. Palffy, E. Pool, S. Baratam, J. F. Kooij, and D. M. Gavrila, "Multi-class road user detection with 3+ 1d radar in the view-of-delft dataset," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 4961–4968, 2022.
- [38] L. Zheng, Z. Ma, X. Zhu, B. Tan, S. Li, K. Long, W. Sun, S. Chen, L. Zhang, M. Wan *et al.*, "Tj4dradset: A 4d radar dataset for autonomous driving," in *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2022, pp. 493–498.
- [39] L. Zheng, S. Li, B. Tan, L. Yang, S. Chen, L. Huang, J. Bai, X. Zhu, and Z. Ma, "Rcfusion: Fusing 4d radar and camera with bird's-eye view features for 3d object detection," *IEEE Transactions on Instrumentation and Measurement*, 2023.
- [40] W. Xiong, J. Liu, T. Huang, Q.-L. Han, Y. Xia, and B. Zhu, "Lxl: Lidar excluded lean 3d object detection with 4d imaging radar and camera fusion," *IEEE Transactions on Intelligent Vehicles*, 2023.
- [41] M. Kristan, V. S. Kenk, S. Kovačič, and J. Perš, "Fast image-based obstacle detection from unmanned surface vehicles," *IEEE transactions on cybernetics*, vol. 46, no. 3, pp. 641–654, 2015.
- [42] B. Bovcon, J. Perš, M. Kristan *et al.*, "Stereo obstacle detection for unmanned surface vehicles by imu-assisted semantic segmentation," *Robotics and Autonomous Systems*, vol. 104, pp. 1–13, 2018.
- [43] S. Moosbauer, D. König, J. Jakel, and M. Teutsch, "A benchmark for deep learning based object detection in maritime environments," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2019, pp. 0–0.
- [44] J. Liu, H. Li, J. Luo, S. Xie, and Y. Sun, "Efficient obstacle detection based on prior estimation network and spatially constrained mixture model for unmanned surface vehicles," *Journal of Field Robotics*, vol. 38, no. 2, pp. 212–228, 2021.
- [45] L. Züst, J. Perš, and M. Kristan, "Lars: A diverse panoptic maritime obstacle detection dataset and benchmark," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 20 304–20 314.
- [46] N. Wang, Y. Wang, Y. Wei, B. Han, and Y. Feng, "Marine vessel detection dataset and benchmark for unmanned surface vehicles," *Applied Ocean Research*, vol. 142, p. 103835, 2024.
- [47] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nusenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.
- [48] J. Domhof, J. F. Kooij, and D. M. Gavrila, "A joint extrinsic calibration tool for radar, camera and lidar," *IEEE Transactions on Intelligent Vehicles*, vol. 6, no. 3, pp. 571–582, 2021.
- [49] M. Meyer and G. Kusch, "Deep learning based 3d object detection for automotive radar and camera," in *2019 16th European Radar Conference (EuRAD)*. IEEE, 2019, pp. 133–136.
- [50] L. Stäcker, P. Heidenreich, J. Rambach, and D. Stricker, "Fusion point pruning for optimized 2d object detection with radar-camera fusion," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 3087–3094.
- [51] X. Zhou, D. Wang, and P. Krähenhöh, "Objects as points," *arXiv preprint arXiv:1904.07850*, 2019.
- [52] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *arXiv preprint arXiv:2010.04159*, 2020.
- [53] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.
- [54] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "Yolox: Exceeding yolo series in 2021," *arXiv preprint arXiv:2107.08430*, 2021.
- [55] G. Jocher, A. Chaurasia, and J. Qiu, "YOLO by Ultralytics," Jan. 2023. [Online]. Available: <https://github.com/ultralytics/ultralytics>
- [56] Q.-L. Zhang and Y.-B. Yang, "Sa-net: Shuffle attention for deep convolutional neural networks," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 2235–2239.
- [57] X. Ma, C. Qin, H. You, H. Ran, and Y. Fu, "Rethinking network design and local geometry in point cloud: A simple residual mlp framework," *arXiv preprint arXiv:2202.07123*, 2022.
- [58] R. Zhang, L. Wang, Z. Guo, Y. Wang, P. Gao, H. Li, and J. Shi, "Parameter is not all you need: Starting from non-parametric networks for 3d point cloud analysis," *arXiv preprint arXiv:2303.08134*, 2023.
- [59] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *Advances in neural information processing systems*, vol. 30, 2017.

