Received 9 December 2024; revised 12 April 2025 and 21 July 2025; accepted 13 August 2025.
Date of publication 26 August 2025; date of current version 26 September 2025.
This article was recommended by Executive Editor Philippe Giguere.

Digital Object Identifier 10.1109/TFR.2025.3602937

SeePerSea: Multimodal Perception Dataset of In-Water Objects for Autonomous Surface Vehicles

MINGI JEONG[1] (Member, IEEE), ARIHANT CHADDA[2], ZIANG REN[3], LUYANG ZHAO[1] (Member, IEEE), HAOWEN LIU[4], AIWEI ZHANG[1], YITAO JIANG[1], SABRIEL ACHONG[1], SAMUEL LENSGRAF[5], MONIKA ROZNERE[6], AND ALBERTO QUATTRINI LI[1] (Member, IEEE)

[1]Department of Computer Science, Dartmouth College, Hanover, NH 03755 USA
[2]IQT Labs, Tysons, VA 94025 USA
[3]Department of Computer Science, Columbia University, New York City, NY 10027 USA
[4]Department of Computer Science, University of Maryland at College Park, College Park, MD 20742 USA
[5]The Institute for Human and Machine Cognition, The University of West Florida, Pensacola, FL 32502 USA
[6]School of Computing, Binghamton University, Binghamton, NY 13902 USA

CORRESPONDING AUTHOR: MINGI JEONG (e-mail: mingi.jeong.gr@dartmouth.edu) AND ALBERTO QUATTRINI LI (e-mail: alberto.quattrini.li@dartmouth.edu)

This work was supported in part by the Burke Research Initiation Award; in part by NSF under Grant CNS-1919647, Grant 2144624, and Grant OIA1923004; and in part by the National Oceanic and Atmospheric Administration (NOAA) NH Sea Grant.

(Regular Article)

ABSTRACT This article introduces the first publicly accessible labeled multimodal perception dataset for autonomous maritime navigation, focusing on in-water obstacles within the aquatic environment to enhance situational awareness for autonomous surface vehicles (ASVs). This dataset, collected over four years and consisting of diverse objects encountered under varying environmental conditions, aims to bridge the research gap in ASVs by providing a multimodal, annotated, and ego-centric perception dataset, for object detection and classification. We also show the applicability of the proposed dataset by training and testing current deep learning-based open-source perception algorithms that have shown success in the autonomous ground vehicle domain. With the training and testing results, we discuss open challenges for existing datasets and methods, identifying future research directions. We expect that our dataset will contribute to the development of future marine autonomy pipelines and marine (field) robotics. This dataset is open source and found at https://seepersea.github.io/

INDEX TERMS Autonomous surface vehicle (ASV), maritime perception, multimodal dataset, obstacle classification, obstacle detection, situational awareness.

I. INTRODUCTION

LEARNING-BASED, multimodal algorithms have shown terrestrial domain success for self-driving cars on the road to autonomy. The precondition(s) to this success fundamentally rest on the availability of relevant, labeled datasets [1], [2], [3]. Equivalent success in marine autonomous surface vehicles (ASVs) is, unsurprisingly, hampered by the lack of relevant multimodal perception datasets. Thus, the goal of this article is to **create the first publicly available labeled, multimodal 3-D perception dataset for autonomous maritime navigation** (see Fig. 1). This dataset, consisting of in-water obstacles, aims to enhance ASVs' situational awareness. Situational awareness is a foundational task that undergirds autonomy, which is increasing in importance given the focus on ASVs for tasks such as environmental monitoring and automated transportation. This importance will only grow as marine trade increases to 90% of the share of world trade [4] and, accordingly,

the expected size of the ASV market will grow to 2.7*B* USD by 2032 [5].

Understanding the locations of static and dynamic objects in the aquatic domain (**object detection**) and determining the types of these objects (**object classification**) are crucial tasks for *data association*—to understand the speed and heading of approaching objects. Such processes are integral for *navigational decision-making*, i.e., collision avoidance. However, aquatic domain challenges, including 1) unstructured navigational environments and 2) the limited maneuverability of marine vehicles, raise the importance of *early* and *accurate* state estimation of in-water obstacles for safe and efficient navigation that minimizes detection errors (e.g., false negatives). Among human error-driven marine accidents, over 70% are attributed to improper situational awareness [6]. Consequently, marine vehicles, even human-driven vessels, naturally rely on **multimodal** data for situational awareness, which aligns with the regulations (e.g., rule 5 *look-out*) explicitly covered by the maritime rules of the road [7].

The scarcity of multimodal labeled 3-D perception datasets for ASVs is attributed to the high operational costs and the extensive labeling effort required [8]. Among the few existing datasets in the aquatic domain, the open-source ones primarily consist of either 1) **single-modality** data that is typically image-based [9], [10], [11], [12], [13], [14], [15], [16], or 2) multiple modalities but lacking **object labels** across modalities [17], [18], which are essential for ground-truth evaluation [19]. This absence of multimodal and ground-truth annotations significantly hinders the development of crucial ASV capabilities, as noted in [16] and [19].

Accordingly, we release the first multimodal labeled maritime dataset. Our dataset includes expeditions from 2021 to 2024 using our ASV platform *Catabot* and a human-driven vessel in different locations (USA, Barbados, and South Korea) covering various environments (both sea and fresh water), conditions (e.g., dusk and daylight), and encounters (e.g., head-on and crossing) with various objects. The proposed dataset includes navigation-oriented three class (ship, buoy, and other) labeled objects for detection and classification. We selected these labels according to the international traffic rule [7] and buoyage system [20]. In summary, the dataset is composed of 11 561 frames of LiDAR point clouds and RGB images. We also demonstrate the utility of the proposed dataset using deep learning-based open-source perception algorithms – both single-modality and multimodal fusion—that have shown success in the terrestrial domain, with both quantitative and qualitative evaluations: highlighting success in some scenarios, but also current gaps.

We release our dataset publicly (https://seepersea.github.io/) for the community and expect it will have the following contributions.

1) SeePerSea, the first LiDAR-camera dataset in aquatic environments with object labels across two modalities, will foster the development of robust fusion perception pipelines for ASV autonomy.
2) SeePerSea, covering various environments and day conditions, will help ensure that developed perception pipelines are increasingly generalizable.

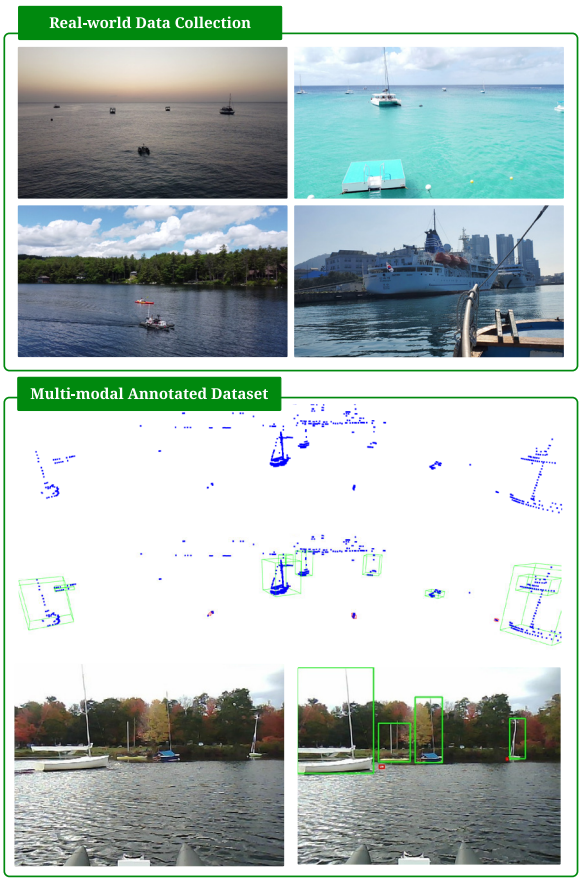


FIGURE 1. **Real-world data collection of in-water objects by ASV and a human-driven boat in operation at different geographic locations and conditions. We provide a multimodal annotated dataset (LiDAR and RGB camera) for marine autonomy.**

Overall, the SeePerSea dataset will contribute to the development of state-of-the-art marine autonomy pipelines and accelerate the future of marine (field) robotics.

The structure of this article is as follows. Section II discusses datasets both in the ground and maritime domains. Section III describes how the data was collected, annotated, and structured. Section IV provides an analysis of the dataset characteristics. Section V presents the results from current deep learning pipelines trained on the provided dataset, and Section VI discusses lessons learned and current gaps. Finally, Section VII summarizes the article and highlights future work.

II. RELATED WORK

Self-driving car datasets focused on 3-D perception, including [1], [2], and [3], have been crucial for progress in terrestrial robotic perception, especially for tasks like object detection, classification, segmentation, and tracking. These collections frequently feature a range of sensors, employing

TABLE 1. Comparison of the state-of-the-art dataset in the maritime domain.

Dataset	Modality		Object Label	On-board Data	Area		Application	Sensors
	Image	Range			Coastal	Fresh		
MassMIND [8]	Y		Y	Y	Y	Y	Object Segmentation	IR cam
MaSTr1325, MODD [14]–[16], [21]	Y		Y	Y	Y		Object Segmentation	RGB cam, IMU
VAIS [9]	Y		Y		Y		Object Classification	IR cam, RGB cam
MARVEL [10]	Y		Y		N/A*	N/A	Object Classification	RGB cam
SeaShips [11]	Y		Y		Y		**Object Detection, Object Classification**	RGB cam
WSODD [12]	Y		Y		Y	Y	**Object Detection, Object Classification**	RGB cam
USVInland [22]	Y	Y		Y		Y	SLAM, Water segmentation, Stereo matching	LiDAR, Stereo cam, RADAR, IMU
NTNU [23]	Y	Y		Y	Y		Object Tracking**	LiDAR, RADAR, EO and IR cam
Pohang [24]	Y	Y		Y	Y		SLAM	LiDAR, Stereo cam, AHRS, GPS, IR cam, RADAR
Ours	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Object Detection, Object Classification**	**LiDAR, RGB cam, IMU, GPS**

* The images contain ships but collected by data mining from web sources.

