LaRS: A Diverse Panoptic Maritime Obstacle Detection Dataset and Benchmark

Supplementary material

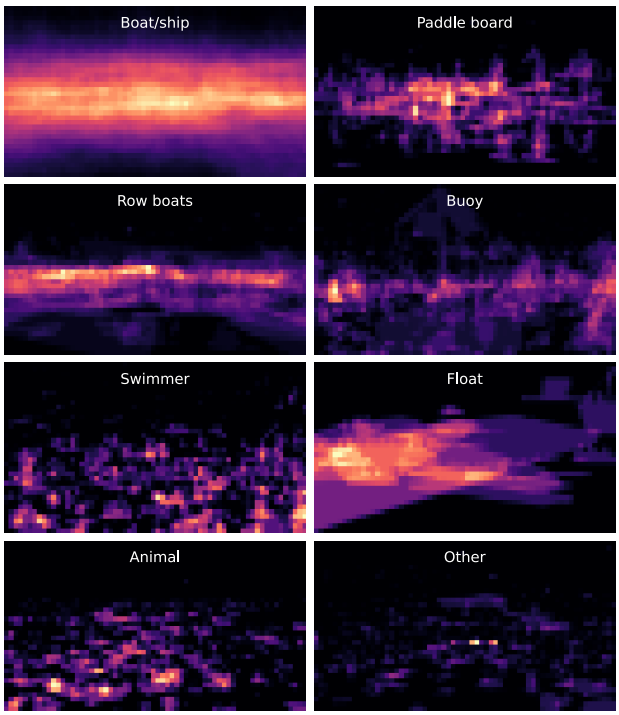

Figure A1: Heatmaps of obstacle categories. Dark and bright colors denote areas with low and high frequency of obstacles, respectively.

A. Obstacle heatmaps

To visualize the spatial distribution of obstacles of different categories, we compute heatmaps of obstacle positions. We divide the image into 64×36 spatial bins. In each bin we count the number of obstacle segments which intersect with the bin. The heatmaps in Figure A1 are normalized with respect to the most populated bin in each category.

Most categories are evenly distributed across the image. Various vessel categories (*e.g.* boat/ship, row boats) vertically most commonly appear near the center of the image, which coincides with the usual position of the horizon. Some smaller categories such as swimmer, animal and buoy contain many instances closer to the camera as well.

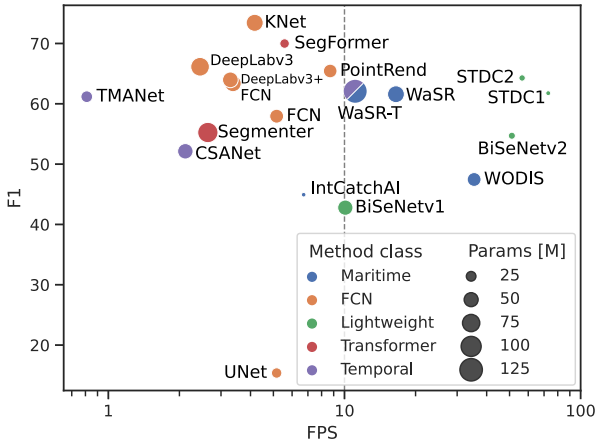

Figure B1: The performance of semantic segmentation method (F1) with respect to their efficiency, measured in FPS. Dashed line denotes the real-time boundary of 10 FPS.

B. Additional semantic segmentation results

B.1. Method detection efficiency

To better understand the trade-offs between detection performance and speed we plot the obstacle detection F1 score of methods with respect to their inference time measured in FPS in Figure B1. Most methods do not reach the real-time inference speed requirement of 10 FPS including top-performing KNet [12] and SegFormer [10]. WaSR-T [13] and WaSR [1] both perform on the limit of this requirement and achieve an F1 score of over 60. Among the real-time methods, the STDC [5] family offers the best trade-off between speed and performance by a large margin achieving best results in both.

