AUTOASSESS

AI & robotics for safe vessel inspection

D1.2 DATA MANAGEMENT PLAN

27/03/2024

AUTOASSESS
AI & robotics for safe vessel inspection

PARTNERS

Co-funded by
the European Union

AUTOASSESS
AI & robotics for safe vessel inspection

Grant Agreement No.: 101120732
Call: HORIZON-CL4-2022-DIGITAL-EMERGING-02
Topic: HORIZON-CL4-2022-DIGITAL-EMERGING-02-07
Type of action: HORIZON-IA

D1.2 DATA MANAGEMENT PLAN

Autoassess

Work package	WP 1
Task	1.2
Type deliverable	DMP – Data Management Plan
Dissemination Level	PU – Public
Due date	31/03/2024
Submission date	27/03/2024
Deliverable lead	DTU
Version	1.0
Authors	Evangelos Boukas, Melanie Brunhofer
Reviewers	Rasmus Eckholdt Andersen
Abstract	One paragraph
Keywords	Keyword 1, Keyword 2

Co-funded by the European Union

DOCUMENT REVISION HISTORY

Version	Date	Description of change	List of contributor(s)
0.1	2024/3/26	Initial Version	Evangelos Boukas, Melanie Brunhofer
1	2024/3/27	Final Version	Evangelos Boukas, Melanie Brunhofer

DISCLAIMER

Co-funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Commission. Neither the European Union nor the granting authority can be held responsible for them.

COPYRIGHT NOTICE

The Consortium is the following:

Participant number	Participant organisation name	Short name	Country
1	DANMARKS TEKNISKE UNIVERSITET	DTU	DK
2	NORGES TEKNISK-NATURVITENSKAPELIGE UNIVERSITET NTNU	NTNU	NO
3	TECHNISCHE UNIVERSITAET MUENCHEN	TUM	DE
4	UNIVERSITEIT TWENTE	UT	NL
5	SCOUTDI AS	SDI	NO
6	COGNITE AS	CGN	NO
7	FAYARD AS	FAY	DK
8	GLAFCOS MARINE EPE	GLC	EL
9	F6S NETWORK IRELAND LIMITED	F6S	IE
10	DNV AS	DNV	NO
11	EURONAV NV	ERN	BE
12	DANAOS SHIPPING COMPANY LIMITED	DAN	CY
13	KLAVENESS SHIP MANAGEMENT AS	KLV	NO
14	UNIVERSITAT ZURICH	UZH	CH
15	FLYABILITY SA	FLY	CH
16	SENSIMA INSPECTION SARL	SEN	CH

EXECUTIVE SUMMARY

This is the first version of the Data Management Plan (DMP) for the AUTOASSESS project. The DMP is a live document that is updated throughout the life of the project. You can ask the PM of the project for the most recent version of the DMP by email to: admin@autoassess.eu

The DMP includes answers to the following questions:

- What is the purpose of collecting data in the AUTOASSESS project?
- What data will be collected in the AUTOASSESS project?
- What are the formats and metadata of the data in the AUTOASSESS project?
- How do we maximize the interoperability and re-use of the data in the AUTOASSESS project?
- What is the size of the data to be stored/shared?
- Where will the data be stored in the public/private AUTOASSESS repositories?
- For how long will the data be stored in the AUTOASSESS project?
- Who has access to the data generated by the AUTOASSESS project?
- What licenses will be used for the data in the AUTOASSESS project?
- How do we handle security for the data in the AUTOASSESS project?
- How do we handle GDPR in the AUTOASSESS project?

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

DMP	Data Management Plan
EUAS	Exploration Unmanned Autonomous System
IUAS	Inspection Unmanned Autonomous System
T-IUAS	Tethered-inspection Unmanned Autonomous System
TOF	Time of Flight
IMUs	Inertial Measurement Units
UTM	Ultrasonic Thickness Measurement
NDT	Non-destructive Testing
ROS	Robot Operating System
ORP	Open Research Platform
PIDs	Persistent Identifiers
DOI	Digital Object Identifier

1 DATA SUMMARY

The Data Management Plan (DMP) describes a systematic approach to manage the data generated throughout the AUTOASSESS project's activities, focusing primarily on real sensor readings and post processing data resulting from scientific laboratory/field experimentation. It includes methodologies and guidelines for data collection, storage, and dissemination, with a strong emphasis on long-term preservation and FAIR principles.

