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 & 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 in 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 unifying 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.

### 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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Frames</th>
<th>T</th>
<th>Env.</th>
<th>Ann.</th>
<th>St.</th>
<th>Th.</th>
<th>Im.</th>
</tr>
</thead>
<tbody>
<tr>
<td>MODD [26]</td>
<td>4454</td>
<td>-</td>
<td>S</td>
<td>B</td>
<td>1</td>
<td>2</td>
<td>-</td>
</tr>
<tr>
<td>MODD2 [3]</td>
<td>11,675</td>
<td>-</td>
<td>S</td>
<td>B</td>
<td>1</td>
<td>2</td>
<td>-</td>
</tr>
<tr>
<td>SMD [37]</td>
<td>16,000</td>
<td>-</td>
<td>S</td>
<td>B</td>
<td>1</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>MODS [5]</td>
<td>8175</td>
<td>9</td>
<td>S</td>
<td>B</td>
<td>1</td>
<td>3</td>
<td>-</td>
</tr>
<tr>
<td>FloW-Img [16]</td>
<td>2000</td>
<td>-</td>
<td>B</td>
<td>S</td>
<td>1</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>Waterline [42]</td>
<td>400</td>
<td>-</td>
<td>B</td>
<td>1, 3, 6</td>
<td>2</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Tamp-WS [44]</td>
<td>600</td>
<td>-</td>
<td>B</td>
<td>1, 3, 6</td>
<td>2</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>USVI-WS [15]</td>
<td>700</td>
<td>-</td>
<td>B</td>
<td>1, 3, 6</td>
<td>2</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ROSEBUD [38]</td>
<td>549</td>
<td>-</td>
<td>B</td>
<td>1, 3, 6</td>
<td>7</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MaStTr1325 [4]</td>
<td>1325</td>
<td>-</td>
<td>B</td>
<td>1, 3, 6</td>
<td>4</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MaSTr1478 [57]</td>
<td>1478</td>
<td>5</td>
<td>B, 1, 3, 6</td>
<td>4</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>MarPS-1395 [38]</td>
<td>1395</td>
<td>-</td>
<td>B</td>
<td>3</td>
<td>3</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>LaRS</td>
<td>4006</td>
<td>9</td>
<td>B, 1, 3, 6</td>
<td>3</td>
<td>8</td>
<td>19</td>
<td></td>
</tr>
</tbody>
</table>

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: B - lake, 1 - river, 3 - sea. Obstacle labels: B - bounding box, S - semantic seg., P - panoptic seg.
for water, obstacle and sky. Several smaller datasets following the same annotation protocol (Waterline [42], Tampere-WaterSeg [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 with failures such as false negative obstacle segmentation and failures on reflections 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.

We reviewed sequences from existing maritime datasets spanning different tasks [37, 15, 19] and selected 96 of the most challenging scenes – 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 segmentation 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

<table>
<thead>
<tr>
<th>Source</th>
<th>Sequences</th>
<th>Key Frames</th>
</tr>
</thead>
<tbody>
<tr>
<td>In-house</td>
<td>494</td>
<td>1354 (33.8 %)</td>
</tr>
<tr>
<td>Web videos</td>
<td>396</td>
<td>897 (22.4 %)</td>
</tr>
<tr>
<td>MaStr1325 [4]</td>
<td>-</td>
<td>1323 (33.0 %)</td>
</tr>
<tr>
<td>USV Inland [15]</td>
<td>29</td>
<td>211 (5.3 %)</td>
</tr>
<tr>
<td>MIT Sea Grant [19]</td>
<td>35</td>
<td>122 (3.0 %)</td>
</tr>
<tr>
<td>SMD [37]</td>
<td>32</td>
<td>99 (2.5 %)</td>
</tr>
</tbody>
</table>

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

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

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):

\[
PQ = \frac{\sum_{(p,g) \in TP} \text{IoU}(p, g)}{|\text{TP}|} \times \frac{|\text{TP}|}{|\text{TP}| + \frac{1}{2}|\text{FP}| + \frac{1}{2}|\text{FN}|},
\]

(2)

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

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 \((\cdot)^{\text{Th}_0}\).

5. Experimental results

5.1. Semantic segmentation methods

We considered 19 methods. Three single-frame state-of-the-art maritime-specific 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]).

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.](image-url)

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
The methods were trained on 2 x NVIDIA V100 GPUs with cation methods MaX-Deeplab [49] and Mask2Former [14].

the-art representatives of transformer-based mask classification methods MaX-Deeplab [49] and Mask2Former [14]. Two state-of-the-art representatives of convolution net family with strong baseline performance include 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.

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_{St} and PQ_{Th}, 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 PQ_{Th} = 27.7%. High PQ_{Th} requires accurate detection obstacles as well as correct classification. Ignoring the classification errors (RQ_{Th}) 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 obstacle.
We thus also evaluate the methods with 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
of 12.8%). This suggests 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, toolkit 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


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

<table>
<thead>
<tr>
<th>Train dataset</th>
<th>µ</th>
<th>Pr</th>
<th>Re</th>
<th>F1</th>
<th>mIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>MaSTr1325</td>
<td>62.2</td>
<td>28.2</td>
<td>69.9</td>
<td>40.2</td>
<td>87.5</td>
</tr>
<tr>
<td>MaSTr1478</td>
<td>72.5</td>
<td>52.1</td>
<td>67.0</td>
<td>58.6</td>
<td>93.6</td>
</tr>
<tr>
<td>ROSEBUD</td>
<td>64.5</td>
<td>30.1</td>
<td>57.2</td>
<td>39.5</td>
<td>81.6</td>
</tr>
<tr>
<td>MaSTr1478 + ROSEBUD</td>
<td>72.2</td>
<td>55.0</td>
<td>67.5</td>
<td>60.6</td>
<td>92.9</td>
</tr>
<tr>
<td>LaRS</td>
<td>78.8</td>
<td>67.6</td>
<td>80.4</td>
<td>73.4</td>
<td>97.2</td>
</tr>
</tbody>
</table>


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

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


Benjamin Kiefer, Matej Kristan, Janez Perš, Lojze Žust, Fabio Poiesi, Fabio Andrade, Alexandre Bernardino, Matthew Dawkins, Jenni Raitoharju, Yitong Quan, Adem Fabio Poiesi, Fabio Andrade, Alexandre Bernardino, Pattern Recognition (CVPR), pages 9796–9805, June 2020. 5

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


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

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