- [60] H. Zhao, L. Jiang, J. Jia, P. H. Torr, and V. Koltun, "Point transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 16 259–16 268.
- [61] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.
- [62] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.
- [63] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3349–3364, 2020.
- [64] M.-H. Guo, C.-Z. Lu, Q. Hou, Z. Liu, M.-M. Cheng, and S.-M. Hu, "Segnext: Rethinking convolutional attention design for semantic segmentation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 1140–1156, 2022.
- [65] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in neural information processing systems*, vol. 34, pp. 12 077–12 090, 2021.
- [66] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 1290–1299.
- [67] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "Yolact: Real-time instance segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9157–9166.
- [68] X. Wang, T. Kong, C. Shen, Y. Jiang, and L. Li, "Solo: Segmenting objects by locations," in *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*. Springer, 2020, pp. 649–665.
- [69] G. Jocher, "YOLOv5 by Ultralytics," May 2020. [Online]. Available: <https://github.com/ultralytics/yolov5>
- [70] D. Wu, M.-W. Liao, W.-T. Zhang, X.-G. Wang, X. Bai, W.-Q. Cheng, and W.-Y. Liu, "Yolop: You only look once for panoptic driving perception," *Machine Intelligence Research*, vol. 19, no. 6, pp. 550–562, 2022.
- [71] D. Vu, B. Ngo, and H. Phan, "Hybridnets: End-to-end perception network," *arXiv preprint arXiv:2203.09035*, 2022.
- [72] R. Guan, S. Yao, X. Zhu, K. L. Man, E. G. Lim, J. Smith, Y. Yue, and Y. Yue, "Achelus: A fast unified water-surface panoptic perception framework based on fusion of monocular camera and 4d mmwave radar," in *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2023, pp. 182–188.
- [73] S. Mehta and M. Rastegari, "Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer," *arXiv preprint arXiv:2110.02178*, 2021.
- [74] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7482–7491.
- [75] X. Bai, Z. Hu, X. Zhu, Q. Huang, Y. Chen, H. Fu, and C.-L. Tai, "Transfusion: Robust lidar-camera fusion for 3d object detection with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 1090–1099.
- [76] S. Zhou, C. Li, and C. Change Loy, "Lednet: Joint low-light enhancement and deblurring in the dark," in *European conference on computer vision*. Springer Nature Switzerland Cham, 2022, pp. 573–589.
- [77] Q. Wen, Y. Wu, and Q. Chen, "Video waterdrop removal via spatio-temporal fusion in driving scenes," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 10 003–10 009.
- [78] R. Li, R. T. Tan, and L.-F. Cheong, "All in one bad weather removal using architectural search," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 3175–3185.
- [79] Y. Cheng, J. Su, H. Chen, and Y. Liu, "A new automotive radar 4d point clouds detector by using deep learning," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 8398–8402.
- [80] Y. Cheng and Y. Liu, "Person reidentification based on automotive radar point clouds," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2021.
- [81] Y. Hu, J. Yang, L. Chen, K. Li, C. Sima, X. Zhu, S. Chai, S. Du, T. Lin, W. Wang *et al.*, "Planning-oriented autonomous driving," in

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 17 853–17 862.

- [82] T. Ye, W. Jing, C. Hu, S. Huang, L. Gao, F. Li, J. Wang, K. Guo, W. Xiao, W. Mao *et al.*, "Fusionad: Multi-modality fusion for prediction and planning tasks of autonomous driving," *arXiv preprint arXiv:2308.01006*, 2023.
- [83] X. Liang, Y. Wu, J. Han, H. Xu, C. Xu, and X. Liang, "Effective adaptation in multi-task co-training for unified autonomous driving," *Advances in Neural Information Processing Systems*, vol. 35, pp. 19 645–19 658, 2022.



Shanliang Yao (Student Member, IEEE) received the B.E. degree in 2016 from the School of Computer Science and Technology, Soochow University, Suzhou, China, and the M.S. degree in 2021 from the Faculty of Science and Engineering, University of Liverpool, Liverpool, U.K. He is currently a joint Ph.D. student of University of Liverpool, Xi'an Jiaotong-Liverpool University and Institute of Deep Perception Technology, Jiangsu Industrial Technology Research Institute. His current research is centered on multi-modal perception using deep learning approach for autonomous driving. He is also interested in robotics, intelligent vehicles and intelligent transportation systems.



Runwei Guan (Student Member, IEEE) received his M.S. degree in Data Science from University of Southampton, Southampton, United Kingdom, in 2021. He is currently a joint Ph.D. student of University of Liverpool, Xi'an Jiaotong-Liverpool University and Institute of Deep Perception Technology, Jiangsu Industrial Technology Research Institute. His research interests include visual grounding, panoptic perception based on the fusion of radar and camera, lightweight neural network, multi-task learning and statistical machine learning. He serves as the peer reviewer of IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, Engineering Applications of Artificial Intelligence, Journal of Supercomputing, IJCNN, etc.