** The public data contains trajectories of detected vehicles, not the raw data of sensors.

either individual or combined data from cameras, LiDAR, and RADAR. Given the importance of these datasets, there is a push to develop specialized datasets for the marine domain to support the advancement of marine autonomy.

Maritime object detection and classification datasets mainly consist of a **single sensor modality**, i.e., camera sensors, used for different purposes. Key datasets include the first visible and infrared (IR) ship image dataset for autonomous navigation compliance [9], a large-scale maritime dataset with over 2 million images detailing vessel information from a community site [10], and a dataset of common ship types from coastal surveillance [11]. Zhou et al. [12] introduced more variety with different water surface objects. However, most datasets were from stationary platforms, not from an **ego-centric perspective**. A significant onboard camera dataset exists [13] but is not public. Public datasets [14], [15], [16], [21] consist of several annotated videos collected by a real ASV platform, but these primarily focus on object segmentation with four classes—sea (water), sky, environment, and obstacle—lacking differentiation of in-water objects like buoys and ships. Nirgudkar et al. [8] presents a long wave IR (LWIR) dataset with categories including sky, water, obstacle, but still limited to a **single modality**.

Several **multimodal** datasets [22], [23], [24] are available, targeting different aspects of marine perception but not directly focusing on **object detection** and **classification**. Cheng et al. [22] covered inland waterway scenes using LiDAR, stereo cameras, RADAR, GPS, and IMUs, for water segmentation, SLAM, and stereo matching. Helgesen et al. [23] combined data from ten cameras, RADAR, and LiDAR for object tracking. Chung et al. [24] collected data from a diverse set of sensors over a 7.5-km route, aiming at SLAM and docking. Table 1 provides an overview of the discussed datasets compared to ours. This lack of datasets in the marine domain, specifically missing the key situational awareness tasks previously described, hampers progress in marine autonomy.

III. DATASET GENERATION

A. SENSOR CONFIGURATIONS

As shown in Fig. 2, we used our custom ASV *Catabot* (in three different configurations) and a human-driven boat installed with a sensor platform. The different configurations allow us to collect diverse data that includes different vehicle dynamics. The *Catabot* dimensions range from 1.08 to 2.68 m long, and from 1.40 to 1.67 m wide. The human-driven boat is 8.27 m long, 2.34 m wide. Both include a global positioning system (GPS)/compass and inertial measurement unit (IMU) with a flight controller unit, installed at the center line of the vehicle, to record proprioceptive data. We used a low-cost u-blox M8N GPS/Compass module. The flight controller hardware we used was a *Pixhawk 4* coupled with a 32-Bit Arm Cortex-M7 microcontroller with a 216-MHz clock speed and 2 MB of flash memory and 512 kB of RAM.

For exteroceptive data, we installed an RGB camera (Full-HD 1080P with CMOS OV2710 image sensor that can support IR during the nighttime) and a 64 channel LiDAR

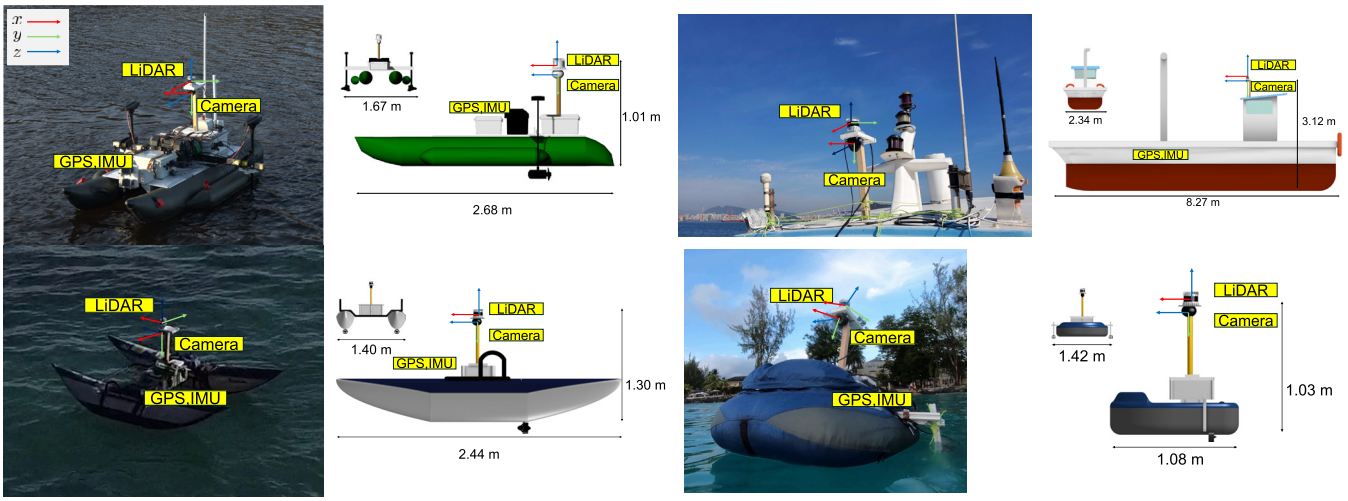
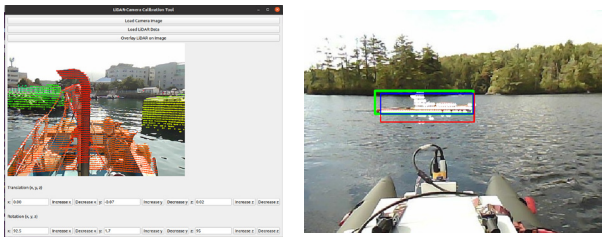


FIGURE 2. Data collection platform (top left): our custom ASV *Catabot*2, (top right): human-driven ship equipped with sensors, (bottom left): our custom ASV *Catabot*1, and (bottom right): our custom ASV *Catabot*5.



(a) (b)

FIGURE 3. Sensor suite calibration and annotation checking tool. (a) LiDAR and camera extrinsic calibration; and (b) point cloud (white) overlaid on the corresponding RGB image to check consistency over labels (green: image label, red: point cloud label, and blue: intersection).

(Ouster OS1-64 gen2). The two exteroceptive sensors were located at the center line of the vehicles to ensure a sufficient horizontal field of view (FoV, camera—91.8°; LiDAR—360° except for the blind sector due to occlusion caused by the vehicle structure) and vertical FoV (camera—75.5°; LiDAR—45°). The LiDAR has a range of 120 m with a horizontal resolution of 0.35° and vertical resolution of 0.7°, while the camera sensor has a 640 × 480 pixel resolution.

We performed intrinsic calibration of each sensor and an extrinsic calibration between camera and LiDAR based on [25] and [26]. We provide a custom tool for checking the extrinsic calibration parameters and overlay of multimodal data as shown in Fig. 3(a). We report the result of the calibration parameters for each sequence of the dataset.

B. DATA COLLECTION AND PROCESSING

We used a companion computer system (Intel NUC) and recorded proprioceptive (GPS, compass, and IMU) and exteroceptive (RGB camera and LiDAR) data via the robot operating system (ROS). Our Intel NUC computer with Ubuntu 18.04 installed has an Intel Core i7-8559U Processor (8M Cache, up to 4.50 GHz) with 1 TB of storage. The heterogeneous sensors operate at different time frequencies: we used a camera with a frequency of 30 Hz and LiDAR with a frequency of 10 Hz.

We collected relevant data in {sea, fresh} waters with varying environmental conditions {dusk, day, night}. We controlled the ASV via either 1) autonomous waypoint following or 2) manual driving, while we manually navigated the human-driven boat. Fig. 4 shows the trajectories during data collection. Our dataset covers collections conducted between 2021 to 2024 in different geographic locations: Lake Sunapee, NH, USA; Lake Mascoma, NH, USA; Busan Port, South Korea; and Holetown, Barbados.

We postprocess the camera and LiDAR data by extracting raw images and point clouds under time synchronization using the `MessageFilter` package [27].

C. GROUNDTRUTH GENERATION

We provide annotations of three in-water object classes based on the domain knowledge and navigation-oriented categorization: **ship**, **buoy**, and **other**, within the camera's FoV as well as the LiDAR's FoV. More specifically, 1) the **ship class** represents all marine vehicles defined according to the international traffic rule [7] as "every description of watercraft used or capable of being used as a means of transportation on water," including examples such as power-driven vessels, fishing boats, kayaks, yachts, and sailboats; 2) the **buoy class** represents floating objects as defined by the International Maritime Buoyage System [20] and includes any artificial objects serving as "aids to navigation," like cardinal, lateral, safe water, isolated danger, and special buoys with varying colors and shapes, such as ball and pillar types; and 3) the **other class** represents any in-water objects that can be risky to maritime navigation, for example, floating docks, and

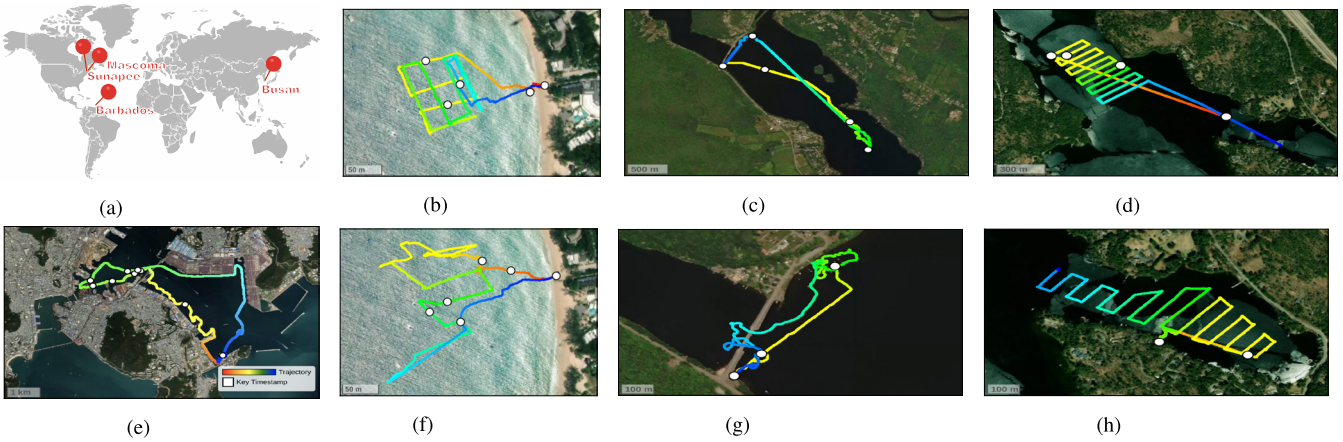


FIGURE 4. Data collection trajectories in different locations, navigating from red to blue. White points are key frames with objects encountered and corresponding annotations in the dataset, defined as "sequences." (a) Geographic locations. (b) Sea—Barbados 1. (c) Lake—Mascoma 1. (d) Lake—Sunapee 1. (e) Sea—Busan. (f) Sea—Barbados 2. (g) Lake—Mascoma 2. (h) Lake—Sunapee 2.

fishing nets. We provide ontology documentation for labeling annotation consistency and dataset usage.

