LaRS: A Diverse Panoptic Maritime Obstacle Detection Dataset and Benchmark

Lojze Žust, Janez Perš, Matej Kristan
University of Ljubljana
{lojze.zust,matej.kristan}@fri.uni-lj.si, janez.pers@fe.uni-lj.si

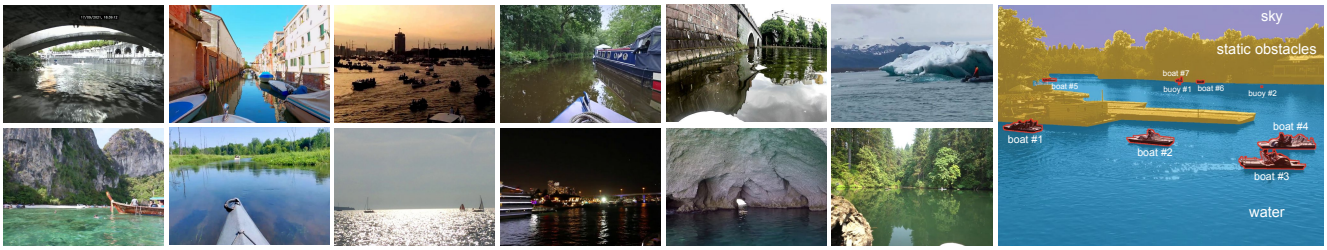

Figure 1: LaRS features diverse and challenging USV-centric scenes with per-pixel panoptic annotations (right).

Abstract

The progress in maritime obstacle detection is hindered by the lack of a diverse dataset that adequately captures the complexity of general maritime environments. We present the first maritime panoptic obstacle detection benchmark LaRS, featuring scenes from Lakes, Rivers and Seas. Our major contribution is the new dataset, which boasts the largest diversity in recording locations, scene types, obstacle classes, and acquisition conditions among the related datasets. LaRS is composed of over 4000 per-pixel labeled key frames with nine preceding frames to allow utilization of the temporal texture, amounting to over 40k frames. Each key frame is annotated with 8 thing, 3 stuff classes and 19 global scene attributes. We report the results of 27 semantic and panoptic segmentation methods, along with several performance insights and future research directions. To enable objective evaluation, we have implemented an online evaluation server. The LaRS dataset, evaluation toolkit and benchmark are publicly available at: https://lojzezust.github.io/lars-dataset

1. Introduction

The maritime industry is undergoing a fundamental transformation. With over 90% of goods being moved over water, substantial efforts are being invested in development of autonomous unmanned surface vessels (USV) [19, 15]. These autonomous boats serve a wide range of purposes, ranging from automated inspection, environmental monitoring, waste cleanup, cargo shipping, to civilian transportation. The autonomy of USVs critically depends on obstacle detection capability for timely collision avoidance. Similarly to the automotive domain [18, 21], cameras have been extensively explored for this task [26, 7, 36, 3, 35, 2].

There are several challenges associated with maritime obstacle detection. The appearance of the navigable surface (water) is dynamic and reflects the environment, often containing strong mirroring and sun glitter (Figure 1). Although modern detectors [45, 46, 8] can accurately detect common dynamic obstacles such as ships and boats, the appearance of obstacles such as buoys, people and animals can vary significantly, bringing the task closer to anomaly detection [30, 9]. Furthermore, background static obstacles, such as shorelines and piers, cannot be addressed by these methods.

The currently dominant approach [26, 2] instead employs semantic segmentation to decompose the scene into three semantic classes (water, obstacles and sky), which jointly address static and dynamic obstacles. Nevertheless, the recent detection benchmark [5] indicates that segmentation methods could benefit from the detection approach. A natural approach that combines these two principles is panoptic segmentation [24], which has proven highly effective in the related field of autonomous ground vehicles [18, 21, 14, 55]. Unfortunately, panoptic segmentation has not been fully explored for maritime perception, primarily due to the lack of a diverse, publicly available, curated panoptic dataset.

Several maritime evaluation [37, 3, 5] and training [4, 15] datasets have been proposed, as shown in Table 1. However, a common drawback of the major evaluation datasets

is that the dynamic obstacles are annotated only with bounding boxes, limiting the evaluation capability. Additionally, the current segmentation training datasets [4, 15] are modest in size and diversity, and the only reported RGB-based maritime panoptic dataset [38] is private and cannot be utilized by the community. Moreover, the scene diversity in individual datasets is fairly low, since they are all captured in limited geographic locations, which hampers the development of robust maritime obstacle detection methods capable of handling general maritime environments.

We address the aforementioned drawbacks by proposing the first maritime panoptic obstacle detection benchmark. Our major contribution is the Lakes Rivers and Seas (LaRS) dataset (see Figure 1). LaRS surpasses existing datasets in terms of diversity, obstacle types and acquisition conditions. The dataset is composed of over 4000 key frames with panoptic labels for 3 stuff and 8 thing categories, and 19 global scene attributes. Each key frame is equipped with the preceding nine frames to facilitate the development of methods that exploit temporal texture. To ensure equal attribute distribution, the training, validation, and test splits were carefully constructed, and we have implemented an online evaluation server to mitigate test-set overfitting.

In addition to the LaRS dataset, our second contribution is the analysis of 19 recent semantic segmentation networks and 8 panoptic segmentation networks. We highlight several limitations of these methods and identify opportunities for their improvement. The dataset, benchmark, and evaluation toolkit will be publicly released, to enable the research community to utilize and build upon our work.

2. Related Work

Maritime obstacle detection. The early works in camera-based obstacle detection include statistical semantic segmentation methods [26], handcrafted saliency estimation [6], background subtraction [36] and stereo reconstruction [48, 33]. These methods, however, typically fail on mirroring, glitter and other visual ambiguities. The general-purpose CNN-based object detectors [39, 32, 5] have shown a much better resilience, but do not cope well with long-tail distribution object types and cannot address background static obstacles.

The current dominant line of research stems from the early statistical method [27], which proposed segmenting the scene into navigable and non-navigable regions (i.e., water and obstacles), thus jointly addressing dynamic and static obstacles. Several works [7, 4] have shown that semantic segmentation networks from the AGV domain underperform in the maritime setup and a number of maritime-specific segmentation networks have been proposed since, most notably [42, 2, 12, 51]. A recent work [57] proposed exploiting the temporal texture to address reflections, while several works considered alternative visual modalities such as thermal imaging [40, 34]. [38] reported some success of a maritime panoptic ship and buoy detection network on a private RGB dataset. Recently the Maritime Computer Vision (MaCVi) initiative has been introduced [22] with the goal of uniting the community and moving the field towards common goals. Notably, it features USV-based obstacle detection and segmentation challenges with several teams contributing approaches surpassing the previous state-of-the-art.