B.2. Additional qualitative results

Figure B2 showcases additional qualitative results for semantic segmentation methods, including low-visibility and night scenes (rows 1, 2 and 7), foggy and rainy scenes (rows 3 and 4), small obstacles (row 5) and reflections (row 6). Methods like UNet [8], BiSeNetv2 [11] and SegFormer [10] are prone to obstacle hallucinations in high-ambiguity scenes (rows 2, 6 and 7). WaSR-T [13] and KNet [12] are the most robust to these ambiguities. Small obstacles (row 5) are only picked up by SegFormer [10] and KNet [12]. Interestingly, even the best performing methods

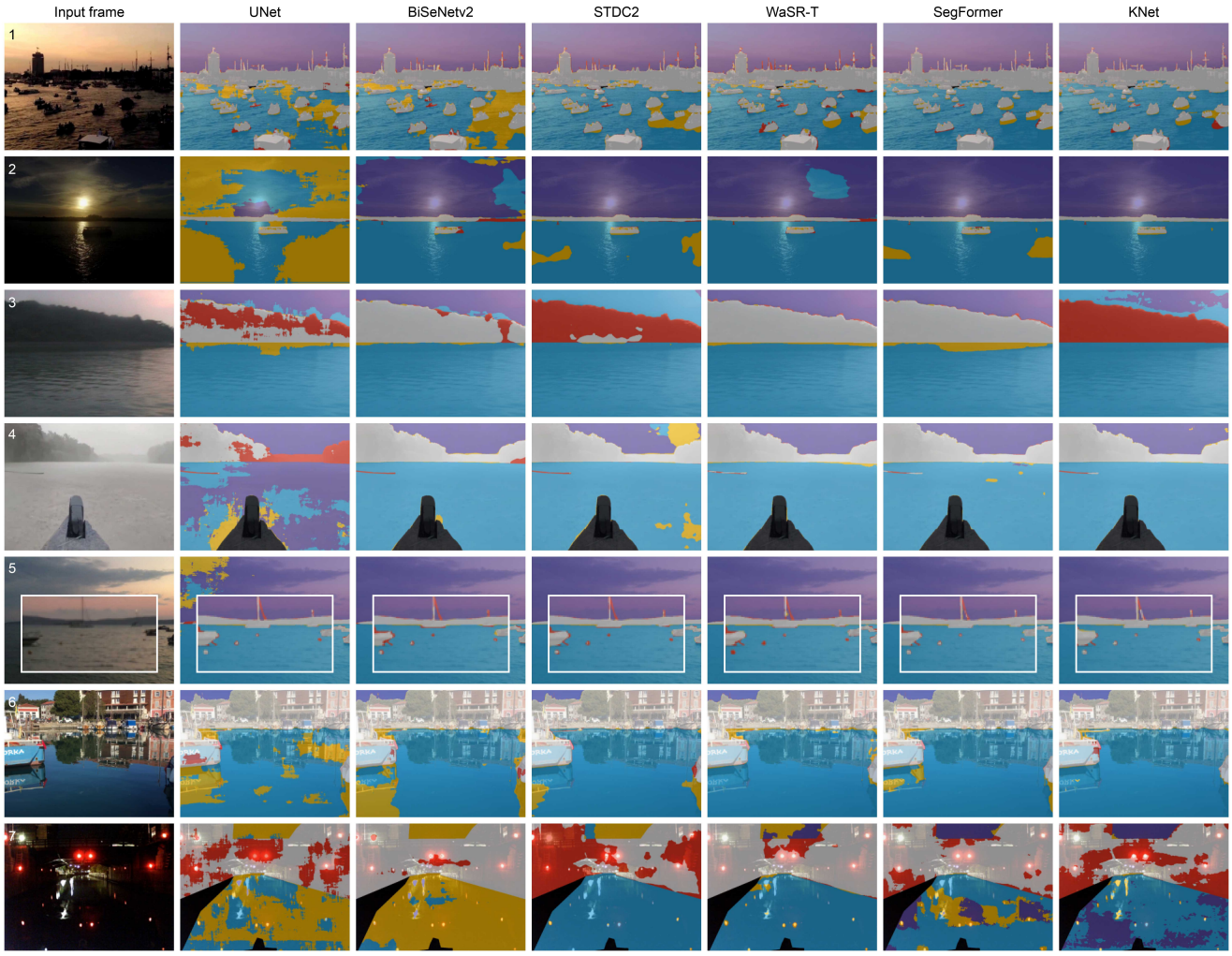

Figure B2: Additional qualitative semantic segmentation results. Sky and water classes are shown in purple and blue, respectively. TP, FN and FP obstacle predictions are shown in white, red and yellow, respectively, while black indicates ignore region. White rectangles show zoomed-in parts of the image.

(*e.g.* KNet [12]) sometimes fail in seemingly simple situations such as the scene in row 3, where the slight ambient fog leads to complete misclassification of the large landmass as water in several methods.