The data strategy for the AUTOASSESS project prioritizes the generation of new data in addition to reusing existing datasets. This decision is in accordance to the project's 7 key objectives and ensures the alignment between data collection efforts and the overall AUTOASSESS objective. While incomplete, due incorporating a subset of the AUTOASSESS sensors and postprocessing, pre-existing data will be employed, specifically coming from our past projects INSPECTRONE and REDHUS, which will be provided with a non-restrictive license by DTU and NTNU.

The AUTOASSESS project will collect a range of data types essential for: a) autonomous robot operation in confined spaces, as well as b) comprehensive evaluation of large vessels. Visual data captured by RGB cameras will provide detailed insights for the vessel surface condition and its structural integrity. Additionally, 3D laser scanners as well as multi-view images with postprocessing (eg photogrammetry) will generate 3D survey data, allowing the creation of detailed geometric models for autonomous robot operation and geometric anomaly detection. Event-based cameras will be used to allow the agile navigation of the EUAS. Moreover, ultrasonic thickness sensors will collect data using non-destructive testing, ensuring the reliability and safety of maritime assets. The AUTOASSESS datasets will be stored in various formats, including image files, video streams, point clouds, radiance fields, timeseries data, etc, to be provided as input to the AUTOASSESS software components, analytical tools and mission methodologies.

In the AUTOASSESS project, the primary goal is the development and validation of autonomous aerial inspection technologies for large vessels. Real-world data from maritime environments is carefully collected and analyzed to improve the efficiency, accuracy, and safety of vessel inspections. Each dataset is a direct contribution to project objectives, enabling autonomous exploration, comprehensive inspections, and enhanced defect detection.

The collected AUTOASSESS data originates from a multitude of sensors, onboard UASs, capturing real-world data within the dynamic and challenging maritime environments. Including data directly from these environments ensures authenticity, relevance, and reliability, increasing the performance of the AUTOASSESS autonomous solution. Moreover, the inherent provenance of the data underscores its integrity and fidelity. Additionally, data collected in the lab provide edge cases for the AUTOASSESS system to achieve generalization.

While the main purpose of the AUTOASSESS data is to train our automated robotic operations and inspection technologies, they can be useful for a broader audience. The inspection data, can be used to retrain surveyors to identify defect in images rather than in person Additionally our data can be useful as a training dataset to assist in the generalization of the feature spaces of wide spread deep learning models, specifically those that tackle industrial environments.

Table 1 provides an overview of the types of experiments planned for this project, alongside a non-exhaustive list of examples of experimental parameters that will be adjusted, and the anticipated data to be collected. The data, along with metadata necessary for reproduction, processing, and evaluation, will be stored in result files, all in digital format. Non-digital data will be converted to digital formats, such as picture or text format, wherever possible.

Type of Experiment	Parameters	Data and Metadata	Data Type	Data Format
ROS-based Experiments	ROS messages, robot status	ROS messages, sensor data, control logs	ROS Message	rosbag
Dataset Collection Field Test	Recording all data from sensors	Sensor data, environmental conditions, robot status	Numerical, Categorical	MP4, JPEG, CSV, JSON, rosbag
Teleoperated Field Tests	Control commands, robot status	Video feeds, sensor data, control logs	Numerical, Video	MP4, JPEG, CSV, rosbag
Learning to fly	Trajectories while navigating	Camera images, Pose estimation (velocity, angular velocity, attitude) acceleration, motor commands (body-rates, thrust)	Video Image, Numerical	MP4, JPEG, rosbag
Fully Autonomous Flight Tests	Predefined flight paths, robot status	Video feed, sensor data, flight logs	Numerical, Video	MP4, CSV, rosbag
Corrosion Detection	Image segmentation of corrosion	Corrosion location, size, severity	Image, Numerical	JPEG, CSV, rosbag
Overall Classification of Vessels	Scene identification in images	Scene labels (good, fair, poor)	Image, Categorical	JPEG, JSON, rosbag
Structural Segmentation of Vessels	Component identification in images and 3D point clouds	Component labels (web frame, longitudinal stiffener, etc.)	Image, 3D Data, Categorical	JPEG, PCD, PLY, JSON, rosbag
Mapping of an Area	3D point cloud generation, 3D mesh generation, radiance fields	3D maps, radiance fields	3D Data	PCD, STL, OBJ, LAS, rosbag
Agile Navigation	Trajectory optimization for agile flight	Pose estimation, gate/manhole parameters	Image, Numerical	rosbag
Contact of Aerial Robot's End Effector (Probe) to Target Walls	Contact detection	Contact location, force data	Numerical	CSV, rosbag
Thickness Measurements Using Aerial Robot's Probe	Thickness measurement	Thickness data	Numerical	CSV, rosbag