Dataset	Frames	T	Env.	Ann.	Classes St.	Th.	Im.
MODD [26]	4454	-	S	B	1	2	-
MODD2 [3]	11,675	-	S	B	1	2	-
SMD [37]	16,000	-	S	B	1	1	-
MODS [5]	8175	9	S	B	1	3	-
FloW-Img [16]	2000	-	L,R	B	-	1	-
Waterline [42]	400	-	L,R	S	2	-	-
Tamp-WS [44]	600	-	L,R	S	2	-	-
USVI-WS [15]	700	-	L,R	S	2	-	-
ROSEBUD [28]	549	-	R	S	7	-	-
MaSTr1325 [4]	1325	-	S	S	4	-	-
MaSTr1478 [57]	1478	5	L,R,S	S	4	-	-
MarPS-1395 [38]	1395	-	S	P	3	3	-
LaRS	4006	9	L,R,S	P	3	8	19

Table 1: Comparison of RGB-based maritime obstacle detection datasets in the number of annotated frames (Frames) and temporal context frames (T), environment types (Env.), number of stuff (St.), thing (Th.) and image-level (Im.) classes. Grayed out datasets are not publicly available. *Environments*: L - lake, R - river, S - sea. *Obstacle labels*: B - bounding box, S - semantic seg., P - panoptic seg.

Maritime datasets. The existing RGB maritime obstacle detection datasets are summarized in Table 1. Several datasets annotate only dynamic obstacles using bounding boxes and often focus on a specific class of objects such as ships (SMD [37]) or floating waste (FloW-IMG [16]). MODD [26] and MODD2 [3] feature more diverse dynamic obstacles annotated by bounding boxes and annotate the static obstacles by lines separating them from the water. A recent evaluation-only dataset MODS [5] surpasses its predecessors in the number of annotated obstacles and proposes an evaluation protocol for both object detection- and segmentation-based maritime methods. The evaluation emphasizes performance aspects important for USV navigation. Two maritime datasets have been recently released in the robotics domain [19, 1], but are not annotated for obstacle detection.

Several segmentation-oriented datasets have been proposed. A training dataset MaSTr1325 [4] is captured in a maritime environment and annotated with per-pixel labels

for water, obstacle and sky. Several smaller datasets following the same annotation protocol (Waterline [42], TampereWaterSeg [44] and USVInland-WS [15]) were captured on inland waters, where reflections are more commonly present due to calmer waters. ROSEBUD [28] extends the number of segmentation classes, but is among the smallest datasets. Recently MaSTr1478 [57] temporally extended [4] with preceding frames and included additional 153 images from inland environments featuring scenes with strong reflections. This is currently the largest maritime segmentation training dataset for obstacle detection. Only two panoptic maritime obstacle detection datasets have been published: MarPS-1395 [38] and MassMIND [34]. However, MarPS-1395 is not publicly available and MassMIND addresses thermal imaging only.

In short, existing public maritime datasets either lack annotations for panoptic obstacle detection or are too small for training and testing modern deep learning methods. Furthermore, they lack scene diversity since they are recorded at a single geographic location. The LaRS dataset, which we present next, overcomes these limitations and fills the gap to enable the development of the next generation of maritime obstacle detection methods.

3. LaRS: Lakes Rivers and Seas dataset

A wide range of sources was considered to ensure the visual diversity of LaRS.Specifically, we (i) collected scenes from public online videos featuring various activities captured from boats around the world, (ii) recorded new sequences in a number of different geographic locations ourselves and (iii) included the most challenging scenes from existing maritime datasets.

The collection of public videos was guided using search prompts related to underrepresented scenes in the existing datasets. This includes canals (*e.g.* "canal tour"), exotic locations (*e.g.* "tropic boat tour", "polar kayaking"), crowded scenes (*e.g.* "boat parade"), strong reflections (*e.g.* "still lake"), and poor visibility conditions (*e.g.* "boat ride in the rain", "night-time boat ride"). At least one key frame was extracted from each of the collected 396 sequences, to ensure visual diversity. In addition, a state-of-the-art obstacle segmentation network [2] on the collected sequences to identify additional difficult key frames. Namely, we manually inspected the predicted segmentation and included examples with failures such as false negative obstacle segmentation and false positives on reflections to increase the difficulty level. In this way, a set of 897 representative key frames spanning diverse and challenging scenes was selected.

Next, we manually recorded videos at various locations on lakes, rivers and seas. From these, we identified 494 challenging sequences, and using the same process as for online videos, we identified 1354 diverse and challenging key frames.

Source	Sequences	Key Frames
In-house	494	1354 (33.8 %)
Web videos	396	897 (22.4 %)
MaSTr1325 [4]	-	1323 (33.0 %)
USV Inland [15]	29	211 (5.3 %)
MIT Sea Grant [19]	35	122 (3.0 %)
SMD [37]	32	99 (2.5 %)

Table 2: LaRS data sources with number of the sourced sequences, the number of selected frames and their percentage in the final dataset.

We reviewed sequences from existing maritime datasets spanning different tasks [37, 15, 19] and selected 96 of the most challenging sequences – of these, 432 key frames were selected. We also included 1323 frames from the major USV-oriented segmentation training dataset [4]. The collection process thus yielded a set of 4006 key frames. The contributions of individual data sources to the final set are summarized in Table 2.

Following [57], to facilitate future development of detection methods that might exploit the temporal texture, we equipped all 4k key frames with the preceding 9 frames.The total number of images in LaRS is thus over 40k. Faces were de-identified in all frames by running a face detector and blurring, followed by manual inspection.

Dataset annotation. All 4k selected key frames were manually annotated with per-pixel panoptic labels by a professional labeling company. In particular, *water*, *sky* and *static obstacles* like shores and piers were annotated as stuff classes, while the dynamic obstacles instances were segmented and classified into 8 different object categories (see Figure 3): *boat*, *row boat*, *paddle board*, *buoy*, *swimmer*, *animal*, *float* and an open-world *other* class to cover the remaining obstacles. Following a standard practice [29] *group labels* were used to group multiple hard-to-delineate neighbouring instances of the same category. Regions that could not be reliably manually segmented were labeled with the *ignore* class. Global attributes were assigned to key frames, to indicate *environment type*, *illumination conditions*, *presence of reflections*, *surface roughness* and *scene conditions*. Examples of scenes corresponding to the 19 global attribute labels are shown in Figure 2.

Annotation correctness was further analyzed to ensure the highest quality of the dataset. In the first pass, state-of-the-art semantic segmentation and panoptic segmentatation methods were trained and run on the entire dataset to identify major annotation errors. Visual inspection of large FP and FN predictions revealed annotation errors in 210 images, which were manually corrected. Finally, we manually

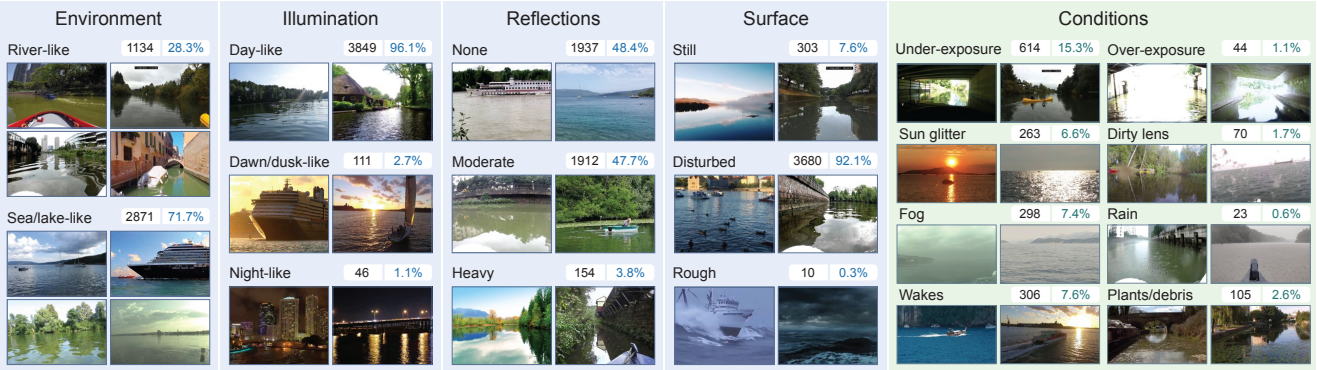

Figure 2: LaRS frames are labeled with 19 global attributes relevant for navigation. Mutually exclusive and mutually non-exclusive groups are indicated in blue and green, respectively. The numbers indicate the amount of frames in the dataset.

