

# DATASHEET:

## SeaClips

This document is based on *Datasheets for Datasets* by Gebru *et al.* [1]. 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.

The dataset is created for research on maritime obstacle detection, especially under consideration of temporal context. The goal is to advance robustness and increase safety of autonomous maritime navigation.

**Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**

The dataset is created by SEA.AI GmbH.

**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 work is supported by the Austrian Research Promotion Agency (FFG) within the research project “SEA-OD” (# 921850).

**Any other comments?**

-

### COMPOSITION

**What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?**

The dataset consists of videos showing maritime scenarios.

**How many instances are there in total (of each type, if appropriate)?**

There are 74 videos and 31,606 frames in total. Each video has an average duration of 14 seconds. There are 129,198 annotated bounding boxes.

**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 74 videos are selected from footage recorded over 14

different days.

**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 video has a duration between 10 and 30 seconds, and is recorded at 30 FPS.

**Is there a label or target associated with each instance?** If so, please provide a description.

In each video, each frame is annotated with bounding boxes that mark each navigational obstacle occurring in the frame. On average, there are approximately four bounding boxes per frame.

**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.

All frames are annotated with bounding boxes and classes for each bounding box. Although the annotation was done with the highest precision, humans cannot achieve 0% error rate. Thus, we cannot guarantee that no objects were missed.

**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.

No.

**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.

A manually defined training, validation, and testing split is provided to ensure independence between splits and comparable results in future work.

**Are there any errors, sources of noise, or redundancies in the dataset?** If so, please provide a description.

Annotations are performed by industry experts and each annotation is manually verified. Nevertheless, annotation inaccuracies cannot be ruled out for sure, as humans can make mistakes.

**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.

It 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.

**Any other comments?**

-

## 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 videos were recorded by three cameras mounted on boats or on shore.

**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 videos were recorded on in total 14 different days, spanning four different months in 2024 and 2025.

**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 videos are recorded with RGB sensors in three different cameras: an e-CAM50 CUNX with a resolution of 1,280 x 960 pixels, a FLIR M364C, recorded at a resolution of 1,280 x 720 pixels, and an Axis Q8752-E with 1,920 x 1,020 pixels.

**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. See Strubell *et al.*[2] for approaches in this area.)

The resource cost is unknown.

**If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**

Videos are manually selected from the recorded footage to ensure video diversity and reduce redundancy between the videos.

**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)?**

Industry experts were involved in the data recording and annotation process and they were compensated with industry-appropriate salaries.

**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.

An internal review made sure that no personal data is within the dataset. To ensure this, the following two safety measures were established: First, as the focus of the dataset is on maritime scenarios, humans in general appear only from a distance, which makes them not identifiable. The exception to this distance-based safety measure is one recording scenario where man-over-board scenarios were simulated, in which the participants explicitly consented to the recording, and they are anonymized by blurring their faces if visible. Second, boat registration numbers and logos are blurred so that the identification of data related to any person or organization is not possible.

**Does the dataset relate to people?** If not, you may skip the remainder of the questions in this section.

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.

No.

## USES

**Has the dataset been used for any tasks already?** If so, please provide a description.

The dataset is evaluated for image-based and video object detection in maritime scenarios.

**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.

The data is released as a Gated Dataset on HuggingFace, which means that the email addresses of researchers downloading the dataset are stored.

**What (other) tasks could the dataset be used for?**

The dataset should not be used for any task other than research on maritime obstacle detection. For more information, have a look at the license file in the HuggingFace dataset.

**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?

Videos do not contain night scenarios and stormy weather. Models trained on the dataset cannot be expected to work in such conditions. The validation set misses objects of one bounding box class, which could not be included, as the independence between the splits would have been violated otherwise. Recordings are limited to two geographical locations along the European Atlantic coast, which makes the dataset biased towards these regions.

**Are there tasks for which the dataset should not be used?** If so, please provide a description.

The dataset is only allowed to be used for research on maritime obstacle detection to improve safety at sea. Other use cases are not allowed. For more information, have a look at the license file in the HuggingFace dataset, which details allowed and prohibited use cases.

**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.

The dataset is released under gated access 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 released via HuggingFace Gated Datasets, under the following URL: <https://huggingface.co/datasets/SEA-AI/SeaClips>.

**When will the dataset be distributed?**

Upon paper acceptance.

**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. Please refer to the license file.

**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?**

-

## MAINTENANCE

**Who is supporting/hosting/maintaining the dataset?**

The dataset is maintained by SEA.AI GmbH.

**How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**

Please contact us via [datasets@sea.ai](mailto:datasets@sea.ai) with the tag [SeaClips] in the subject line.

**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)?

If errors, which impede the proper usage of the dataset are found, they will be corrected. No new instances will be added, already existing instances can be deleted upon reasonable request.

**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.

If the dataset will be updated, it is because of errors. Older versions will therefore not be maintained.

**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.

If you want to contribute to the dataset, please contact [datasets@sea.ai](mailto:datasets@sea.ai) with the tag [SeaClips] in the subject line.

**Any other comments?**

No.

## REFERENCES

- [1] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Dauméé III, and Kate Crawford. Datasheets for Datasets. *arXiv:1803.09010 [cs]*, January 2020.
- [2] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and Policy Considerations for Deep Learning in NLP. *arXiv:1906.02243 [cs]*, June 2019.