For the images, we used the third party Amazon AWS Mechanical Turk annotation service in addition to the annotation by team members using the open-source Anylabeling [28] tool and model-assisted labeling using Meta Research's Segment Anything Model (SAM) [29]. For the LiDAR point clouds, we adapted an open-source labeling tool [30] for our purpose. We first conducted manual annotations and then resized them to bounding boxes that tightly contains the point cloud within it, while maintaining the yaw of the manually annotated bounding boxes. For both, we ran three rounds of annotation review by the expert team members for quality control.

We provide the label format in a standardized way along with converter implementations, such as You Only Look Once (YOLO) format, KITTI format, and unified normative, so that users can apply the dataset to different applications. The point cloud label contains $\{x, y, z, dx, dy, dz, \text{yaw}, \text{class}\}$ information. We only provide the yaw angle, assuming the roll and pitch remain approximately zero. Even if in rough water conditions, this assumption might not hold, roll and pitch information is typically not necessary for ASV 2-D navigation. For consistency of labeling in one frame of an image and a point cloud with its quality, we used a custom tool to extract the same object across the modalities [see Fig. 3(b)]. For the KITTI label format, we consider the annotation of an object as valid, only if they are located within the FoV of both camera and LiDAR, following the KITTI benchmark guideline [1].

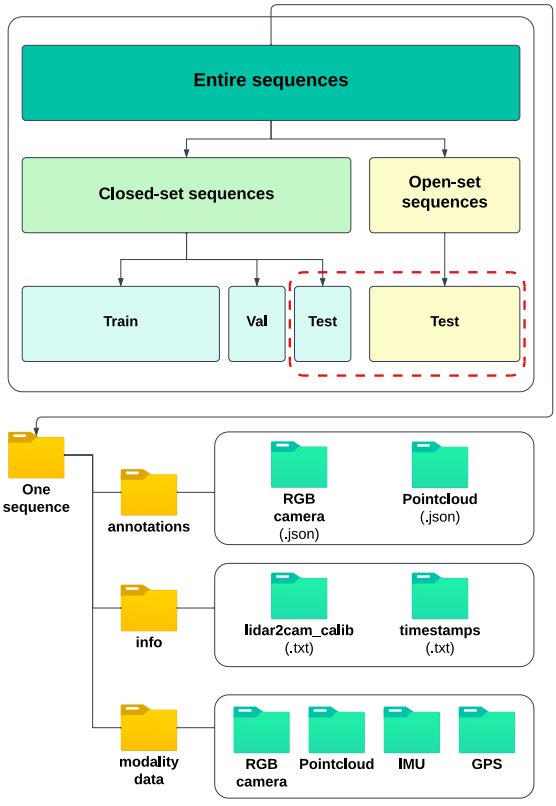


FIGURE 5. Overall dataset structure.

D. DATASET STRUCTURE

Fig. 5 shows the overall structure of our dataset, divided into three subsets: *train*, *validation*, and *test*. We define a sequence as a 60-s event involving object encounters at a specific geographical location, including Barbados, Busan, Lake Sunapee, and Lake Mascoma. For each sequence, we establish subdirectories based on annotations, information, and sensor modalities. In addition, we categorize sequences into *closed-set* (used for training and evaluation) and *open-set* (excluded from training and used only for evaluation).

Given the geographical coverage of the dataset—{**sea**: Barbados, Busan, **fresh**: Sunapee, Mascoma}—we first construct the *open-set* by selecting one sequence from

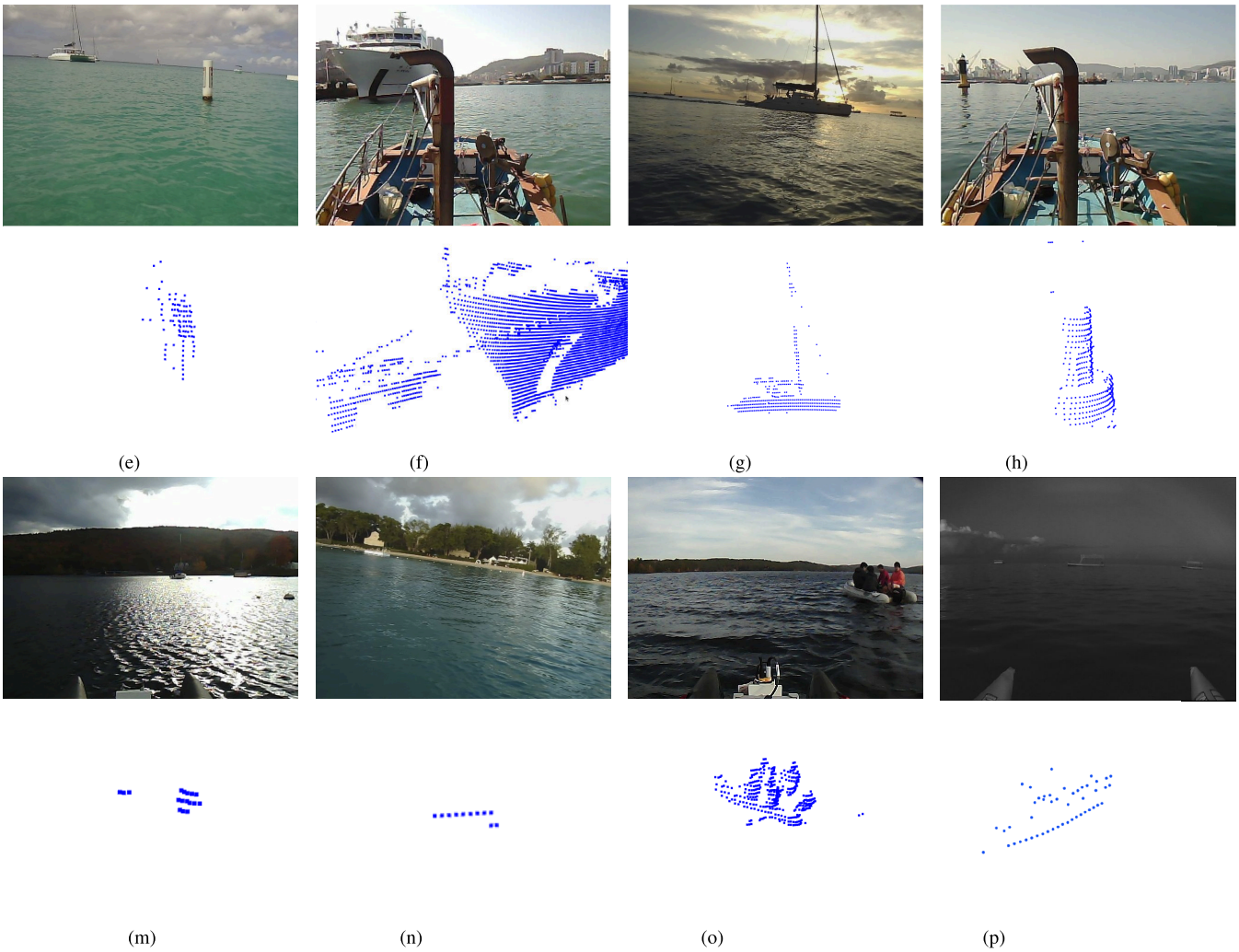


FIGURE 6. **In-water objects under varying environmental conditions in our dataset. Top: images. Bottom: point clouds. Note that the view angle of the point clouds is adjusted for the best visualization, regardless of the corresponding image of the object. (a) Class buoy—pillar. (b) Class ship—large vessel. (c) Class ship—yacht. (d) Class buoy—cardinal. (e) Class buoy—ball. (f) Class other—floating dock. (g) Class ship—raft with people. (h) Class ship—boat by IR.**

TABLE 2. **Labeled objects by class present in the dataset in the RGB image modality and the LiDAR modality.**

Class Name	Ship	Buoy	Other
Image Obj. Count	22874	11337	1833
LiDAR Obj. Count	22251	15692	1636

each location, resulting in a total of four sequences and 1376 frames total. Each selected sequence was chosen to reflect a challenging condition specific to its environment: Sunapee features multiple kayaks at far distances; Mascoma includes many boats and buoys (more than ten) under water surface glare; Barbados captures a sunset scenario; and Busan contains both a buoy and a boat approaching from a distance to close proximity. For the remaining 10 185 frames, which form the *closed-set*, at the sequence level, we randomly shuffle and split into *train*, *validation*, and *test* subsets using a 0.70, 0.15, and 0.15 ratio. This partitioning strategy enables fair quantifiable evaluation of both model performance and generalization capabilities for learning-based algorithms [31].

IV. DATASET CHARACTERISTICS

A. DATASET COMPOSITION

As shown in Fig. 6, our maritime perception dataset consists of various objects in water under varying conditions collected by the sensor platforms onboard ASVs or onboard a human-driven ship. This annotated, ego-perspective dataset is the first in the maritime domain, to the best of our knowledge with sufficiently large number of annotated frames (total 11 561). We believe it will be useful for training, validating, and benchmarking maritime perception.