Co-funded by the European Union

VIO testing on fully-actuated UAS	Camera calibration	Pose estimation ROS messages	ROS Message	CSV, rosbag
User Inputs to User Interface and Decision Support System	User commands, structured & unstructured interviews	User command logs	Categorical	JSON, rosbag, free text
User Experience Feedback	User feedback, structured, unstructured interviews	User feedback data	Textual	TXT, DOCX, rosbag, free text
Data quality evaluation	camera position, lighting	Images, ground truth	Image, Video	JPEG, MP4
Electromagnetic thickness measurements	Calibration standard, inspection parameters	Raw electromagnetic data, thickness data, inspection settings, probe information	Numerical, categorical	TXT, CSV, JSON, binary arrays
Ultrasound thickness measurements	Calibration standard, inspection parameters	UT sensor data, thickness data, inspection settings, probe information	Numerical, categorical	TXT, CSV, JSON, binary arrays, rosbag

TABLE 1 : NON-EXHAUSTIVE LIST OF EXPERIMENT TYPES, RELATED EXPERIMENTAL PARAMETERS, DATA RESULTS AND ASSOCIATED METADATA

The data presented in Table 1 stem from a multitude of sensors. These sensors include RGB cameras, time-of-flight (TOF) cameras, event-based cameras, ultrasonic distance sensors, time of flight distance sensors, 3D laser scanners, multi-degree-of-freedom inertial measurement units (IMUs), ultrasonic thickness measurement probes and electromagnetic thickness measurement probes. Table 2 summarizes all sensors and their associated data types.

Sensor Type	Data Type	File Format
RGB and TOF Cameras	Numerical (pixel values for RGB images), Binary (image files)	JPEG (RGB images), PNG (depth images)
Event-Based Cameras	Numerical (events representing changes in brightness)	TBD (dependent on implementation)
3D Laser Scanner	Numerical (point cloud coordinates), Binary (point cloud files)	PLY, LAS, PCD
Ultrasonic Distance Sensors and Ultrasonic Thickness Measurement Probe	Numerical (distance and thickness measurements)	CSV (text-based)
IMU	Numerical (inertial measurements such as acceleration and angular velocity)	CSV, JSON (text-based)
Control Signals	Numerical (e.g., PWM signals)	TBD (dependent on implementation)
User Inputs and User Experience Feedback	Textual (input text, selections, ratings)	TXT, CSV (text-based)
Electromagnetic Thickness Measurements and Ultrasound Thickness Measurements	Numerical (thickness measurements)	Electromagnetic Thickness Measurements: Proprietary format, Ultrasound Thickness Measurements: CSV, XML (text-based)

TABLE 2 : AUTOASSESS SENSOR DATA OVERVIEW

2 FAIR DATA

The AUTOASSESS data management strategy will comply with the FAIR principles to enhance knowledge integration, promote sharing and reuse of data and help data and metadata to be 'machine-readable'. The following sections tackle different aspects of our data management strategy.

2.1 MAKING DATA FINDABLE, INCLUDING PROVISIONS FOR METADATA

Both researchers working on the AUTOASSESS project as well as the data provided will use digital identifiers. All relevant AUTOASSESS researchers will use an ORCID personal identifier. Digital

identifiers (persistent identifiers - PIDs) will be attached to all AUTOASSESS datasets as provided by the selected data repository. While other public access platforms might be used, the main data repositories for AUTOASSESS are the Zenodo[1] platform and the DTU Data platform. A digital object identifier (DOI) is automatically assigned to all Zenodo files. The same is true for the DTU Data platform.