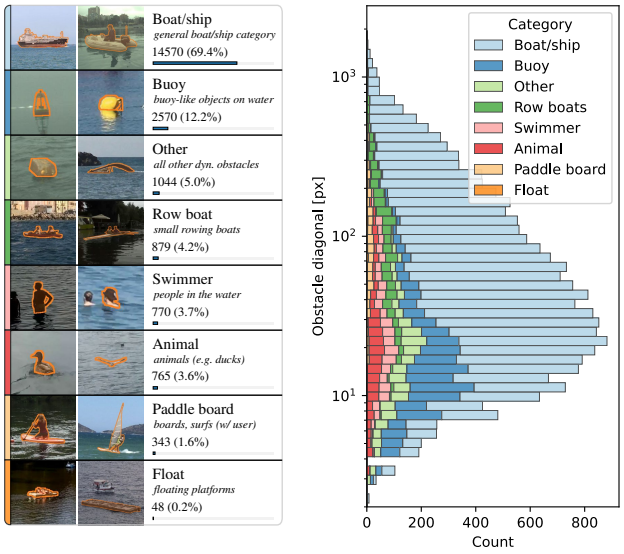

Figure 3: Statistics of dynamic obstacle classes in LaRS (left) with respect to their size (right).

inspected all ground truth instance labels of the dynamic obstacles and identified and corrected approximately 3600 annotation errors. The statistics of the final dynamic obstacle categories their instance distribution by size are shown in Figure 3.

Dataset evaluation splits. The dataset was split into training (65 %), validation (5 %) and test (30 %) sets. To prevent overfitting, we made sure there was no overlap between the sets, i.e., that all key frames extracted from a single sequence are contained within the same set. We also ensured that the distribution of the resolution, reflection levels and scene types is similar across the dataset splits. This was done by computing histograms over the aforementioned properties within each set and computing the Hellinger distances between all three pairs of image sets. A randomized search was then applied to create splits that minimized the average Hellinger distance. The training and test splits will be publicly released along with the ground truth. For the test set, only the frames will be released, while the ground truth is withheld and an evaluation server has been set-up to provide automated and unbiased evaluation.

4. Evaluation protocol

The methods are trained on the training set, the validation set is used for stopping criterion and the performance is evaluated on the test set. The evaluation protocol includes two tasks: (i) the classical semantic-segmentation-based obstacle detection and (ii) panoptic-segmentation-based obstacle detection. The respective performance measures are described next.

4.1. Semantic segmentation performance measures

The standard maritime obstacle detection evaluation protocol MODS [5] is applied to analyze the methods based on semantic segmentation. This protocol considers three semantic classes: water, sky and obstacle. The first two are directly obtained from the ground truth panoptic labels, while the last is obtained by combining all dynamic and static obstacle annotations. In addition to MODS domain-specific primary measures, we also compute the mean intersection-over-union (mIoU), a commonly used measure in general semantic segmentation [29, 18, 21].

The MODS primary performance measures are (i) water-edge estimation accuracy computed from boundary between water and static obstacles and (ii) dynamic obstacle detection accuracy. The ground truth panoptic labels simplify the water-edge estimation accuracy measure, which we define as per-pixel classification accuracy evaluated within a d pixels thick region around the ground-truth water

edge, G_d, i.e.,

$$\mu = \frac{1}{|G_d|} \sum_{(p,g)\in G_d} [p = g], \tag{1}$$

where p and g are predicted and ground-truth labels of pixels in G_d.

The MODS dynamic obstacle detection accuracy is determined by precision (Pr), recall (Re) and F1 score calculated in correspondence to the practical use of the methods. The method iterates over all ground truth dynamic obstacles. If the coverage of the predicted *obstacle* pixels exceeds $\theta = 0.7$, the dynamic obstacle is counted as a true-positive, otherwise it counts as a false-negative. The number of false-positives is estimated as the number of predicted obstacle segments (computed by connected components) in the ground-truth water mask. Please see [5] for further details.

4.2. Panoptic segmentation performance measures

Standard panoptic performance evaluation measures [24] are used: segmentation quality (SQ), recognition quality (RQ) and the combined panoptic quality (PQ):

$$\mathrm{PQ} = \underbrace{\frac{\sum_{(p,g)\in \mathrm{TP}} \mathrm{IoU}(p,g)}{|\mathrm{TP}|}}_{\text{segmentation quality (SQ)}} \times \underbrace{\frac{|\mathrm{TP}|}{|\mathrm{TP}| + \frac{1}{2}|\mathrm{FP}| + \frac{1}{2}|\mathrm{FN}|}}_{\text{recognition quality (RQ)}}. \tag{2}$$

The individual metrics are also reported separately for *thing* and *stuff* classes indicated by superscripts $(\cdot)^{\mathrm{Th}}$ and $(\cdot)^{\mathrm{St}}$.

It should be noted that, from the perspective of obstacle detection, additional instance detections on static obstacles are not considered false positives. Additionally, misclassification of an obstacle type is considered less critical than failing to detect the obstacle altogether. Therefore, we also report obstacle-class-agnostic variants of the metrics, which ignore the class label, denoted by $(\cdot)^{\mathrm{Th}_a}$.

5. Experimental results

5.1. Semantic segmentation methods

We considered 19 methods. Three single-frame state-of-the-art maritime-specific obstacle detection methods (WaSR [2], WODIS [12], IntCatchAI [42]) and several general semantic segmentation methods, i.e., four FNC-style classical methods (FCN [31], UNet [41], DeepLabv3 [10], DeepLabv3+ [11], PointRend [25], KNet [56]), three modern lightweight convolutional methods (BiSeNetv1 [53], BiSeNetv2 [52], STDC [20]) and two transformer-based methods (SegFormer [50], Segmenter [43]). The selection also includes two recent temporal semantic segmentation methods from the AGV domain (CSANet [54], TMANet [47]) and one from maritime domain (WaSR-T [57]).