Table 2 shows the annotated class breakdown in the RGB camera data and LiDAR data, where the predominant class

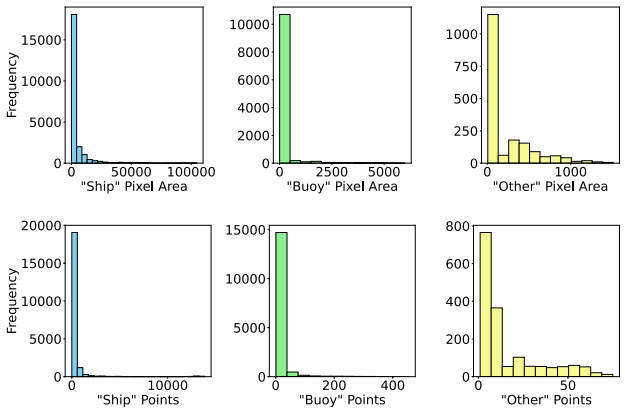


FIGURE 7. Distribution of labeled object pixel area in the RGB camera modality (top) and of LiDAR points (bottom) by class. Both pixel area and number of LiDAR points exhibit similar distributions.

in both modalities is "ship," followed by "buoy," and then "other." While both modalities of the dataset exhibit a class imbalance between the "ship" annotations and the other two classes, this imbalance naturally reflects the characteristics of coastal navigation environments represented in the dataset. Training performant learning-based models using this dataset may require strategies to address this natural imbalance—see Section VI. We characterize the annotation resolution, made via 2-D and 3-D bounding boxes, based on its pixel area [see Fig. 7(top)] and the number of LiDAR points [see Fig. 7(bottom)], respectively. This resolution is inherently limited by the underlying sensor resolution as well as other confounders related to the modality (e.g., illumination for RGB cameras) and others related to the maritime domain (e.g., in-water dynamics). Still, this approach gives insight into the amount of available sensor information upon which to detect and classify objects.

For the majority of objects, the annotation resolution is in the lowest bin, where the ship class has the highest average pixel area (mean: 4197.1, standard deviation: 10 194.2, and median: 794.0), followed by other (mean: 157.4, standard deviation: 551.7, and median: 28.0), and finally buoy (mean: 218.3, standard deviation: 301.2, and median: 38.0). Generally, the point cloud data follows the same trend where ships have the highest average point-cloud points (mean: 360.1, standard deviation: 1477.2, and median: 37.0), followed by other (mean: 15.8, standard deviation: 17.8, and median: 8.0), and then buoy (mean: 11.0, standard deviation: 35.6, and median: 2.0). Of note is the long-tailed nature of the distributions in Fig. 7, meaning that there is a large amount of heterogeneity within the same class.

In terms of environmental conditions, the data is composed of 79.9% for "day," 14.9% for "dusk," and 5.2% for "night." Dusk and night are imbalanced given the challenges in collecting data during that time. While we envision future work expanding the dataset to include a broader range of lighting conditions, we provide suggestions in Section VI to address this challenge together with the class imbalance challenge.

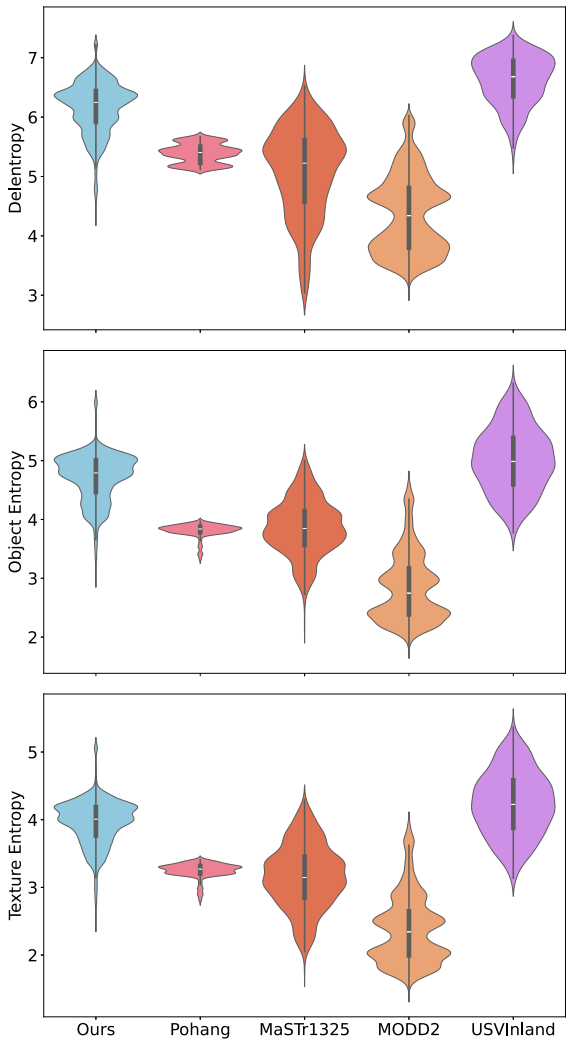


FIGURE 8. RGB image complexity comparisons between our dataset (light blue) and four other maritime perception image datasets Pohang [24], MaSTr1325 [21], MODD2 [15], and USVInland [22].

B. DATASET COMPLEXITY

As described in detail below, we propose novel metrics [e.g., birds-eye-view entropy with pillars (BEVE-Ps) and voxels (BEVE-V), and distance variability entropy (DVE)] in addition to existing metrics (e.g., image entropy and occlusion percentage) in the literature that quantitatively evaluate the dataset's characteristics with respect to the maritime domain to help analyze future benchmark algorithms.

1) IMAGE COMPLEXITY

Image entropy indicates the variation or complexity of an image at the grayscale distribution. In general, a low value corresponds to less edges and corners and possibly fewer

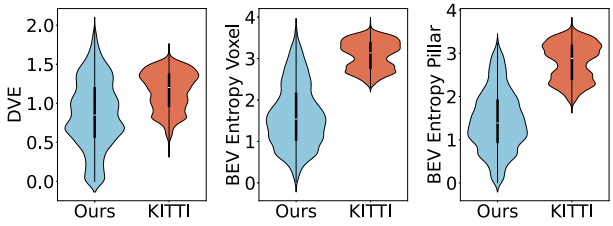


FIGURE 9. Point cloud complexity comparison between our dataset (light blue) and KITTI [1] (red-orange). Across DVE, birds-eye-view (BEV) entropy voxel, and BEV entropy pillar, our dataset shows a greater range of point cloud complexities.

interesting features, while a high value corresponds to an image with a significant amount of texture.

We evaluate image complexity with three entropy metrics: delentropy, object-level entropy, and texture-level entropy. For **delentropy** metric, we first applied the Sobel operator to approximate the gradients along the vertical and horizontal directions, and afterward calculated the Shannon entropy. We take inspiration from the evaluation criterion in the work by [32], an underwater dataset—where image-based object detection algorithms typically implement some preliminary edge detection processing. Note, instead of the Sobel filter, another edge detection algorithm, such as the Canny Edge detector, can work as well. The traditional **object entropy** and **texture entropy** metrics are similar in that they are directly calculating the Shannon entropy, but with different-sized template disks—object-level with a disk of 10 pixel radius and feature-level with a disk of 5 pixel radius. Here, there is no prior applied edge-detection-based filter.

Fig. 8 depicts the results of image complexity, according to the above three entropy metrics, for our dataset as well as for four other comparison datasets: Pohang [24], MaSTR1325 [21], MODD2 [15], and USVInland [22]. Compared to the Pohang dataset, our dataset includes more diverse imagery scenes. On the other hand, the image complexity of the USVInland dataset is comparable to our dataset—not surprising, given the various textures of nearby trees, rocks, tunnels, and houses in inland waters. While the MaSTR1325 and MODD2 datasets (both from the same authors) have a greater range of complexity compared to our dataset—much of their images have small objects (relative to image size) and due to observable off white-balancing, the pixel intensity values are within a smaller range—leading to many images corresponding to low entropy values. Our dataset shows a wide diversity of images, and with better on-camera white-balancing, our images have greater pixel intensity variations.

2) LiDAR COMPLEXITY

We introduce three entropy-based metrics to evaluate the spatial complexity of LiDAR-derived point clouds: BEVE-Ps, BEVE-Vs, and DVE. These metrics gauge how the point distribution spans discretized bins (pillars, voxels, or distance intervals), offering a detailed view of spatial variability in LiDAR data. Lower entropy values indicate that scene objects are more concentrated and clustered within a certain region. Conversely, higher entropy values indicate that there are objects more widely distributed or densely spread across the sensor range. A dataset with variation in entropy values represents its richness and complexity of the data.

BEVE-P and BEVE-V: These metrics measure *point cloud complexity* based on a discretized representation (pillar or voxel bins) of the LiDAR data. Formally, we define the metric as

$$\text{BEV} = -\sum_{i=1}^{M}\left[\frac{k_i}{K}\log\left(\frac{k_i}{K}\right)\right] \tag{1}$$

where i indexes each pillar or voxel, M is the total number of pillars or voxels, K represents the total number of points in the frame, and k_i is the count of points in each respective pillar or voxel. A lower BEV indicates that there is a concentration of objects in fewer bins; while a higher BEV suggests that there is a broader distribution of objects across multiple bins, reflecting a richer spatial arrangement.

DVE: This metric evaluates *point cloud complexity* based on radial distance from the LiDAR sensor, measuring how points are distributed across predefined radial distance intervals. We define the metric as

$$\text{DVE} = -\sum_{i=1}^{R}\left[\frac{n_i}{N}\log\left(\frac{n_i}{N}\right)\right] \tag{2}$$

where i indexes each predefined radial distance interval, R is the total number of distance intervals, N is the total number of points within the frame, and n_i is the number of points in each respective radial distance ring. A lower DVE suggests that points lie within fewer radial bands (indicating a simpler or more concentrated layout), whereas a higher DVE indicates that points are spread more extensively across different distances (denoting a more complex and broad-ranging scene).

Fig. 9 shows the proposed complexity metrics of the dataset within the collected point clouds, indicating that we have varying spatial distributions of in-water objects.