Zhaodong Wu is an undergraduate student from Xi'an Jiaotong-Liverpool University and will receive his bachelor's degree in 2024. His primary research interests are AI-supported clinical decision support, especially medical image processing, and computer vision for autonomous driving.



Yi Ni is an undergraduate student at Xi'an Jiaotong-Liverpool University (XJTLU), majoring in Information and Computing Science. His research interests include computer vision, panoptic perception based on the fusion of radar and camera, medical image segmentation and classification, and multi-task learning. He has participated in research including multi-modal perception based on the fusion of radar and camera, predicting neurological recovery from coma after cardiac arrest using EEG recordings, and cervical spine fracture detection through two-stage approach of mask segmentation and windowing.



Zile Huang is currently an undergraduate student at the University of Liverpool, Xi'an Jiaotong-Liverpool University. He is pursuing a degree in Information and Computing Science. His research interests include real-time object detection, multimodal learning, and scene understanding.



Ryan Wen Liu (Member, IEEE) received the B.Sc. degree (Hons.) in Information and Computing Science from the Department of Mathematics, Wuhan University of Technology, Wuhan, China, in 2009, and the Ph.D. degree from The Chinese University of Hong Kong, Hong Kong, in 2015. He is currently a Professor with the School of Navigation, Wuhan University of Technology. His research interests include intelligent waterborne transportation systems and intelligent marine vehicles. He is an Associate Editor of the *IET Intelligent Transport Systems*,

International Journal of Intelligent Transportation Systems Research, and IEEE OPEN JOURNAL OF VEHICULAR TECHNOLOGY.



Yong Yue Fellow of Institution of Engineering and Technology (FIET), received the B.Eng. degree in mechanical engineering from Northeastern University, Shenyang, China, in 1982, and the Ph.D. degree in computer aided design from Heriot-Watt University, Edinburgh, U.K., in 1994. He worked in the industry for eight years and followed experience in academia with the University of Nottingham, Cardiff University, and the University of Bedfordshire, U.K. He is currently a Professor and Director with the Virtual Engineering Centre, Xi'an Jiaotong-Liverpool

University, Suzhou, China. His current research interests include computer graphics, virtual reality, and robot navigation.



Weiping Ding (M'16-SM'19) received the Ph.D. degree in Computer Science, Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2013. In 2016, He was a Visiting Scholar at National University of Singapore, Singapore. From 2017 to 2018, he was a Visiting Professor at University of Technology Sydney, Australia. He is a Full Professor with the School of Information Science and Technology, Nantong University, Nantong, China, and also the supervisor of Ph.D postgraduate by the Faculty of Data Science at City University of Macau,

China. His main research directions involve deep neural networks, multimodal machine learning, and medical images analysis. He ranked within the top 2% Ranking of Scientists in the World by Stanford University (2020-2023). He has published over 250 articles, including over 100 IEEE Transactions papers. His fifteen authored/co-authored papers have been selected as ESI Highly Cited Papers. He serves as an Associate Editor/Editorial Board member of IEEE Transactions on Neural Networks and Learning Systems, IEEE Transactions on Fuzzy Systems, IEEE/CAA Journal of Automatica Sinica, IEEE Transactions on Intelligent Transportation Systems, IEEE Transactions on Intelligent Vehicles, IEEE Transactions on Emerging Topics in Computational Intelligence, IEEE Transactions on Artificial Intelligence, Information Fusion, Information Sciences, Neurocomputing, Applied Soft Computing. He is the Leading Guest Editor of Special Issues in several prestigious journals, including IEEE Transactions on Evolutionary Computation, IEEE Transactions on Fuzzy Systems, and Information Fusion.



Eng Gee Lim (Senior Member, IEEE) received the B.Eng. (Hons.) and Ph.D. degrees in Electrical and Electronic Engineering (EEE) from Northumbria University, Newcastle, U.K., in 1998 and 2002, respectively. He worked for Andrew Ltd., Coventry, U.K., a leading communications systems company from 2002 to 2007. Since 2007, he has been with Xi'an Jiaotong-Liverpool University, Suzhou, China, where he was the Head of the EEE Department, and the University Dean of research and graduate studies. He is currently the School Dean of Advanced Technology, the Director of the AI University Research Centre, and a Professor with the Department of EEE. He has authored or coauthored over 100 refereed international journals and conference papers. His research interests are artificial intelligence (AI), robotics, AI+ health care, international standard (ISO/IEC) in robotics, antennas, RF/microwave engineering, EM measurements/simulations, energy harvesting, power/energy transfer, smart-grid communication, and wireless communication networks for smart and green cities. He is a Chartered Engineer and a fellow of The Institution of Engineering and Technology (IET) and Engineers Australia. He is also a Senior Fellow of Higher Education Academy (HEA).