Architecture	Bbone	μ	Pr	Re	F1	mIoU	FPS	GMacs
UNet [41]	S5	75.7	8.6	70.6	15.4	90.1	5.2	1621
FCN [31]	RN-50	76.8	50.1	68.7	57.9	92.6	5.2	1582
FCN [31]	RN-101	77.4	59.0	68.5	63.4	95.0	3.4	2203
DeepLabv3 [10]	RN-101	77.5	61.1	72.0	66.1	95.2	2.4	2779
DeepLabv3+ [11]	RN-101	77.8	57.8	71.7	64.0	95.4	3.3	2031
PointRend [25]	RN-101	77.5	60.6	71.1	65.4	94.9	8.7	521
BiSeNetv1 [53]	RN-50	73.3	31.6	66.3	42.8	92.2	10.1	792
BiSeNetv2 [52]	-	73.9	48.2	63.2	54.7	93.5	51.1	98.4
STDC1 [20]	-	75.6	58.6	65.3	61.8	93.6	**72.9**	67.7
STDC2 [20]	-	76.5	64.3	64.3	64.3	94.5	56.5	94.1
SegFormer [50]	MiT-B2	78.6	63.8	77.5	70.0	96.8	5.6	144
Segmenter [43]	ViT-B	72.2	51.6	59.5	55.2	95.1	2.6	556
KNet [56]	Swin-T	**78.8**	**67.6**	**80.4**	**73.4**	**97.2**	4.2	1973
WaSR [2]	RN-101	71.0	59.9	63.4	61.6	96.6	16.5	399
WODIS [12]	RN-101	63.0	38.8	61.1	47.5	85.7	35.4	61.8
IntCatchAI [42]	-	62.4	40.6	50.2	44.9	45.6	6.7	**4.7**
WaSR-T [57]	RN-101	71.1	59.7	64.7	62.1	96.7	11.1	579
CSANet [54]	RN-101	63.7	47.2	58.2	52.1	94.2	2.1	3912
TMANet [47]	RN-50	77.1	52.5	73.4	61.2	94.1	0.8	5193

Table 3: Performance of single-image state-of-the-art general (top), maritime (middle) and temporal (bottom) semantic segmentation methods on LaRS. Gold, silver and bronze indicate the top three scores in each category.

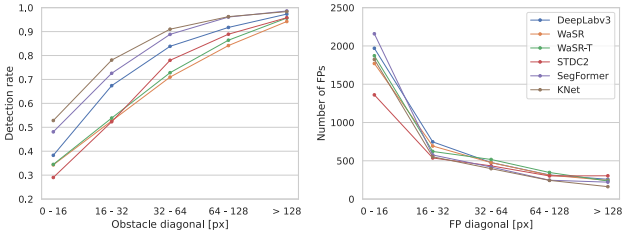

Figure 4: Segmentation-based obstacle detection rate (left) and number of false positives (right) w.r.t. the obstacle size.

WaSR [2], WaSR-T [57], CSANet [54], TMANet [47], WODIS [12] and IntCatchAI [42] were trained using their official configurations. All other methods were trained using *MMSegmentation* [17] with their Cityscapes configurations adapted to LaRS. The methods were trained on 2 x NVIDIA V100 GPUs with a batch size of 8. Runtimes were estimated in frames per second (FPS) on a single GPU.

The results are reported in Table 3. KNet [56] achieves the best water-edge accuracy (78.8 %), followed by SegFormer [50] (-0.2%), which implies a very good segmentation accuracy. This is supported by mIoU, which ranks these two methods at the top. More importantly, these two methods also outperform all other methods in F1 score by a large margin, indicating very good dynamic obstacle detection performance. Specifically, KNet ranks first, followed by SegFormer (-3.4% F1 score) and DeepLabv3 [10] (-7.3% F1 score).

Note that the best-performing methods are relatively

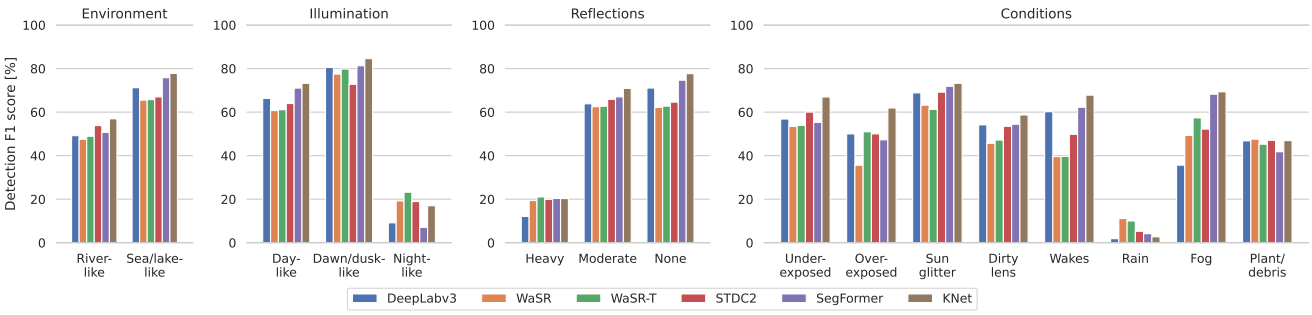

Figure 5: Semantic segmentation detection performance (F1) with respect to global attributes.

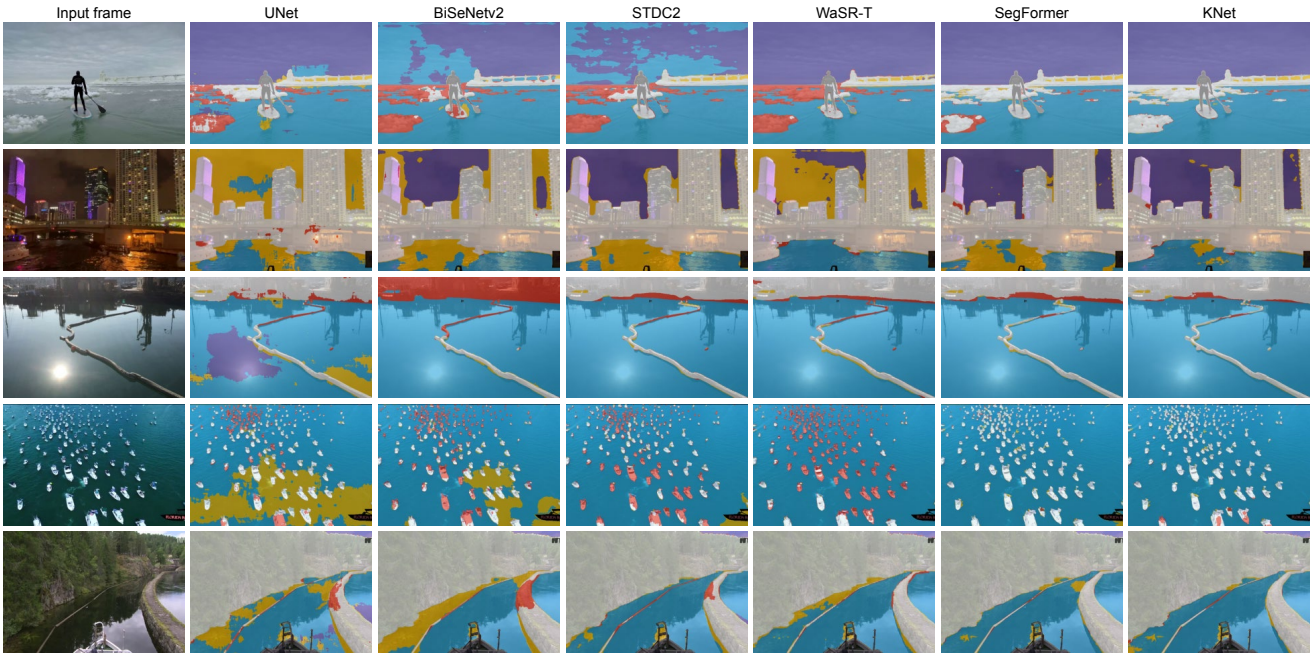

Figure 6: Qualitative semantic segmentation results on LaRS. Sky and water classes are shown in purple and blue, respectively. TP, FN and FP obstacle predictions are shown in white, red and yellow, respectively, while black indicates the ignore region.