V. PERCEPTION BENCHMARKS

We ran the perception benchmark on our proposed datasets on detection tasks, i.e., **object detection** and **object classification** and developed the necessary conversion tools. We used a computer equipped with an Intel i7-7820X 8-core 3.6-GHz processor, 32-GB RAM, and NVIDIA GPU RTX 3090 Ti with 24-GB VRAM. We evaluated benchmark algorithms, offering insights into the applicability and adaptability of these benchmarks in the maritime domain.

A. IMAGE-BASED BENCHMARKS

While many real-time (RT) RGB image object detection approaches exist, we selected two representative models to provide a benchmarking of this dataset upon: YOLOv9 [33] and RT-DETR [34]. We specifically benchmark using RT

TABLE 3. Performance breakdown of 2 benchmark 2-D image object detectors by class. mAP is reported via the aggregated IoU threshold from (0.5 to 0.95) per class.

Model	Aggregated mAP (0.5:0.95)		"ship" mAP (0.5:0.95)		"buoy" mAP (0.5:0.95)		"other" mAP (0.5:0.95)	
	Val	Test	Val	Test	Val	Test	Val	Test
YOLOv9 [33]	**0.54**	**0.42**	**0.83**	**0.65**	**0.42**	**0.34**	**0.36**	**0.27**
RT-DETR [34]	0.21	0.16	0.45	0.36	0.13	0.10	0.04	0.01

TABLE 4. Comparison of validation and test results for LiDAR-based benchmarks (IoU thresholds of 0.7 and 0.5) for ship class objects. Green highlights the best performance across fusion methods, while yellow indicates the best performance among LiDAR-only methods.

Model	Modality	BEV AP (0.7)		BEV AP (0.5)		3D AP (0.7)		3D AP (0.5)	
		Val	Test	Val	Test	Val	Test	Val	Test
PointPillars [35]	LiDAR-only	28.01	17.32	57.22	50.77	4.23	3.14	30.30	30.07
SECOND [36]	LiDAR-only	34.29	27.68	56.95	52.36	8.93	10.17	40.70	40.67
PointRCNN [37]	LiDAR-only	3.24	3.11	23.93	21.48	0.32	0.43	2.91	2.66
PV-RCNN [38]	LiDAR-only	19.11	9.99	42.40	38.65	3.79	9.09	23.64	16.54
Voxel-RCNN [39]	LiDAR-only	33.36	27.50	54.55	50.69	12.90	13.46	41.96	43.03
TED-S [40]	LiDAR-only	**49.64**	**37.36**	**70.56**	**55.09**	**36.88**	**26.99**	**60.82**	**46.10**
PointPainting [41]	Fusion	30.51	25.42	57.92	46.10	10.05	12.95	42.54	37.67
CLOCs [42]	Fusion	32.02	21.76	56.68	49.47	9.69	8.28	45.38	41.10
Focal Conv-F [43]	Fusion	37.48	**36.36**	**61.69**	**54.55**	19.83	15.58	47.79	**45.45**
TED-M [40]	Fusion	**50.32**	32.40	54.05	43.64	**30.24**	**27.69**	**53.87**	42.91

detectors as the ego-centric ASV perspective of this dataset lends itself to use in RT, on-board object detection use cases. Based on that criteria, we selected a model from the popular YOLO object detector lineage, which uses a convolutional neural network (CNN) backbone approach and a newer transformer-backbone approach based on the detection transformer [44] (DETR), that was adapted for RT use.

We trained both models for 300 epochs and used the default hyperparameters from the YOLOv9 [45] and RT-DETR with HGNetv2 backbone [46] open-source implementations. From their reference implementation, we applied a confidence threshold of 0.25 and an intersection over union (IoU) threshold of 0.45 for nonmaximum suppression to postprocess outputs before compiling results. For consistent comparison across 2-D object detection methods, we used the mean average precision (mAP) metric. Validation and test set results are in Table 3 and example detections are in Fig. 10. The qualitative examples are from 3 of the dataset's locations: Barbados, Lake Mascoma, and Busan Port, to show several multiobject encounters with ship, buoy, and other labeled objects.

From the results of both models, qualitative and quantitative, out-of-the-box models have room for improvement—especially on the buoy and other classes. It is clear that (a) there are relevant image-only features to train object detection models and (b) that this dataset represents a challenging detection task, characteristic of the maritime ASV environment.

The heterogeneity of maritime objects, variable environmental conditions, and in-water dynamics make this a difficult RGB-camera-only robotic vision problem—one that the addition of LiDAR data can help address.

B. LiDAR-BASED DEEP LEARNING BENCHMARKS

We analyzed the performance of LiDAR-based methods on 3-D and BEV detection in the maritime domain. We selected six state-of-the-art LiDAR-only 3-D object detection models—based on the following categorizations.

1) Voxel-Based: PointPillars [35], SECOND [36], Voxel-RCNN [39], and TED-S [40].
2) Point-Based: PointRCNN [37].
3) Point-Voxel-Based: PV-RCNN [38].

We adapted open-source libraries, including OpenPCDet [47], as well as implementations from other repositories [36], [40], [42], to enable benchmark comparisons tailored to our maritime dataset.

We followed each paper's guideline on setting the hyperparameters and used the suggested values when possible. We increased the point cloud range and the voxel size to account for the longer distances between the ASV and obstacles. We trained each method for 200 epochs with early stopping once the model stopped improving.

For consistent comparisons across LiDAR-only and fusion methods, we evaluated and reported performance for objects within both the camera and LiDAR FoV, consistent with KITTI benchmarks [1]. Note that our ground truth labeling provides a 360° FoV from the LiDAR used on our platforms. We compared performance based on average precision (AP) at IoU thresholds of 0.7 and 0.5, evaluated for both BEV

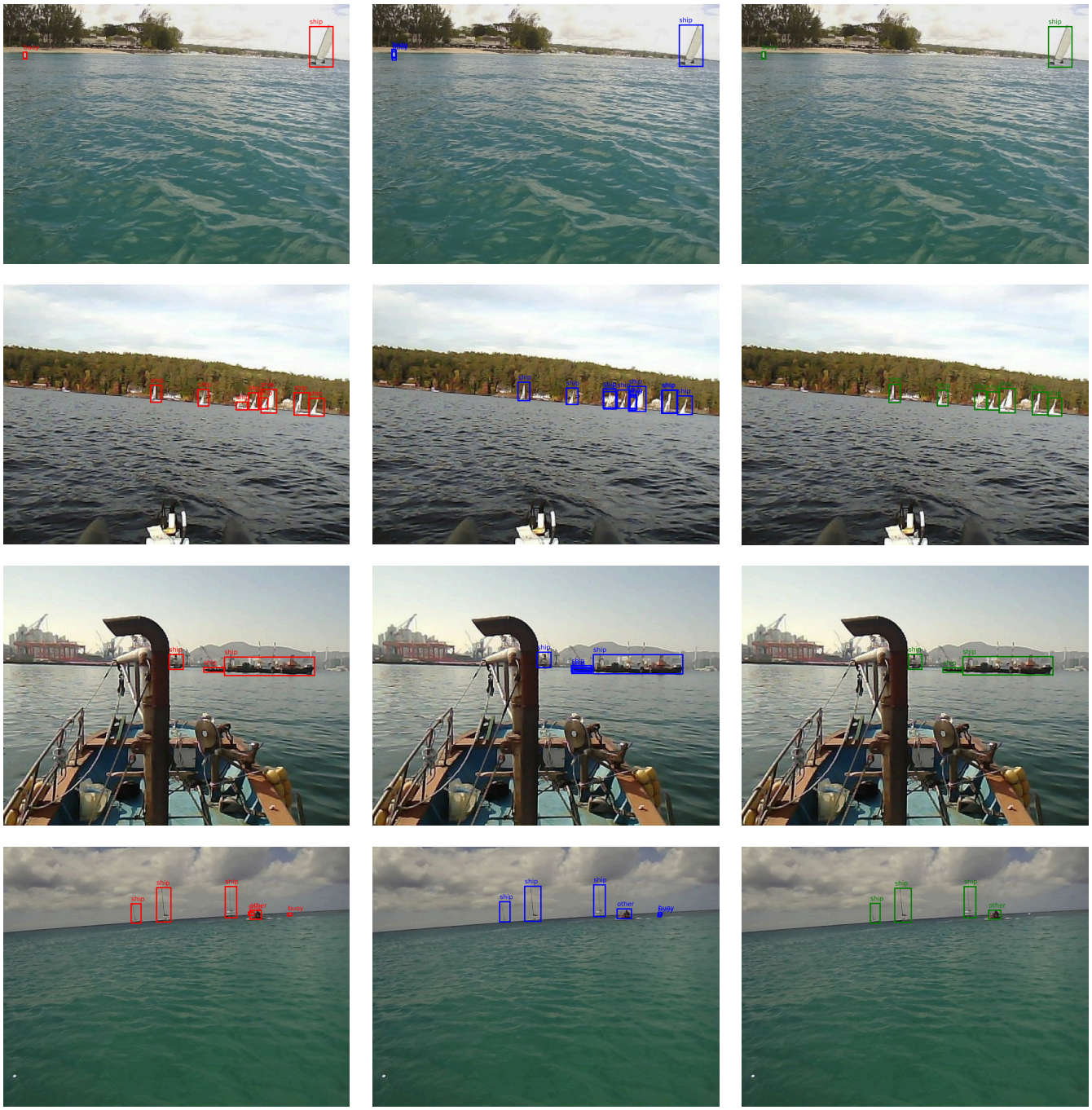


FIGURE 10. **Detection examples from the benchmark image models trained on the RGB image portion of the dataset, where the left column is groundtruth (red), the middle column is RT-DETR (blue), and the right column is YOLOv9 (green). While the models did learn to predict many of the classes, there is still much room for improvement that robotic perception methods adapted to the ASV domain could begin to address using this dataset.**

and 3-D detection. We focused on the **ship** class for performance comparison due to the sparsity and challenges posed by features associated with small objects in the **buoy** and **other** classes, which LiDAR typically returns as 1–2 points, as shown in Fig. 7(bottom).