C. Additional panoptic segmentation results

C.1. Performance by obstacle size

Figure C1 shows the performance of panoptic methods (PQ) with respect to the obstacle size. Similarly to semantic segmentation methods, the best performance is observed on large obstacles. However, on very small obstacles, the performance drops to almost zero. We believe there is a large potential for improving panoptic methods in this regard. Note also, that the PQ metric is much more sensitive to minor mask shifts on small obstacles compared to large ones, which also impacts these results. In contrast to semantic segmentation methods, large false-positive detections are more common than small ones, with the exception of Panoptic DeepLab, which also produces a large number

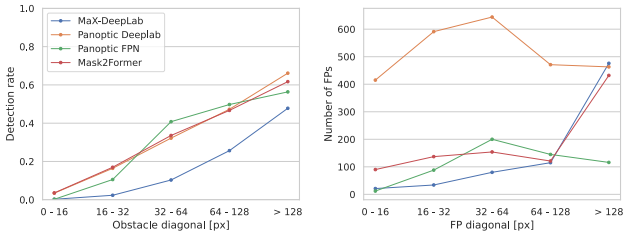

Figure C1: Performance (PQ) of panoptic methods by obstacle size.

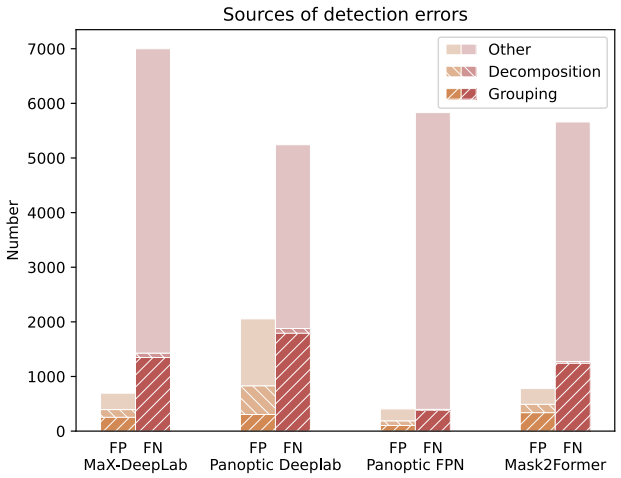

Figure C2: Common sources of FP and FN errors for panoptic detection methods.

of smaller false-positive detections.

C.2. Source of detection errors

To further explore the problem of object grouping and object decomposition errors discussed in Section 5.2, we investigate the frequency of these problems across the different methods. Specifically, we inspect the source of false-positive and false-negative detections in the obstacle-class-agnostic case. A FP segment is counted as result of a decomposition error, if it is largely (more than 70% of its area) contained within a single ground-truth obstacle segment. A FP segment is similarly counted as a result of a grouping error if it covers more than one ground-truth obstacle segments. A FN ground-truth segment is counted as a result of decomposition, if the combined coverage of all predicted obstacle segments exceeds the threshold of 50%, and as a result of grouping if there exists any single predicted segment, that that covers more than 50% of the FN segment.

The proportion of the two sources of errors for each method are reported in Figure C2. We observe that a sizeable amount of FP and FN detections are the result of these errors. Object grouping is an especially common source of false-negative detections. Addressing the issue of object grouping would thus lead to substantial performance improvements.

C.3. Obstacle confusion matrices

Similarly to the Figure 8 of the main paper, we plot the confusion matrices for the remaining panoptic methods (Panoptic DeepLab [2], Panoptic FPN [7] and MaX-DeepLab [9]) in Figure C3. Compared to Mask2Former [3], Panoptic DeepLab [2] does not rely on void predictions and correctly identifies more obstacle instances. However, the