The AUTOASSESS data will be grouped into folders (or datasets) that are self-sufficient and correspond to a specific field test/experiment. Data can be raw as collected by the sensors, or postprocessed. The data can be numerical (including range data for LiDARs and thickness data for NDT UTM), multimedia (images and videos), categorical or even textual. The data formats used are dictated by the type of data. Rosbags are collected as much as possible for their ability to recreate an active mission —more on that in section 2.3. The data types along with the data formats used in the AUTOASSESS project can be found in Table 2.

Each data file in an AUTOASSESS dataset is complemented by metadata. The metadata include but are not limited to location, date and time, partner collecting data, data responsible, test parameters, sensor calibration and test conditions. Data that are not generated by sensors need to be accompanied by a document description file (usually a .txt) where all the metadata of the file should be provided.

To maximize the likelihood of discovery and possible reuse, keywords will be carefully chosen and included to the metadata. These keywords, which include important ideas, variables, and techniques used in data collecting and analysis, will be chosen according to how well they relate to the study field. Furthermore, the metadata structure will allow easy harvesting and indexing by data repositories and search engines. By adhering to standardized metadata formats and protocols, such as Dublin Core or schema.org, the metadata will be readily ingestible by automated indexing systems. This ensures that the datasets are discoverable not only within the project consortium but also by the broader scientific community, promoting collaboration and knowledge exchange on a global scale.

To provide ease of use and better findability, the file naming scheme of AUTOASSESS raw data will include information such as: purpose, partner date and sensor ID. Post-processed data should use a prefix starting with the file name of the raw data, suffixed by the type of postprocessing. Finally, all data should end with a version number.

2.2 MAKING DATA ACCESSIBLE

It is the intention of the consortium to produce publicly available data to benefit the academic community and the EU industrial community. According to the grant Agreement, the beneficiaries need to provide digital data, including associated metadata, to validate the outcomes of the project as well as the peer-reviewed scientific publications. The deposit of the data will happen as soon as feasible (Article 17 of the Grant Agreement).

To some extent, the publication of data and results might be delayed by our efforts to satisfy the requirements of the ANNEX 5 of the Grant Agreement of the AUTOASSESS project, which includes the following restrictions and specifications:

[1] https://zenodo.org/

Protection of Results (Article 16): Beneficiaries that have received funding under the grant must adequately protect their results for an appropriate period and with appropriate territorial coverage. This is subject to considerations such as commercial exploitation prospects, the legitimate interests of other beneficiaries, and any other legitimate interests.

Exploitation of Results (Article 16): Beneficiaries that have received funding under the grant must use their best efforts to exploit their results directly or have them exploited indirectly by another entity, up to four years after the end of the action. If the results are not exploited within one year after the end of the action, the beneficiaries must use the Horizon Results Platform to find interested parties to exploit the results.

Dissemination of Results (Article 17): Beneficiaries must disseminate their results as soon as feasible, in a publicly available format, subject to any restrictions due to the protection of intellectual property, security rules, or legitimate interests. Any other beneficiary may object within 15 days of receiving notification if it can show that its legitimate interests in relation to the results or background would be significantly harmed.

Open Science: Research Data Management (Article 17): Beneficiaries must manage the digital research data generated in the action responsibly, in line with the FAIR principles. This includes establishing a data management plan (DMP), depositing the data in a trusted repository as soon as possible, and ensuring open access to the deposited data, under certain conditions.

Taking the aforementioned clauses into account, due time will be granted to each beneficiary producing data to evaluate whether the produced data needs to be protected for exploitation, no-harm or other legitimate interests. The final time for this evaluation cannot exceed 1 year after the data collection. In the meantime, the data will be freely shared among partners to allow the efficient execution of the project. The AUTOASSESS consortium will use a multitude of means to store and serve the data, including both private and public options, as shown in Table 3.