The evaluation results (see Table 4) for BEV detection are comparable to those of previous work [17], which used simulation results tested on 2-D. Instead, our benchmark comparison extends applicability to the 3-D domain with real-world data. Among LiDAR-only methods, TED-S achieved the highest performance across both BEV AP and 3-D AP metrics, outperforming other state-of-the-art approaches. These strong results may be attributed to its transformation-equivariant sparse convolution pooling

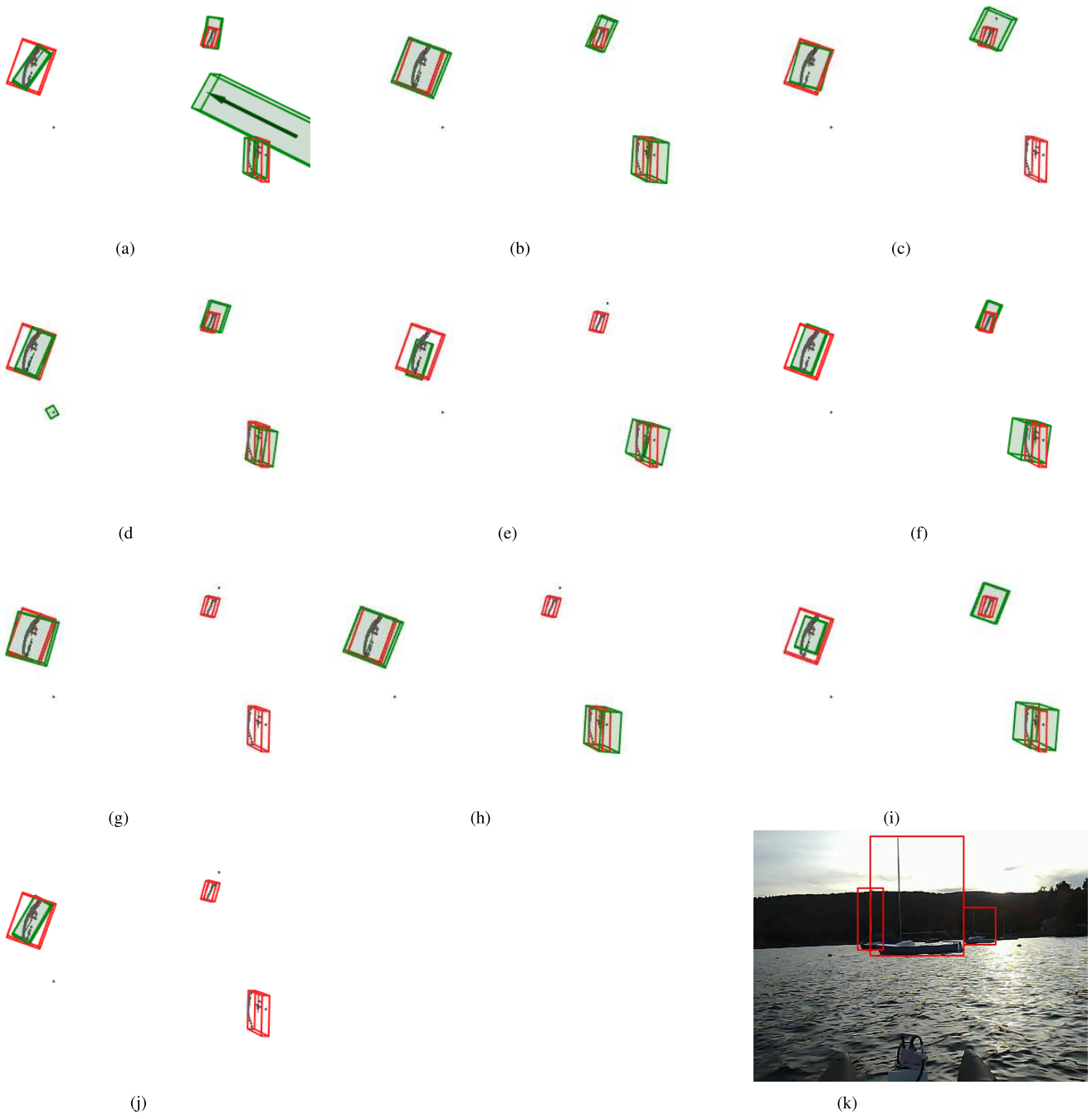


FIGURE 11. Qualitative comparison of LiDAR-based and fusion object detection benchmarks tested on our dataset, shown from a bird's-eye view. The evaluated objects belong to the ship class within the FoV of both the camera and LiDAR, with ground truth bounding boxes depicted in red and predicted bounding boxes in green. (a) PointPillars. (b) SECOND. (c) PointRCNN. (d) PV-RCNN. (e) Voxel-RCNN. (f) TED-S. (g) PointPainting. (h) CLOCs. (i) Focals Conv-F. (j) TED-M. (k) Image of evaluated objects.

and transformation-invariant voxel pooling modules, which enable learning of robust, transformation-equivariant voxel features. In addition, its distance-aware data augmentation strategy enhances detection of distant objects—an important characteristic for in-water maritime scenarios. Aside from TED-S, SECOND consistently demonstrated strong BEV AP performance. This may be attributed to its voxel-based representation and efficient sparse convolution, which effectively captures large-scale geometric features. These characteristics make SECOND particularly robust for BEV representations, where preserving spatial structure is critical. In contrast, Voxel-RCNN performed relatively well in 3-D AP metrics.

Its performance stems from leveraging high-resolution voxel grids combined with an accurate region proposal network, enabling more precise object localization in 3-D space. On the other hand, PointRCNN, which relies solely on raw point clouds and bypasses voxelization, is limited in its ability to efficiently extract global features, making it less effective in sparse maritime environments. Meanwhile, PV-RCNN, employing a hybrid approach that combines voxel-based feature extraction (for global context) with raw point-cloud features (for local precision), was better than PointRCNN by balancing global and local feature extraction. PointPillars showed relatively lower performance, particularly in 3-D AP. This is likely due to its reliance on a pillar-based pseudo-image representation that flattens vertical structure early in the pipeline.

C. FUSION-BASED DEEP LEARNING BENCHMARKS

We evaluated the following three state-of-the-art 3-D object detection fusion methods.

1) *Sequential Fusion:* PointPainting [41] based on DeepLabV3 [48] and PointPillars.
2) *Decision-Level Fusion:* CLOCs [42] based on the detection of YOLOv9 [33] and SECOND.
3) *Feature-Level Fusion:* Focal Conv-F [43] and TED-M [40].

As shown in Table 4, Focal Conv-F and TED-M achieved the best overall results among fusion-based methods across both BEV AP and 3-D AP metrics. Focal Conv-F's effectiveness may be attributed to the integration of complementary sensor modalities through focal sparse convolutions, enabling robust spatial reasoning and precise object localization. TED-M builds upon TED-S by incorporating appearance features from RGB images, offering further improvements in some cases. However, as noted in [40] and [49], our results indicate that incorporating camera data does not uniformly enhance detection performance. Notably, TED-M's marginal gains come at the cost of increased system complexity, as it requires generating pseudo-LiDAR points from camera images. These image-derived points depend on depth estimation [50], which can be particularly noisy for distant objects. This noise partly explains why distant or hard-to-detect targets sometimes see minimal benefit—or even slight performance degradation—with fusion. Such drops can also be attributed to sensor misalignment [51], [52], a challenge observed in the maritime domain and further discussed in Section VI. Other fusion methods such as CLOCs and PointPainting performed competitively but their performance lagged at stricter IoU thresholds for 3-D AP. CLOCs, which integrates predictions from multiple backbones, showed reduced performance in scenarios requiring high precision, likely due to a weaker emphasis on fine-grained feature alignment. Similarly, PointPainting's reliance on segmentation quality and alignment resulted in lower performance in 3-D AP metrics compared to Focal Conv-F.

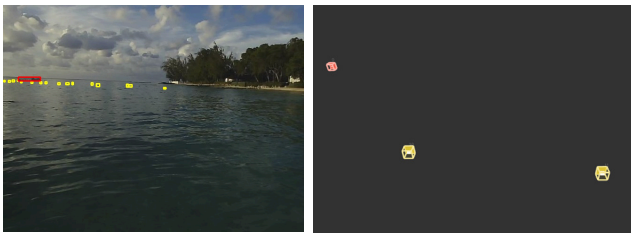


FIGURE 12. Challenging sparsity example from Barbados sequence in our dataset—(yellow) buoys and (orange) floating dock. Although there are many objects in the image (left), the LiDAR measurement has only 3 objects while each of them has 1 point inside the bounding box (right).

We also provide a qualitative analysis across the 3-D object detection benchmarks. Fig. 11 illustrates the results of a sequence (Mascoma Lake) on our *open-set* test split, which were excluded from all training steps. Consistent with the quantitative evaluation, TED-S and Focal Conv-F exhibited strong performance for ship detection.

VI. DISCUSSION

Based on our contribution of the first multimodal dataset in the maritime domain and its utility for deep learning-based approaches, we identify and provide insights into the challenges and open problems for future tasks aimed at enhancing robust perception systems in maritime environments. Furthermore, we hope this dataset will provide the research community with a starting point to develop robust, novel methods for ASV perception. Given this work, it is our continuing hypothesis that multimodal methodologies are essential for the development of robust ASV situational awareness given in-water dynamics, environment heterogeneity, and failures being inadmissible. In the following paragraphs, we will discuss open challenges to the development of these methods, which include: sparsity, generalizability, and misalignment.

Maritime environments often feature **sparse** point clouds due to objects located at long distances and unstable measurements affected by the motions of both ego and target vehicles, as noted by Jeong and Li [53]. Current detection methods struggle to learn features from such minimal data, particularly for buoys and small objects. This highlights the need for models capable of accurately detecting and classifying objects even under sparse conditions. For example, as shown in Fig. 12, even a single LiDAR-detected point representing an object such as a buoy could lead to a collision if ignored. This differs from other domains where they often use thresholds for a minimum number of points.