slow (~ 4-5 FPS) even on high-end hardware and may not be suitable for real-world applications with often limited compute power. Alternatively, STDC1 and STDC2 [20] demonstrate exceptional efficiency (~ 50-70 FPS), while incurring a performance drop of 9-10% in terms of F1 score compared to the top performer KNet.

To further probe the performance of the best-performing and fastest methods, we analyze the detection rate (Re) and the number of FP detections with respect to the obstacle size in Figure 4. The largest performance variance between methods is observed for small obstacles. This is where KNet and SegFormer most substantially stand out from the rest, which is also confirmed by qualitative examples in Figure 6, particularly on thin (third row) and compact small obstacles (fourth row).

Interestingly, compared to single-frame methods, the temporal methods do not appear to benefit from the additional temporal context. For example, the performance of temporal WaSR-T [57] is almost on par (+0.5% F1) with its single-frame counterpart WaSR [2]. Since the prior work [57] on a smaller training set indicated a clear advantage of WaSR-T over WaSR, we speculate that the observed reduced difference is due to the increased size and larger diversity of the LaRS training set.

Figure 5 investigates performance with respect to scene attributes. Overall, river-like environments are more challenging compared to sea/lake-like environments, which may be attributed to a larger quantity of reflections and

Architecture	Backbone	PQ (%)				RQ (%)				SQ (%)				FPS	GMacs
		All	Th	Th_a	St	All	Th	Th_a	St	All	Th	Th_a	St		
Panoptic Deeplab [13]	ResNet-50	34.7	13.4	33.0	91.4	40.3	19.3	**46.3**	96.2	69.5	60.0	71.3	94.9	6.0	**339.3**
Panoptic FPN [23]	ResNet-50	40.1	21.7	**35.5**	89.3	46.9	28.6	45.9	95.8	73.5	66.1	**77.3**	93.1	**21.7**	471.4
Panoptic FPN [23]	ResNet-101	38.7	19.7	**35.5**	89.4	45.0	26.1	46.0	95.5	73.6	66.1	77.1	93.5	16.7	627.2
MaX-DeepLab [49]	MaX-S	31.9	9.5	19.2	91.7	36.1	13.4	26.0	96.6	71.3	62.5	73.7	94.8	3.7	-
Mask2Former [14]	ResNet-50	37.6	17.0	27.9	92.4	43.7	23.6	37.6	97.3	71.3	62.4	74.2	95.0	10.6	464.2
Mask2Former [14]	ResNet-101	37.2	16.3	29.2	92.8	43.0	22.7	38.9	97.1	71.4	62.3	75.0	95.5	5.7	620.0
Mask2Former [14]	Swin-T	39.2	18.8	34.0	93.7	45.5	25.8	45.2	98.1	72.2	63.5	75.2	95.4	5.4	470.7
Mask2Former [14]	Swin-B	**41.7**	**21.8**	33.6	**94.7**	**48.5**	**29.7**	44.6	**98.5**	**78.2**	**71.5**	75.3	**96.2**	4.8	948.0

Table 4: Panoptic quality (PQ), recognition quality (RQ) and segmentation quality (SQ) reported overall (All) and with respect to stuff (St) and things (Th), with Th_a denoting class-agnostic score. The inference speed is reported in FPS.

Architecture	Backbone	μ	F1	mIoU
Panoptic Deeplab [13]	ResNet-50	73.5	64.6	95.4
Panoptic FPN [23]	ResNet-50	67.2	58.9	93.5
Panoptic FPN [23]	ResNet-101	66.9	58.1	93.3
MaX-DeepLab [49]	MaX-S	72.7	60.2	95.4
Mask2Former [14]	ResNet-50	75.1	54.9	95.4
Mask2Former [14]	ResNet-101	75.8	53.2	95.6
Mask2Former [14]	Swin-T	76.2	56.7	96.8
Mask2Former [14]	Swin-B	**77.4**	**71.1**	**97.6**

Table 5: Performance of panoptic methods under the semantic segmentation setup.

background variety of the former. The methods are fairly robust to dusk scenes, with a moderate performance increase compared to daytime scenes. However, the performance substantially drops on night-time scenes. Interestingly, a performance advantage of the temporal WaSR-T is observed over the single-frame counterparts, which indicates the potential for exploiting temporal context in situations with significant visual ambiguity. Moreover, all methods are fairly robust to moderate reflections, while strong reflections lead to substantial performance drops. Of the different scene conditions, the methods perform best on sun glitter, fog, and wakes, while the worst performance is observed in the presence of rain, dirty lenses, and plants/debris.

5.2. Panoptic segmentation methods

Several panoptic methods with various backbones are considered: Panoptic Deeplab [13] and Panoptic FPN [23] as members of conv-net family with strong baseline performance on ground-vehicle-related tasks, and two state-of-the-art representatives of transformer-based mask classification methods MaX-Deeplab [49] and Mask2Former [14]. The methods were trained on 2 x NVIDIA V100 GPUs with a batch size of 4.

Results in Table 4 indicate that the top PQ performance is achieved by Swin-B-based Mask2Former [14] (41.7 %), followed by Panoptic FPN [23] (-1.6 %) and Swin-T-based Mask2Former (-2.5 %). Overall, the methods achieve relatively low PQ scores. Comparing PQ^{Th} and PQ^{St}, we observe that the static obstacles (i.e., stuff class) are well detected ($PQ^{St} = 94.7\%$ for the best method) but methods struggle the detection of dynamic obstacles (i.e., things).

Specifically, the recognition quality for dynamic obstacles of the best method is only $RQ^{Th} = 27.7\%$. High RQ^{Th} requires accurate detection obstacles as well as correct classification. Ignoring the classification errors (RQ^{Th_a}) substantially increases this score (to 44.6 %), which confirms that a major source of errors is obstacle misclassification. We thus plot the confusion matrix between predicted and GT instance classes for the top performing method (Swin-B-based Mask2Former [14]) in Figure 8 and observe significant confusion between boat/ship, row boats, paddle board and float categories. The objects from the rarer classes are often predicted as the more common boat/ship category. In addition, similarly to what we observed in semantic segmentation methods, small obstacles such as buoys, swimmers and animals are often missed and segmented as water. It should be noted that modern panoptic methods use a *void* label for regions without sufficiently confident segment predictions. Void labels account for approximately 24% of all predictions on dynamic obstacles.