Concerning publicly available data, AUTOASSESS will employ at least two repositories. The first option, Zenodo, is a multidisciplinary open repository maintained by CERN. It allows researchers to deposit research papers, data sets, research software, reports, and other digital files. Any file format can be used and, therefore, it poses no restriction to the AUTOASSESS specialized datatypes (see Table). A digital object identifier (DOI) is automatically assigned to all Zenodo files. By using Zenodo we will guarantee compliance with the data management requirements of Horizon Europe. The second repository is the DTU Data repository, which has a long history of data availability and compliance with EU regulations. Additionally, every dataset uploaded at DTU Data is automatically assigned with a DOI.

Concerning the time period that the data remain available and findable as well as the licensing, please see section 2.4. The metadata, described in section 2.1, will be available for a longer period than the data, under the responsibility of the PC.

Data Storage (closed)	Host	Comment
NextCloud	DTU	Internal data management for all beneficiaries (closed)
Github	Microsoft	Subsets of Partners access
Cognite Data Fusion	Cognite (cloud provider Microsoft / Google)	Subsets of Partners access
RaM cloud	UTwente	Subsets of Partners access
OneDrive	UTwente	Subsets of Partners access
Open access	Host	Comment
Zenodo	CERN	Publicly available data
DTU DATA	DTU	Publicly available data
Open access journals	Publisher of open access journal	Open access publishing
Github	Microsoft	Open access deposition
Gitlab		Open access deposition

TABLE 3 : CURRENT LIST OF DATA REPOSITORIES FOR DATA STORAGE AND MANAGEMENT DURING THE PROJECT (CLOSED), OPEN-ACCESS PUBLICATION OF DATA AND FOR ARCHIVING AFTER COMPLETION OF THE PROJECT

2.3 MAKING DATA INTEROPERABLE

Ensuring data interoperability is critical to the AUTOASSESS project to promote data reuse both within the project consortium as well as across disciplines. AUTOASSESS will use community-endorsed interoperability best practices and adhere to established vocabularies, standards, formats, and procedures in order to accomplish this goal.

In addition to standard file formats like CSV or JSON, ROS (Robot Operating System) rosbag files will be used to store raw data from sensors and instruments. Rosbag files facilitate data storage, replay, and exchange within the ROS ecosystem and ensure compatibility and ease of access for researchers and collaborators. Furthermore, for simple sharing and reuse among consortium members, datasets will be exported to widely used formats (See Table 1, Table 2).

The AUTOASSESS project ensures data and metadata interoperability through adherence to established standards, formats, and best practices. The consortium will utilize accepted literature in the field of robotics and inspection as well as standard vocabulary. The project will provide a consistent definition of abbreviations and mappings for project-specific vocabulary. Furthermore, ontologies or vocabularies unique to the project will be made publicly available in order to encourage cooperation and knowledge sharing by allowing the larger research community to reuse, improve, or extend them.

2.4 INCREASE DATA RE-USE

With the exception of data sets that support unfinished peer review publications, patent applications, data that can cause serious commercial harm to the providers, or information that cannot legally be openly accessible, **all data of AUTOASSESS will be made publicly available in open access repositories** (as defined in Table 3) **for at least 10 years**.

As mentioned in section 2.2, when it pertains to data, we will use the Zenodo platform as well as the DTU Data platform. As far as source code is concerned, we will open source as much as possible on the Github platform as well as post packages to the AI on Demand platform, specifically when the software concerns machine learning solutions.

Concerning licensing, the data will be provided with permissive licenses, which will allow its use for research and commercial applications, given the proper acknowledgment is given to the owners. Examples of such licenses are: BSD 3-Clause License, Attribution Share Alike (CC-BY-SA) license.

Regarding the reusability of the data, README files will accommodate data and software utilities —if any. An example of such a README file can be viewed in Figure 1.