Generalizability remains another significant open challenge. For instance, LiDAR-based deep learning models (and many RT image-based deep learning models) use anchors, which represent the predefined dimensions of bounding boxes, to enhance the accuracy and efficiency of object predictions. However, as shown by the range of the length (0.1–123.5 m), width (0.1–81.1 m), and height (0.1–35.7 m)

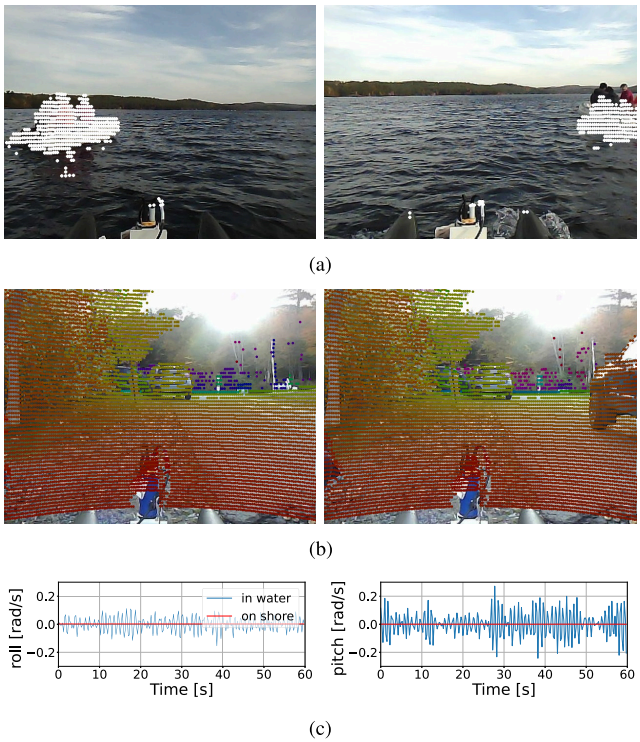


FIGURE 13. Comparison of LiDAR-camera alignment and motion-induced effects during the same deployment date. (a) Sensor misalignment occurs in water due to motion, despite the same calibration. (b) Stable alignment is preserved on shore, and (c) quantitative analysis of motions (roll and pitch) during 60 s as a sequence time. Note that the point clouds are colored for the best visibility.

of the LiDAR annotations, object sizes in the maritime domain vary greatly—from small fishing boats to large commercial ships—all defined as ''ships'' under international maritime traffic rules [7]. Therefore, if one does not carefully choose hyperparameter values, such as anchor dimensions, it may degrade the performance of detection benchmarks. Furthermore, all the point cloud detection benchmarks we utilized in our study relied on preset point cloud ranges. However, we observed cases where detected point clouds lay beyond the sensor's nominal range (e.g., exceeding 120 m), particularly in open sea conditions. Aligned with maritime navigation principles on focusing on early detection and taking large actions in ample time, one must thoughtfully select predefined ranges and sizes. These parameters strictly constrain current learning-based methods, underscoring the need for models, such as anchor-free approaches, which can adapt to varying detection ranges and object dimensions.

Another challenge for generalization is **class imbalance**, as noted in Section IV-A. This imbalance naturally reflects real-world coastal navigation environments. To mitigate its effects during model development, we recommend incorporating class-aware strategies such as targeted data augmentation (e.g., oversampling of rare object classes, copy-paste methods, or simulation-based generation) and loss reweighting approaches (e.g., focal loss or class-balanced loss functions). These methods can enhance detection performance on minority classes without compromising the dataset's representativeness. Furthermore, although the current version represents an initial contribution to the community, there are strategies to address the limited number of samples collected under **low-light conditions**. It is possible to supplement the dataset with style-transfer methods or domain adaptation techniques to simulate nighttime environments and improve model robustness across varying illumination scenarios.

Robustness against **misalignment** presents additional open challenges in the maritime domain. As observed in ground-based applications [51], [52], spatial and temporal misalignment is also prevalent in maritime environments. This misalignment arises from factors such as noisy extrinsic parameters and the relative motion between ego and target vehicles on the water surface, as illustrated in Fig. 13. We compared ASV behavior in in-water versus on-shore conditions. As shown in Fig. 13(c), during the same deployment operation, the in-water scenario exhibited significantly higher motion variability than the on-shore case (Levene's test: p-value $<$ 0.01 for both roll and pitch), leading to misalignments. Furthermore, mechanical misalignments are particularly difficult to correct onboard due to the lack of fixed environmental features and the continuous motion caused by hydrodynamic forces. These challenges underscore the need for online, in-water calibration methods to improve system robustness.

In addition, annotations in the maritime domain naturally suffer from misalignment. Our dataset primarily considers z-axis orientation (i.e., yaw) during the labeling process. However, pitch and roll can significantly impact object detection and state estimation—especially in maritime settings, where dynamic and nonstationary conditions differ greatly from those in other domains (e.g., flat road surfaces). Generating accurate ground truth for pitch and roll remains a major challenge but is essential for improving detection and tracking performance in such environments. Addressing this open problem is likely to be a key prerequisite for developing robust multimodal fusion methods in maritime surface applications.

VII. CONCLUSION

This article introduces the first publicly accessible multimodal perception dataset for autonomous maritime navigation, focusing on in-water obstacles within aquatic environments to enhance situational awareness for ASVs. Our dataset, which includes a diverse range of in-water objects encountered under varying environmental conditions, aims to bridge the research gap in marine robotics by providing a multimodal, annotated, and ego-centric perception dataset for object detection and classification. We also demonstrate the applicability of the proposed dataset using open-source deep learning-based perception algorithms that have proven successful in other domains. In addition, the development

and analysis of this dataset offer foundational insights for advancing perception tasks in the maritime domain.

Future work will focus on designing adaptable and robust deep learning models and systems capable of addressing domain-specific complexities while aligning with maritime best practices. Alongside this, we also plan to integrate additional sensor configurations under diverse weather conditions (e.g., rain and snow), such as marine RADAR and wide-field-of-view cameras for supplementary data collection. Furthermore, we plan to extend this work to the object tracking task and continue to explore multimodal modeling for ASV perception using this dataset. These advancements will be crucial for enhancing situational awareness, safety, and efficiency in real-world autonomous maritime systems, addressing high-impact societal needs such as search and rescue, environmental monitoring, and transportation.

ACKNOWLEDGMENT

The authors would like to thank Kizito Masaba, Julien Blanchet, Haesang Jeong, and Capt. Jinmyeong Lee for help with field experiments, and Korea Maritime and Ocean University, McGill Bellairs Research Institute, and the Eliassen family for access to the experimental sites. They would also like to thank Dongha Chung and Jinhwan Kim from KAIST for providing tips on publicly available datasets.

(Mingi Jeong and Arihant Chadda contributed equally to this work.)

REFERENCES

[1] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1231–1237, Sep. 2013.

[2] P. Sun et al., "Scalability in perception for autonomous driving: Waymo open dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2443–2451.

[3] H. Caesar et al., "NuScenes: A multimodal dataset for autonomous driving," 2019, *arXiv:1903.11027*.

[4] *Review of Maritime Transport 2021*., United Nations Conference on Trade and Development, Geneva, Switzerland, 2022.

[5] *Unmanned Surface Vehicle Market Size, Share, Competitive Landscape and Trend Analysis Report By Size, By Application, By Mode of Operation: Global Opportunity Analysis and Industry Forecast, Pp. 2023–2032*, Allied Market Research, Delaware, USA, 2023, pp. 2023–2032.

[6] C. Dominguez-Péry, L. N. R. Vuddaraju, I. Corbett-Etchevers, and R. Tassabehji, "Reducing maritime accidents in ships by tackling human error: A bibliometric review and research agenda," *J. Shipping Trade*, vol. 6, no. 1, p. 20, Nov. 2021.

[7] *Convention on the International Regulations for Preventing Collisions At Sea, 1972 (COLREGs)*, International Maritime Organization, London, U.K., 1972.

[8] S. Nirgudkar, M. DeFilippo, M. Sacarny, M. Benjamin, and P. Robinette, "MassMIND: Massachusetts maritime INfrared dataset," *Int. J. Robot. Res.*, vol. 42, nos. 1–2, pp. 21–32, Jan. 2023.

[9] M. M. Zhang, J. Choi, K. Daniilidis, M. Wolf, and C. Kanan, "VAIS: A dataset for recognizing maritime imagery in the visible and infrared spectrums," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2015, pp. 10–16.

[10] E. Gundogdu, B. Solmaz, V. Yücesoy, and A. Koç, "MARVEL: A large-scale image dataset for maritime vessels," in *Proc. Asian Conf. Comput. Vis.*, 2017, pp. 165–180.

[11] Z. Shao, W. Wu, Z. Wang, W. Du, and C. Li, "SeaShips: A large-scale precisely annotated dataset for ship detection," *IEEE Trans. Multimedia*, vol. 20, no. 10, pp. 2593–2604, Oct. 2018.

[12] Z. Zhou et al., "An image-based benchmark dataset and a novel object detector for water surface object detection," *Frontiers Neurorobotics*, vol. 15, Sep. 2021, Art. no. 723336.

[13] P. Kaur et al., "Sea situational awareness (SeaSAw) dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 2578–2586.

[14] M. Kristan, V. Sulic Kenk, S. Kovacic, and J. Pers, "Fast image-based obstacle detection from unmanned surface vehicles," *IEEE Trans. Cybern.*, vol. 46, no. 3, pp. 641–654, Mar. 2016.

[15] B. Bovcon, R. Mandeljc, J. Perš, and M. Kristan, "Stereo obstacle detection for unmanned surface vehicles by IMU-assisted semantic segmentation," *Robot. Auto. Syst.*, vol. 104, pp. 1–13, Jun. 2018.

[16] B. Bovcon, J. Muhovic, D. Vranac, D. Mozetic, J. Pers, and M. Kristan, "Mods, "MODS—A USV-oriented object detection and obstacle segmentation benchmark," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 8, pp. 13403–13418, Aug. 2022.

[17] J. Lin, P. Diekmann, C.-E. Framing, R. Zweigel, and D. Abel, "Maritime environment perception based on deep learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 9, pp. 15487–15497, Sep. 2022.