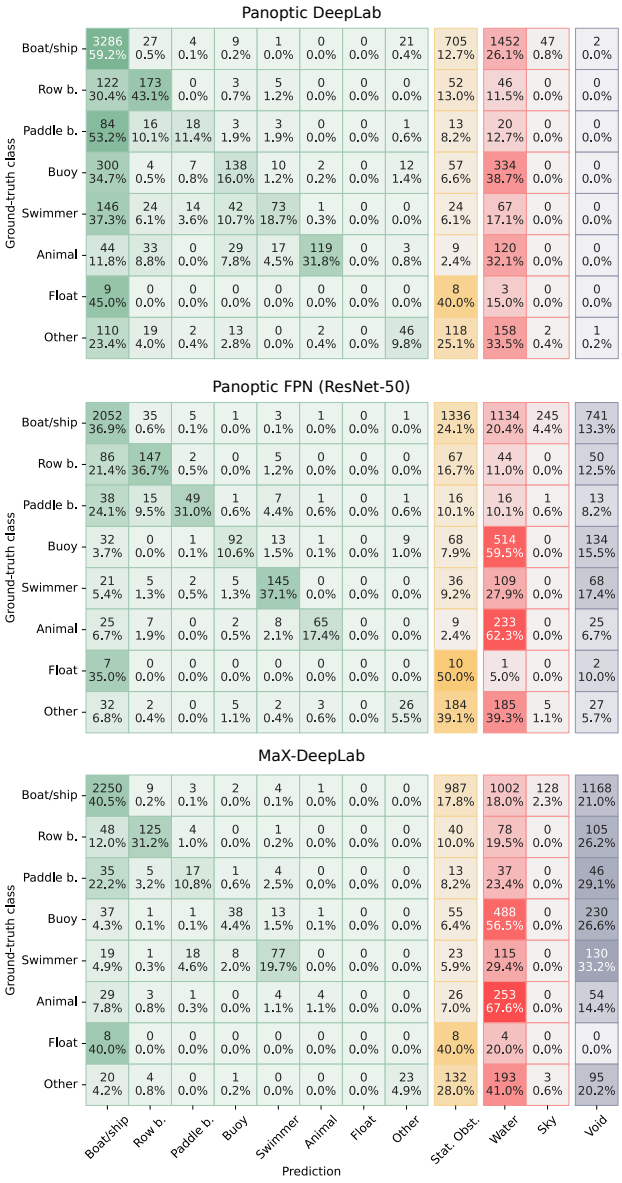

Figure C3: Ground-truth dynamic obstacle confusion matrices for panoptic methods.

confusion between individual obstacle types is much larger. For example, Panoptic DeepLab [2] tends to classify most obstacles to the majority *boat/ship* category. On the other hand, Panoptic FPN [7] and MaX-DeepLab [9] show concerning level of misclassifications of obstacles as water, which is potentially hazardous from the boat navigation perspective. This problem is especially prevalent on smaller obstacle categories such as buoys and animals.