He is currently the School Dean of Advanced Technology, the Director of the AI University Research Centre, and a Professor with the Department of EEE. He has authored or coauthored over 100 refereed international journals and conference papers. His research interests are artificial intelligence (AI), robotics, AI+ health care, international standard (ISO/IEC) in robotics, antennas, RF/microwave engineering, EM measurements/simulations, energy harvesting, power/energy transfer, smart-grid communication, and wireless communication networks for smart and green cities. He is a Chartered Engineer and a fellow of The Institution of Engineering and Technology (IET) and Engineers Australia. He is also a Senior Fellow of Higher Education Academy (HEA).



Hyungjoon Seo (Member, IEEE) received the bachelor's degree in civil engineering from Korea University, Seoul, South Korea, in 2007, and the Ph.D. degree in geotechnical engineering from Korea University in 2013. In 2013, he worked as a research professor in Korea University. He served as a visiting scholar at University of Cambridge, Cambridge, UK, and he worked for engineering department in University of Cambridge as a research associate from 2014 to 2016. In August 2016, he got an assistant professor position in the Department of Civil Engineering

at the Xi'an Jiaotong Liverpool University (XJTLU), China. He has been an assistant professor at the University of Liverpool, UK, from 2020. His research interests are monitoring using artificial intelligence and SMART monitoring system for infrastructure, soil-structure interaction (tunneling, slope stability, pile), Antarctic survey and freezing ground. Hyungjoon is the director of the CSMI (Centre for SMART Monitoring Infrastructure), CSMI is collaborating with University of Cambridge, University of Oxford, University of Bath, UC Berkeley University, Nanjing University, and Tongji University on SMART monitoring. He presented a keynote speech at the 15th European Conference on Soil Mechanics and Geotechnical Engineering in 2015. He is currently appointed editor of the *CivilEng* journal and organized two international conferences. He has published more than 50 scientific papers including a book on Geotechnical Engineering and SMART monitoring.



Ka Lok Man (Member, IEEE) received the Dr. Eng. degree in electronic engineering from the Politecnico di Torino, Turin, Italy, in 1998, and the Ph.D. degree in computer science from Technische Universiteit Eindhoven, Eindhoven, The Netherlands, in 2006. He is currently a Professor in Computer Science and Software Engineering with Xi'an Jiaotong-Liverpool University, Suzhou, China. His research interests include formal methods and process algebras, embedded system design and testing, and photovoltaics.



Jieming Ma received the M.Sc. degree in advanced microelectronic systems engineering from the University of Bristol, UK, in 2010, and received the Ph.D. degree in computer science from the University of Liverpool, UK, in 2014. He is currently working as an Associate Professor at the Xi'an Jiaotong-Liverpool University, China. His research interests include intelligent optimization, machine learning and applications in renewable energy systems.



Xiaohui Zhu (Member, IEEE) received his Ph.D. from the University of Liverpool, UK in 2019. He is currently an associate professor, Ph.D. supervisor and Programme Director with the Department of Computing, School of Advanced Technology, Xi'an Jiaotong-Liverpool University. He focuses on advanced techniques related to autonomous driving, including sensor-fusion perception, fast path planning, autonomous navigation and multi-vehicle collaborative scheduling.



Yutao Yue (Member, IEEE) is an associate professor at the Artificial Intelligence Thrust and Intelligent Transportation Thrust of Hong Kong University of Science and Technology (Guangzhou). He received his Bachelor's degree from the University of Science and Technology of China, and Master and PhD degree from Purdue University. He has a dual background in academia and industry, as the team leader of Guangdong Province Introduced Innovation Scientific Research Team, senior scientist of Kuang-Chi Group, and the founder of the Institute of Deep Perception Technology of JITRI. His research interests include multimodal perception fusion, machine consciousness, artificial general intelligence, causal emergence, etc. He has been engaged in scientific research and technology industrialization for over 20 years. He has co-invented 354 granted Chinese patents, 18 USA patents, and 7 EU patents. He has led 6 major research projects with a total funding of nearly 130 million RMB. He has published over 60 papers, advised 13 postdoc research fellows, and received multiple awards including Wu Wenjun Artificial Intelligence Science and Technology Award.