[18] L. Trinh, S. Mercelis, and A. Anwar, "A comprehensive review of datasets and deep learning techniques for vision in unmanned surface vehicles," 2024, *arXiv:2412.01461*.

[19] T. Clunie, M. DeFilippo, M. Sacarny, and P. Robinette, "Development of a perception system for an autonomous surface vehicle using monocular camera, LiDAR, and marine RADAR," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2021, pp. 14112–14119.

[20] *IALA Maritime Buoyage System*, U.K. Hydrographic Office, Taunton, U.K., 2018.

[21] B. Bovcon, J. Muhovi , J. Perš, and M. Kristan, "The MaSTr1325 dataset for training deep USV obstacle detection models," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Apr. 2019, pp. 3431–3438.

[22] Y. Cheng, M. Jiang, J. Zhu, and Y. Liu, "Are we ready for unmanned surface vehicles in inland waterways? The USVInland multisensor dataset and benchmark," *IEEE Robot. Autom. Lett.*, vol. 6, no. 2, pp. 3964–3970, Apr. 2021.

[23] Ø. K. Helgesen, K. Vasstein, E. F. Brekke, and A. Stahl, "Heterogeneous multi-sensor tracking for an autonomous surface vehicle in a littoral environment," *Ocean Eng.*, vol. 252, May 2022, Art. no. 111168.

[24] D. Chung, J. Kim, C. Lee, and J. Kim, "Pohang canal dataset: A multimodal maritime dataset for autonomous navigation in restricted waters," *Int. J. Robot. Res.*, vol. 42, no. 12, pp. 1104–1114, Oct. 2023.

[25] J. Beltrán, C. Guindel, A. de la Escalera, and F. García, "Automatic extrinsic calibration method for LiDAR and camera sensor setups," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 10, pp. 17677–17689, Oct. 2022.

[26] G. Pandey, J. McBride, S. Savarese, and R. Eustice, "Automatic targetless extrinsic calibration of a 3D LiDAR and camera by maximizing mutual information," in *Proc. AAAI Conf. Artif. Intell.*, Sep. 2021, pp. 2053–2059.

[27] (2023). *ROS Message Filter*. [Online]. Available: https://wiki.ros.org/messagefilters

[28] V. A. Nguyen. (2023). *Anylabeling-Effortless Data Labeling With Ai Support*. [Online]. Available: https://github.com/vietanhdev/anylabeling

[29] A. Kirillov et al., "Segment anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 4015–4026.

[30] W. Zimmer, A. Rangesh, and M. Trivedi, "3D BAT: A semi-automatic, Web-based 3D annotation toolbox for full-surround, multi-modal data streams," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2019, pp. 1816–1821.

[31] M. Najibi et al., "Motion inspired unsupervised perception and prediction in autonomous driving," in *Proc. ECCV*, vol. 13698. Cham, Switzerland: Springer, 2022, pp. 424–443.

[32] O. Álvarez-Tuñón et al., "MIMIR-UW: A multipurpose synthetic dataset for underwater navigation and inspection," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, May 2023, pp. 6141–6148.

[33] C.-Y. Wang, I.-H. Yeh, and H.-Y. Mark Liao, "YOLOv9: Learning what you want to learn using programmable gradient information," 2024, *arXiv:2402.13616*.

[34] Y. Zhao et al., "DETRs beat YOLOs on real-time object detection," 2023, *arXiv:2304.08069*.

[35] A. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "PointPillars: Fast encoders for object detection from point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 12689–12697.

[36] Y. Yan, Y. Mao, and B. Li, "SECOND: Sparsely embedded convolutional detection," *Sensors*, vol. 18, no. 10, p. 3337, Oct. 2018. [Online]. Available: https://www.mdpi.com/1424-8220/18/10/3337
[37] S. Shi, X. Wang, and H. Li, "PointRCNN: 3D object proposal generation and detection from point cloud," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 770–779.
[38] S. Shi et al., "PV-RCNN: Point-voxel feature set abstraction for 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10526–10535.
[39] J. Deng, S. Shi, P. Li, W. Zhou, Y. Zhang, and H. Li, "Voxel R-CNN: Towards high performance voxel-based 3D object detection," 2020, *arXiv:2012.15712*.
[40] H. Wu, C. Wen, W. Li, X. Li, R. Yang, and C. Wang, "Transformation-equivariant 3D object detection for autonomous driving," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, 2023, pp. 2795–2802.
[41] S. Vora, A. H. Lang, B. Helou, and O. Beijbom, "PointPainting: Sequential fusion for 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4604–4612.
[42] S. Pang, D. Morris, and H. Radha, "CLOCs: Camera-LiDAR object candidates fusion for 3D object detection," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2020, pp. 10386–10393.
[43] Y. Chen, Y. Li, X. Zhang, J. Sun, and J. Jia, "Focal sparse convolutional networks for 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5418–5427.
[44] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 213–229, doi: 10.1007/978-3-030-58452-8_13.
[45] C.-Y. Wang and H.-Y. M. Liao. (2023). *YOLOv9: Implementation of Paper*. [Online]. Available: https://github.com/WongKinYiu/yolov9
[46] (2024). *Ultralytics*. [Online]. Available: https://github.com/ultralytics/ultralytics
[47] (2020). *Openpcdet: An Open-source Toolbox for 3D Object Detection From Point Clouds*. [Online]. Available: https://github.com/open-mmlab/OpenPCDet
[48] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.
[49] V. Vats, M. Binta Nizam, and J. Davis, "VaLID: Verification as late integration of detections for LiDAR-camera fusion," 2024, *arXiv:2409.15529*.
[50] M. Hu, S. Wang, B. Li, S. Ning, L. Fan, and X. Gong, "PENet: Towards precise and efficient image guided depth completion," in *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, May 2021, pp. 13656–13662.
[51] K. Yu et al., "Benchmarking the robustness of LiDAR-camera fusion for 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2023, pp. 3188–3198.
[52] S. Pang, D. Morris, and H. Radha, "Fast-CLOCs: Fast camera-LiDAR object candidates fusion for 3D object detection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 3747–3756.
[53] M. Jeong and A. Q. Li, "Efficient LiDAR-based in-water obstacle detection and segmentation by autonomous surface vehicles in aquatic environments," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2021, pp. 5387–5394.

MINGI JEONG (Member, IEEE) is currently pursuing the Ph.D. degree with the Reality and Robotics Laboratory, Department of Computer Science, Dartmouth College, Hanover, NH, USA.

His current research interests include autonomous navigation, multirobot systems, and maritime collision avoidance decision making.

ARIHANT CHADDA received the Bachelor of Arts degree in computer science from the Reality and Robotics Laboratory, Dartmouth College, Hanover, NH, USA, in 2022.

He is currently a Senior Data Scientist of applied research with In-Q-Tel, Tysons, VA, USA. His current research interests include developing robust digital and physical autonomous systems.

ZIANG REN received the M.S. degree from Dartmouth College, Hanover, NH, USA, in 2024. He is currently pursuing the Ph.D. degree in computer science with the MobileX Laboratory, Columbia University, New York City, NY, USA.

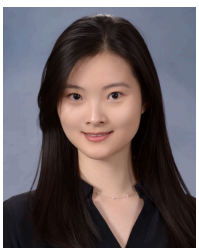
He was at the Reality and Robotics Laboratory, Dartmouth College, during the M.S. degree. His research interests include 3-D vision, computational imaging, and robot perception.

LUYANG ZHAO (Member, IEEE) received the Ph.D. degree in computer science from Dartmouth College, Hanover, NH, USA, in 2025.

She will join the Department of Electrical and Computer Engineering, Clemson University, Clemson, SC, USA, as an Assistant Professor. Her research interests include soft modular robotics, computational design, robot perception, and learning-based control.

HAOWEN LIU received the M.S. degree from Dartmouth College, Hanover, NH, USA, in 2023. She is currently pursuing the Ph.D. degree in computer science with the University of Maryland at College Park, College Park, MD, USA.

She was at the Reality and Robotics Laboratory, Dartmouth College, during the M.S. degree. Her research interests include video understanding and embodied AI.



AIWEI ZHANG received the Bachelor of Arts degree in computer science from the Reality and Robotics Laboratory, Dartmouth College, Hanover, NH, USA, in 2025.

Her current research interests include social computing and machine learning with applications in healthcare, community-centered mental health support, and digital well-being.

YITAO JIANG received the Master of Engineering Management degree from the Thayer School of Engineering, Dartmouth College, Hanover, NH, USA, in 2024, where he is currently pursuing the Ph.D. degree.

His research interests include actuators, modular tensegrity, soft robotic systems, swarm robotics driven by large language models, and robots inspired by biological locomotion.

SABRIEL ACHONG is currently pursuing the bachelor's degree in economics with Dartmouth College, Hanover, NH, USA.

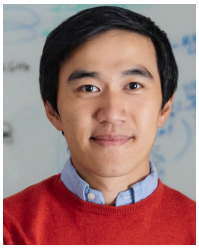
Her current research interests include artificial intelligence, applied statistics, and agent-based modeling.

SAMUEL LENSGRAF received the Ph.D. degree in computer science from Dartmouth College, Hanover, NH, USA, in 2024.

He is currently a Research Scientist with Florida Institute for Human and Machine Cognition, Pensacola, FL, USA, where he works on autonomous underwater systems. His research interests include bimanual underwater manipulation and underwater human–machine teaming.

MONIKA ROZNERE received the Ph.D. degree in computer science from Dartmouth College, Hanover, NH, USA, in 2024.

She is currently an Assistant Professor with the School of Computing, Binghamton University, Binghamton, NY, USA, and leads the Marine Robotics Laboratory. Her research interests include active perception, sensor fusion, and autonomous exploration.



ALBERTO QUATTRINI LI (Member, IEEE) received the Ph.D. degree in computer science and engineering from Polytechnic of Milan, Milan, Italy, in 2015.

He is currently an Associate Professor with the Computer Science Department, Dartmouth College, Hanover, NH, USA, and the Co-Director of the Reality and Robotics Laboratory. His research interests include autonomous mobile robotics, artificial intelligence, and agents and multiagent systems.