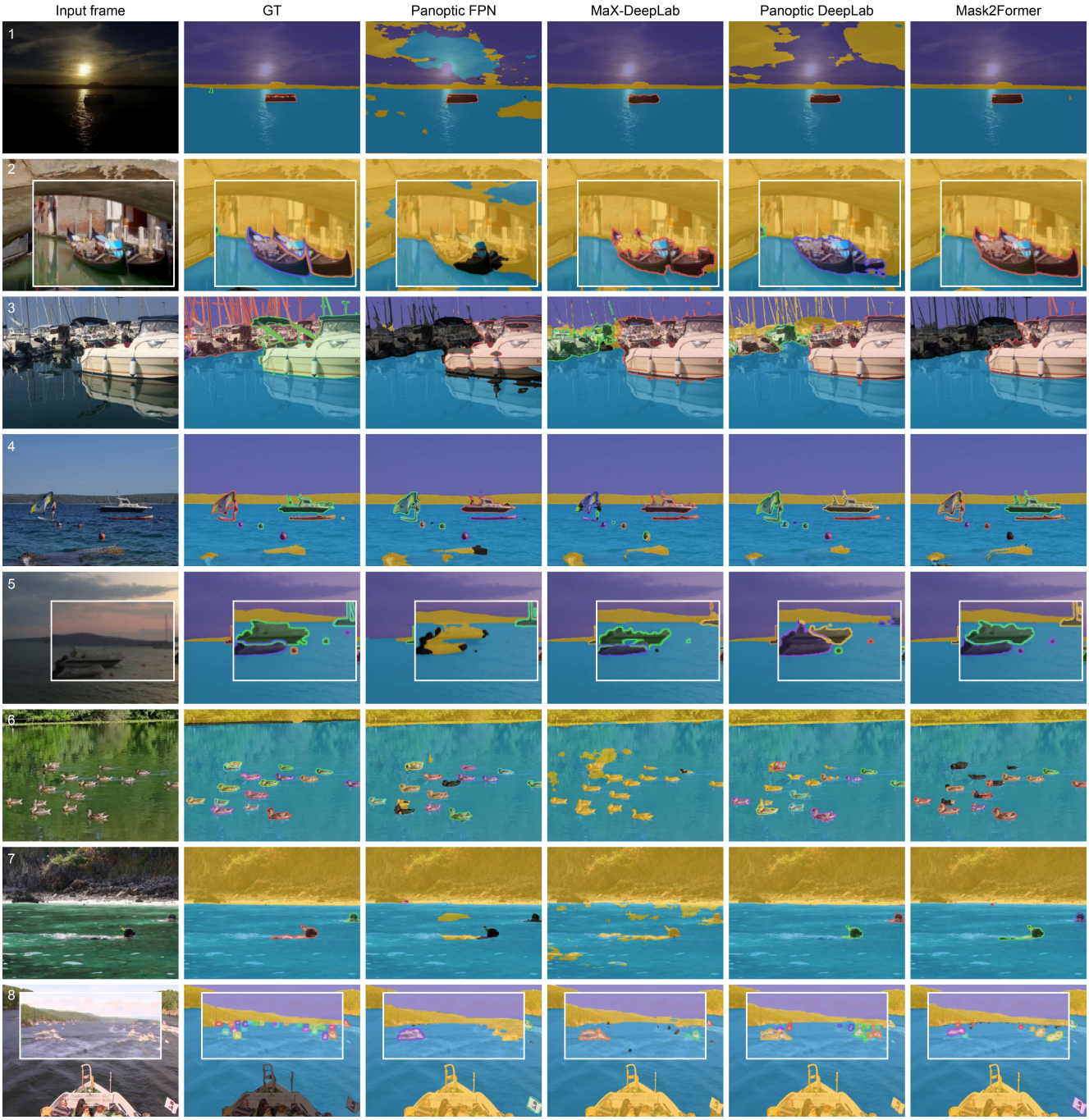

Figure C4: Additional qualitative panoptic results. Individual instance detections are outlined with different colors. Void predictions are colored black. White rectangles show zoomed-in parts of the image.

C.4. Additional qualitative results

Figure C4 showcases additional qualitative results for panoptic segmentation methods. Small or far-away objects (rows 5 and 8) are often missed (Panoptic FPN [7] and Panoptic DeepLab [2]) or labeled as void (MaX-DeepLab [9] and Mask2Former [3]). Similar objects that are close together (rows 2, 5 and 6) are commonly grouped as a single object. Additionally, Mask2Former [3] sometimes groups even far-away objects (*e.g.* buoys in row 5 and ducks in row 6).

D. Datasheet for LaRS

This document is based on *Datasheets for Datasets* by Gebru *et al.* [6]. Please see the most updated version here.

MOTIVATION

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.
LaRS was created as a benchmark for panoptic maritime obstacle detection, to facilitate the development and evaluation of new panoptic (and semantic) segmentation methods for robust obstacle detection under a wide range of conditions and situations.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?
The detaset was created by the ViCoS lab at the University of Ljubljana, Slovenia.

What support was needed to make this dataset? (e.g.who funded the creation of the dataset? If there is an associated grant, provide the name of the grantor and the grant name and number, or if it was supported by a company or government agency, give those details.)
The creation of the dataset was funded by the Slovenian Research Agency program P2-0214 and project J2-2506.

Any other comments?
No.

COMPOSITION

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.
Instances in the dataset are snippets (i.e. scenes) of 10 sequential video frames (photos) depicting maritime scenarios captured from the perspective of a USV.

How many instances are there in total (of each type, if appropriate)?
The dataset contains four thousand instances.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).
The instance were extracted from a larger set of videos. The videos were manually selected to feature diverse scenarios and geographic locations. At least one instance was extracted from each video to ensure visual diversity. Additional challenging instances were extracted through a visual inspection of predictions of a state-of-the-art (SotA) method.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.
Each snippet contains 10 image frames. The image frames were processed to blur faces to protect the identities of individuals in the image.

Is there a label or target associated with each instance? If so, please provide a description.
One *"key"* video frame in the snippet is annotated with panoptic masks. This includes "water", "sky" and "static obstacle" stuff classes and 8 different dynamic obstacle categories (*i.e.* things). The average image has ∼9 masks, totaling ∼36k masks. Each scene is also annotated with 19 different global attributes covering different environment types, reflection levels and other conditions.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.
The annotations of the test set will not be made publicly available to ensure fair comparison between methods. We host an evaluation server (macvi.org) for submitting and evaluating the results of new methods.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.
In some cases, several snippets have been extracted from a single video. We include the ID of the source sequence in the naming of the instance to make this relationship explicit.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.
We provide a recommended data split into training (65 %), validation (5 %) and test (30 %) set. Source sequences are mutually exclusive between sets. We insure equal distribution of resolution, reflection levels and scene types across sets.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.
The annotations were created by human annotators and verified by us. Nonetheless, minor inconsistencies among different human annotators are possible. Annotation errors may be reported to lars.dataset@gmail.com.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete

dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.
The dataset is self-contained.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.
No.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.
No.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.
No.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.
No.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.
No. We blur the faces of people appearing in the images to protect their identity. Issues with anonymization may be reported by email to lars.dataset@gmail.com.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.
No.