Another source of errors is the grouping of objects into a single detection and the decomposition of a single instance into several detections. The qualitative examples in Figure 7 show that incorrect object grouping/splitting is particularly acute in dense scenes (row 3). Interestingly, the best-performing method Mask2Former sometimes incorrectly groups even well-separated instances (rows 2 and 3).

Note that labeling several small water regions as static obstacles substantially affects robotic navigation in practice, since the USV might frequently stop to avoid a possible

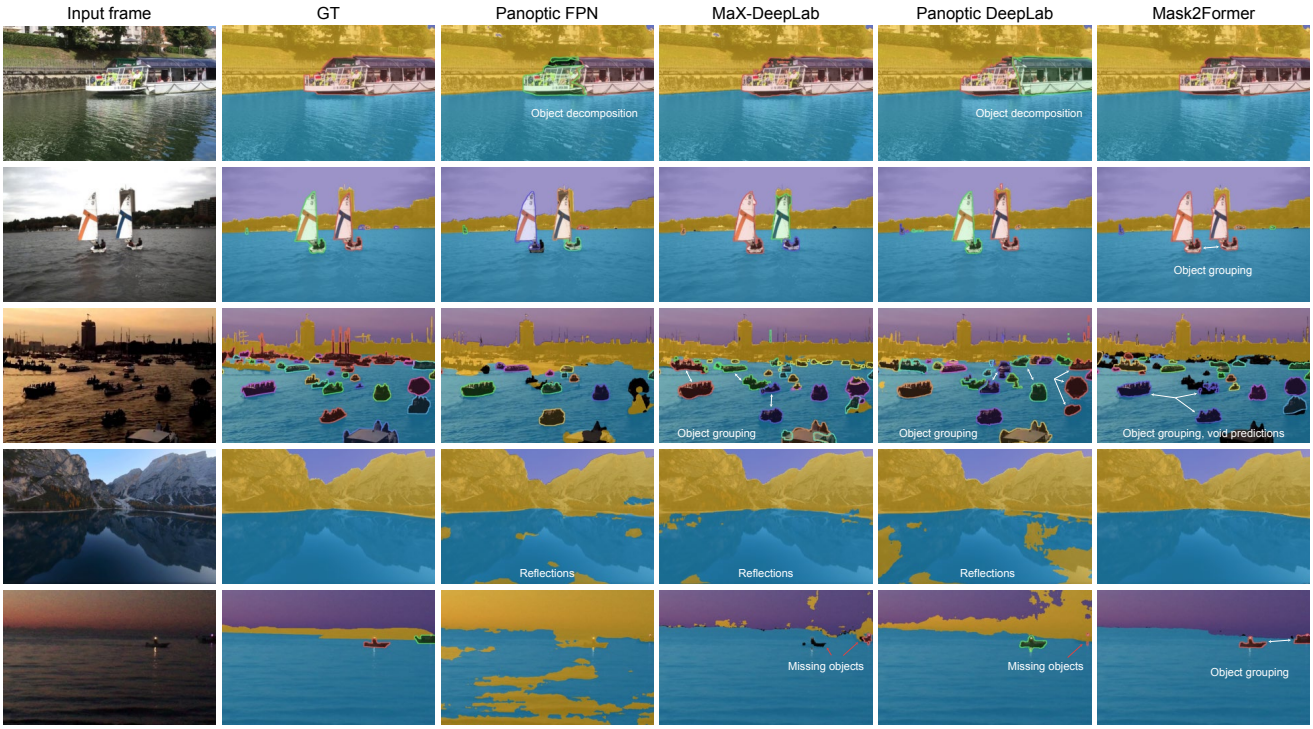

Figure 7: Qualitative panoptic segmentation results. Individual instance detections are outlined with different colors. Void predictions are colored black. Common errors are indicated with white text.

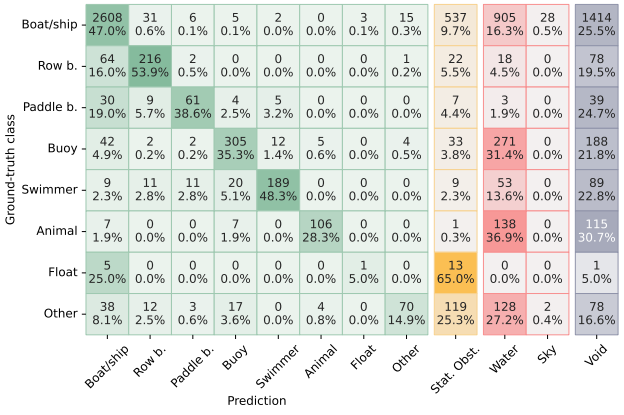

Figure 8: Confusion matrix of ground-truth dynamic obstacles for Mask2Former with the Swin-B backbone.

collision. This is not properly reflected in panoptic performance measures, which would decrease only slightly. Similarly, joining two nearby obstacles into a single instance is not detrimental from a practical obstacle avoidance standpoint, but can significantly reduce the panoptic measures.

We thus also evaluate the methods with semantic segmentation measures from Section 4.1, by assigning all detected static and dynamic obstacles and void predictions to the *obstacle* class. Results in Table 5 reveal that the best panoptic methods perform on par with state-of-the-art semantic segmentation methods under this setup. For example, the best panoptic method (Mask2Former with Swin-B backbone) lags behind the best semantic segmentation method (Table 3) by only -2.3 % in F1 score. This presents a clear opportunity for panoptic methods, whose performance would greatly improve also at the panoptic level by properly addressing the instance detection and separation capability.

5.3. Difficulty level of LaRS

We conduct an experiment to demonstrate the difficulty level of the LaRS benchmark. We train the best performing semantic segmentation method KNet on some of the largest and most diverse existing maritime segmentation datasets MaSTr1325 [4], MaSTr1478 [57] and ROSEBUD [28] and evaluate them on the LaRS test set. Results are presented in Table 6.

We observe a severe performance drop when training on previously available datasets. These datasets are limited in nature and lack the variety required to tackle the LaRS benchmark. For example MaSTr1325 only contains maritime scenes, while ROSEBUD only contains fluvial scenes. Furthermore, even combining all the examples from related datasets for training the network is not enough and leads to subpar performance compared to LaRS training (F1 drop

Train dataset	μ	Pr	Re	F1	mIoU
MaSTr1325	62.2	28.2	69.9	40.2	87.5
MaSTr1478	72.5	52.1	67.0	58.6	93.6
ROSEBUD	64.5	30.1	57.2	39.5	81.6
MaSTr1478 + ROSEBUD	72.2	55.0	67.5	60.6	92.9
LaRS	**78.8**	**67.6**	**80.4**	**73.4**	**97.2**

Table 6: Performance of KNet semantic segmentation on the LaRS test set, when trained with different existing maritime segmentation datasets.

of 12.8 %). This suggest that the current datasets are just not representative enough for general maritime perception and outlines the need for large, diverse datasets like LaRS to move the field forward.