Any other comments?
No.

COLLECTION

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.
The instances in the dataset were collected from a combination of online sources (publicly available videos and datasets) and recordings from members of our lab. The corresponding panoptic masks were annotated by a professional labelling company and verified by us.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created. Finally, list when the dataset was first published.
The instances in the dataset vary in their date of capture over a range of years up to 2023. The date of the first publication of the dataset is 1 August 2023.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?
The instances were captured with a wide range of different consumer-grade and industry-grade RGB cameras.

What was the resource cost of collecting the data? (e.g. what were the required computational resources, and the associated financial costs, and energy consumption - estimate the carbon footprint.)
Since the sources of the dataset instances were pre-existing videos and videos captured during vacation time of our team members, no additional resource cost occurred during the collection process.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?
Short snippets were extracted from longer video sequences. The selected snippets were determined manually based on the visual variety of the scene and difficulty, determined by the performance of existing obstacle detection methods.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?
The panoptic masks and category labels were annotated by a professional annotation service. The annotators were compensated with an hourly wage set by the vendor.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.
We underwent an internal privacy review to evaluate and determine how to mitigate any potential risks with respect to the privacy of people appearing in the photos. Blurring faces protects the privacy of the people in the photos.

Does the dataset relate to people? If not, you may skip the remainder of the questions in this section.
No.

Any other comments?
No.

PREPROCESSING / CLEANING / LABELING

Was any preprocessing/cleaning/labeling of the data done(e.g.,discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.
We blur the faces to preserve the privacy of the individuals. No other preprocessing was done to the photos.

Was the "raw" data saved in addition to the preprocessed/-cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.
No, because we preprocess the data to preserve the privacy of individuals, we do not release raw data.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.
We used the RetinaFace model [4] (https://github.com/serengil/retinaface) to detect faces in the photos.

Any other comments?
No.

USES

Has the dataset been used for any tasks already? If so, please provide a description.
The dataset was used to train and evaluate 27 different semantic and panoptic segmentation methods.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.
No. However, we require the users of the dataset to cite it in their papers, so its use is trackable via citations.

What (other) tasks could the dataset be used for?
The dataset was intended for training and evaluation of semantic and panoptic segmentation methods. However, with minimal effort the dataset could also be used for other task such as instance segmentation and object detection.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?
We do not foresee any such impact of future uses.

Are there tasks for which the dataset should not be used? If so, please provide a description.
Full terms of use for the dataset can be found at https://lojzezust.github.io/lars-dataset.

Any other comments?
No.

DISTRIBUTION

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.
Yes, the dataset will be available to the research community.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?
The dataset is available at https://lojzezust.github.io/lars-dataset

When will the dataset be distributed?
The dataset was released online on 1 August 2023

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.
The licence agreement and terms of use can be found at https://lojzezust.github.io/lars-dataset

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.
No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.
No.

Any other comments?
No.

MAINTENANCE

Who is supporting/hosting/maintaining the dataset?
The dataset will be hosted and maintained by the ViCoS lab,

University of Ljubljana.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
Please email lars.dataset@gmail.com.

Is there an erratum? If so, please provide a link or other access point.
No.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?
The dataset may be updated in case of discovered privacy concerns and major labeling errors. In this case, the version history and changes will be made clear on the dataset website (https://lojzezust.github.io/lars-dataset).

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.
No.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.
We will use a versioning system to keep track of the changes in the annotations. Older versions of annotations will be available for download to ensure reproducibility. In case of detected privacy concerns, we will update the image data accordingly. In this case, older version of the data will not be available for download.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.
We encourage the community to expore other uses of the dataset and extend it with new types of annotations. The users creating the new annotations will be responsible for hosting and distributing their annotations.

Any other comments?
No.

References

[1] 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] 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. 3, 4

[3] 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. 3, 4

[4] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5203–5212, 2020. 7

[5] 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. 1

[6] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, Nov. 2021. 5

[7] 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. 3, 4

[8] 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. 1

[9] 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. 3, 4

[10] 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. 1

[11] 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. 1

[12] 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. 1, 2

[13] 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. 1