6. Conclusion

We presented the first maritime panoptic obstacle detection benchmark LaRS, containing scenes from lakes, rivers, and seas. LaRS is the largest dataset of its kind and exceeds other maritime obstacle detection datasets in terms of the diversity of recording locations, acquisition conditions, obstacle appearances, number of categories and annotation detail. Each key frame is annotated by panoptic segmentation labels, 19 global attributes and additionally equipped with several preceding frames to enable the development of methods exploiting temporal context.

Results for 27 semantic- and panoptic-segmentation-based detection methods reveal that semantic-segmentation methods slightly outperform the panoptic counterparts in overall segmentation quality. We identify several opportunities for improvement of the methods, notably improving the instance separation of panoptic methods and better exploitation of the temporal context in scenes with significant ambiguity. The dataset, tookit and the online evaluation server will be publicly released to foster further advancements in maritime obstacle detection.

Acknowledgments

This work was supported by the Slovenian Research Agency programs P2-0214 and P2-0095, and project J2-2506.

References

[1] Ola Benderius, Christian Berger, and Krister Blanch. Are we ready for beyond-application high-volume data? The Reeds robot perception benchmark dataset, Sept. 2021. 2

[2] Borja Bovcon and Matej Kristan. WaSR–A Water Segmentation and Refinement Maritime Obstacle Detection Network. *IEEE Transactions on Cybernetics*, pages 1–14, July 2021. 1, 2, 3, 5, 6

[3] Borja Bovcon, Rok Mandeljc, Janez Perš, and Matej Kristan. Stereo obstacle detection for unmanned surface vehicles by IMU-assisted semantic segmentation. *Robotics and Autonomous Systems*, 104, 2018. 1, 2

[4] Borja Bovcon, Jon Muhovič, Janez Perš, and Matej Kristan. The MaSTr1325 dataset for training deep USV obstacle detection models. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3431–3438, 2019. 1, 2, 3, 8

[5] Borja Bovcon, Jon Muhovič, Duško Vranac, Dean Mozetič, Janez Perš, and Matej Kristan. MODS – A USV-oriented object detection and obstacle segmentation benchmark. *IEEE Transactions on Intelligent Transportation Systems*, May 2021. 1, 2, 4, 5

[6] Tom Cane and James Ferryman. Saliency-Based Detection for Maritime Object Tracking. In *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1257–1264, June 2016. 2

[7] Tom Cane and James Ferryman. Evaluating deep semantic segmentation networks for object detection in maritime surveillance. In *Proceedings of AVSS 2018 - 2018 15th IEEE International Conference on Advanced Video and Signal-Based Surveillance*, 2019. 1, 2

[8] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-End Object Detection with Transformers. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12346 LNCS:213–229, May 2020. 1

[9] Robin Chan, Krzysztof Lis, Svenja Uhlemeyer, Hermann Blum, Sina Honari, Roland Siegwart, Mathieu Salzmann, Pascal Fua, and Matthias Rottmann. SegmentMeIfYouCan: A Benchmark for Anomaly Segmentation. Apr. 2021. 1

[10] Liang Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, June 2017. 5

[11] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 801–818, Feb. 2018. 5

[12] Xiang Chen, Yuanchang Liu, and Kamalasudhan Achuthan. WODIS: Water Obstacle Detection Network Based on Image Segmentation for Autonomous Surface Vehicles in Maritime Environments. *IEEE Transactions on Instrumentation and Measurement*, 70:1–13, 2021. 2, 5

[13] Bowen Cheng, Maxwell D Collins, Yukun Zhu, Ting Liu, Thomas S Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-DeepLab: A Simple, Strong, and Fast Baseline for Bottom-Up Panoptic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12475–12485, June 2020. 7

[14] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention Mask Transformer for Universal Image Segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1280–1289, 2022. 1, 7

[15] Yuwei Cheng, Mengxin Jiang, Jiannan Zhu, and Yimin Liu. Are We Ready for Unmanned Surface Vehicles in Inland Waterways? The USVInland Multisensor Dataset and Benchmark. *IEEE Robotics and Automation Letters*, 6(2):3964–3970, 2021. 1, 2, 3

[16] Yuwei Cheng, Jiannan Zhu, Mengxin Jiang, Jie Fu, Changsong Pang, Peidong Wang, Kris Sankaran, Olawale Onabola, Yimin Liu, Dianbo Liu, and Yoshua Bengio. FloW: A Dataset and Benchmark for Floating Waste Detection in Inland Waters. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10953–10962, 2021. 2

[17] Mmsegmentation Contributors. MMSegmentation: OpenMMLab semantic segmentation toolbox and benchmark, 2020. 5

[18] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1, 4

[19] Michael DeFilippo, Michael Sacarny, and Paul Robinette. RoboWhaler: A Robotic Vessel for Marine Autonomy and Dataset Collection. In *OCEANS 2021: San Diego – Porto*, pages 1–7, Sept. 2021. 1, 2, 3

[20] Mingyuan Fan, Shenqi Lai, Junshi Huang, Xiaoming Wei, Zhenhua Chai, Junfeng Luo, and Xiaolin Wei. Rethinking BiSeNet For Real-time Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9716–9725, Apr. 2021. 5, 6

[21] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, June 2012. 1, 4

[22] Benjamin Kiefer, Matej Kristan, Janez Perš, Lojze Žust, Fabio Poiesi, Fabio Andrade, Alexandre Bernardino, Matthew Dawkins, Jenni Raitoharju, Yitong Quan, Adem Atmaca, Timon Höfer, Qiming Zhang, Yufei Xu, Jing Zhang, Dacheng Tao, Lars Sommer, Raphael Spraul, Hangyue Zhao, Hongpu Zhang, Yanyun Zhao, Jan Lukas Augustin, Eui-ik Jeon, Impyeong Lee, Luca Zedda, Andrea Loddo, Cecilia Di Ruberto, Sagar Verma, Siddharth Gupta, Shishir Muralidhara, Niharika Hegde, Daitao Xing, Nikolaos Evangeliou, Anthony Tzes, Vojtěch Bartl, Jakub Špaňhel, Adam Herout, Neelanjan Bhowmik, Toby P. Breckon, Shivanand Kundargi, Tejas Anvekar, Ramesh Ashok Tabib, Uma Mudenagudi, Arpita Vats, Yang Song, Delong Liu, Yonglin Li, Shuman Li, Chenhao Tan, Long Lan, Vladimir Somers, Christophe De Vleeschouwer, Alexandre Alahi, Hsiang-Wei Huang, Cheng-Yen Yang, Jenq-Neng Hwang, Pyong-Kun Kim, Kwangju Kim, Kyoungoh Lee, Shuai Jiang, Haiwen Li, Zheng Ziqiang, Tuan-Anh Vu, Hai Nguyen-Truong, Sai-Kit Yeung, Zhuang Jia, Sophia Yang, Chih-Chung Hsu, Xiu-Yu Hou, Yu-An Jhang, Simon Yang, and Mau-Tsuen Yang. 1st Workshop on Maritime Computer Vision (MaCVi) 2023: Challenge Results. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 265–302, 2023. 2

[23] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic Feature Pyramid Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6399–6408, Apr. 2019. 7

[24] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollar. Panoptic segmentation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2019-June, pages 9396–9405. IEEE Computer Society, June 2019. 1, 5

[25] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. PointRend: Image Segmentation As Rendering. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9796–9805, June 2020. 5

[26] Matej Kristan, Vildana Sulić Kenk, Stanislav Kovačič, and Janez Perš. Fast Image-Based Obstacle Detection from Unmanned Surface Vehicles. *IEEE Transactions on Cybernetics*, 46(3), 2016. 1, 2

[27] Matej Kristan, Janez Perš, Vildana Sulič, and Stanislav Kovačič. A Graphical Model for Rapid Obstacle Image-Map Estimation from Unmanned Surface Vehicles. In Daniel Cremers, Ian Reid, Hideo Saito, and Ming-Hsuan Yang, editors, *Computer Vision – ACCV 2014*, Lecture Notes in Computer Science, pages 391–406, Cham, 2015. Springer International Publishing. 2

[28] Reeve Lambert, Jalil Chavez-Galaviz, Jianwen Li, and Nina Mahmoudian. ROSEBUD: A Deep Fluvial Segmentation Dataset for Monocular Vision-Based River Navigation and Obstacle Avoidance. *Sensors*, 22(13):4681, June 2022. 2, 3, 8

[29] Tsung Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 8693 LNCS, pages 740–755. Springer Verlag, May 2014. 3, 4

[30] Krzysztof Lis, Krishna Kanth Nakka, Pascal Fua, and Mathieu Salzmann. Detecting the Unexpected via Image Resynthesis. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2152–2161, Seoul, Korea (South), Oct. 2019. IEEE. 1

[31] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 07-12-June, pages 431–440, 2015. 5

[32] Liyong Ma, Wei Xie, and Haibin Huang. Convolutional neural network based obstacle detection for unmanned surface vehicle. *Mathematical Biosciences and Engineering*, 17(1), 2020. 2

[33] Jon Muhovič, Rok Mandeljc, Borja Bovcon, Matej Kristan, and Janez Perš. Obstacle Tracking for Unmanned Surface Vessels Using 3-D Point Cloud. *IEEE Journal of Oceanic Engineering*, 45(3), 2020. 2

[34] Shailesh Nirgudkar, Michael DeFilippo, Michael Sacarny, Michael Benjamin, and Paul Robinette. MassMIND: Massachusetts Maritime INfrared Dataset, Sept. 2022. 2, 3

[35] Shailesh Nirgudkar and Paul Robinette. Beyond Visible Light: Usage of Long Wave Infrared for Object Detection in Maritime Environment. In *2021 20th International Conference on Advanced Robotics (ICAR)*, pages 1093–1100, Dec. 2021. 1

[36] Dilip K. Prasad, Chandrashekar Krishna Prasath, Deepu Rajan, Lily Rachmawati, Eshan Rajabally, and Chai Quek. Object Detection in a Maritime Environment: Performance Evaluation of Background Subtraction Methods. *IEEE Transactions on Intelligent Transportation Systems*, 20(5):1787–1802, May 2019. 1, 2

[37] Dilip K. Prasad, Deepu Rajan, Lily Rachmawati, Eshan Rajabally, and Chai Quek. Video Processing From Electro-Optical Sensors for Object Detection and Tracking in a Maritime Environment: A Survey. *IEEE Transactions on Intelligent Transportation Systems*, 18(8), 2017. 1, 2, 3

[38] Dalei Qiao, Guangzhong Liu, Wei Li, Taizhi Lyu, and Juan Zhang. Automated Full Scene Parsing for Marine ASVs Using Monocular Vision. *Journal of Intelligent & Robotic Systems*, 104(2):1–20, 2022. 2, 3

[39] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, June 2017. 2

[40] Paul Robinette, Michael Sacarny, Michael Defilippo, Michael Novitzky, and Michael R. Benjamin. Sensor Evaluation for Autonomous Surface Vehicles in Inland Waterways. In *OCEANS 2019 - Marseille, OCEANS Marseille 2019*, volume 2019-June. Institute of Electrical and Electronics Engineers Inc., June 2019. 2

[41] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, volume 9351, pages 234–241, 2015. 5

[42] L. Steccanella, D. D. Bloisi, A. Castellini, and A. Farinelli. Waterline and obstacle detection in images from low-cost autonomous boats for environmental monitoring. *Robotics and Autonomous Systems*, 124, 2020. 2, 3, 5

[43] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for Semantic Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7262–7272, May 2021. 5

[44] Jussi Taipalmaa, Nikolaos Passalis, Honglei Zhang, Moncef Gabbouj, and Jenni Raitoharju. High-Resolution Water Segmentation for Autonomous Unmanned Surface Vehicles: A Novel Dataset and Evaluation. In *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, Pittsburgh, PA, USA, Oct. 2019. IEEE. 2, 3

[45] Mingxing Tan, Ruoming Pang, and Quoc V. Le. EfficientDet: Scalable and Efficient Object Detection. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 10778–10787, Nov. 2019. 1

[46] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: Fully convolutional one-stage object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2019-Octob, 2019. 1

[47] Hao Wang, Weining Wang, and Jing Liu. Temporal Memory Attention for Video Semantic Segmentation. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 2254–2258. IEEE, Sept. 2021. 5

[48] Han Wang and Zhuo Wei. Stereovision based obstacle detection system for unmanned surface vehicle. In *2013 IEEE International Conference on Robotics and Biomimetics, ROBIO 2013*, 2013. 2

[49] Huiyu Wang, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. MaX-DeepLab: End-to-End Panoptic Segmentation with Mask Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5463–5474, Dec. 2020. 7

[50] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. In *Advances in Neural Information Processing Systems*, volume 34, pages 12077–12090, May 2021. 5

[51] L Yao, D Kanoulas, Z Ji, and Y Liu. ShorelineNet: An Efficient Deep Learning Approach for Shoreline Semantic Segmentation for Unmanned Surface Vehicles. In *Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021. 2

[52] Changqian Yu, Changxin Gao, Jingbo Wang, Gang Yu, Chunhua Shen, and Nong Sang. BiSeNet V2: Bilateral Network with Guided Aggregation for Real-Time Semantic Segmentation. *International Journal of Computer Vision*, 129(11):3051–3068, Nov. 2021. 5

[53] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. BiSeNet: Bilateral segmentation network for real-time semantic segmentation. In *Computer Vision - ECCV 2018*, pages 334–349, Aug. 2018. 5

[54] Yichen Yuan, Lijun Wang, and Yifan Wang. CSANet for Video Semantic Segmentation With Inter-Frame Mutual Learning. *IEEE Signal Processing Letters*, 28:1675–1679, 2021. 5

[55] Oliver Zendel, Matthias Schörghuber, Bernhard Rainer, Markus Murschitz, and Csaba Beleznai. Unifying Panoptic Segmentation for Autonomous Driving. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21319–21328, June 2022. 1

[56] Wenwei Zhang, Jiangmiao Pang, Kai Chen, and Chen Change Loy. K-Net: Towards Unified Image Segmentation. In *Advances in Neural Information Processing Systems*, Oct. 2021. 5

[57] Lojze Žust and Matej Kristan. Temporal Context for Robust Maritime Obstacle Detection. In *2022 IEEE/RJS International Conference on Intelligent Robots and Systems (IROS)*, 2022. 2, 3, 5, 6, 8