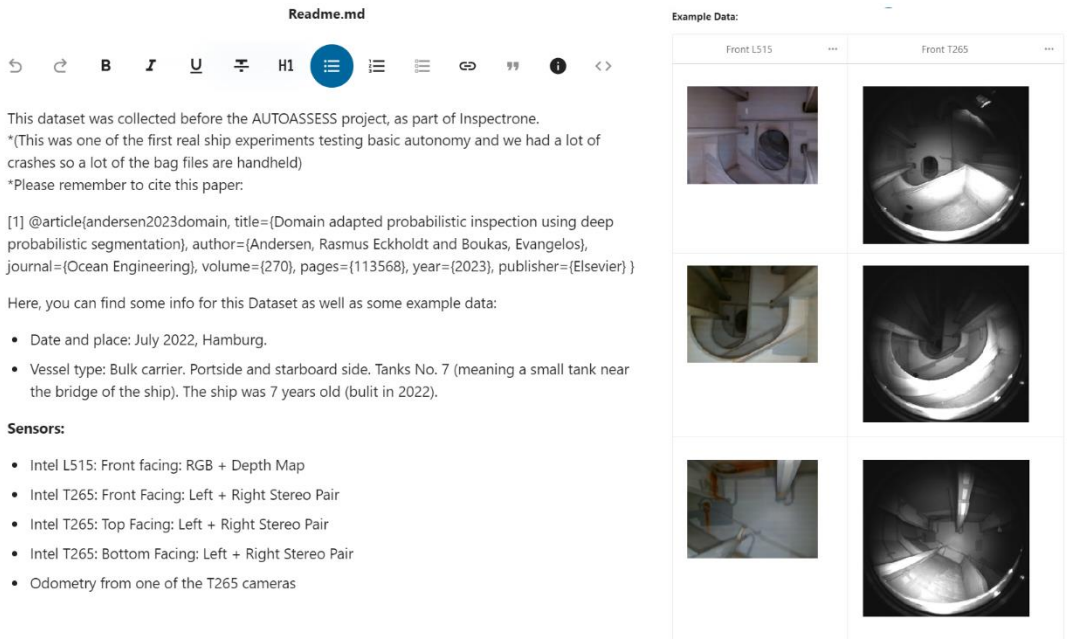

Figure 1: Example of a README file in one of our currently "under processing" datasets. It includes metadata, as well as sample images and videos.

2.5 OTHER RESEARCH OUTPUTS

In addition to the management of data, we foresee the publication of scientific articles in reputable journal and conference publications. When it pertains to these scientific documents,

we will abide by the Article 17 of the HEU Model Grant Agreement and HEU Programme Guide, as follows:

- Early and open sharing
- research output management
- open access
- open peer-review
- involvement of all relevant knowledge actors

Early and open sharing will be achieved through deposition of pre-print (Arxiv, DTU Data) and immediate open access scientific publications. Open access will be ensured through a focus toward publication in open-access journals or Gold Open Access journals. All the publications, with the exception of minor PR material, will be submitted to peer-review journals. With regards to open peer review, we will advise the partners to publish on the Open Research Platform[2] (OPR) of the European Commission. As soon as the OPR gets indexed by Web of Science this will become a stricter requirement. Finally, to address the involvement of all relevant knowledge actors we will try to include our partners in the internal pre-submission review of our scientific publications, with the appropriate acknowledgement in the publication.

3 ALLOCATION OF RESOURCES

The responsibility of guaranteeing FAIR status on the data collected and processed, including its metadata, lies with the consortium members who collect the data. The responsibility to host the data lies with the coordinator partner (DTU) and the PC (Evangelos Boukas). The cost of maintaining the data for the duration of the project will be provided within the project and, thereafter, will be covered by the department of Electrical and Computer Engineering at DTU. Since the solution has been self-maintained by DTU, the total cost for data will be less than 8000€ for the duration of the data availability period.

The costs for data availability during the project execution, when the data is live (meaning editing and processing) will be covered by the project budget. The cost of data after the project duration, while in frozen state, will be minimal and covered by the department of Electrical and Computer Engineering at DTU.

The total period when the preservation of the AUTOASSESS data is ensured is at minimum 10 years after the project end. There are multiple pathways to guarantee the long-term availability of the AUTOASSESS data. The AUTOASSESS will utilize all of them. At minimum, the safe Zenodo platform as well as the long-standing DTU Data platform will be used.

4 DATA SECURITY

Data security for the AUTOASSESS starts at acquisition. As soon as a mission (field test or demonstration) is finished the data will be transferred to two physical mediums, an external storage and the ground control station, as well as the cloud solution of Cognite. At a later stage,

2 https://open-research-europe.ec.europa.eu/

the data should be transferred to our consortium's file storage, i.e.: the AUTOASEESS NextCloud[3]. This system is backed up daily at another DTU server. Given the release from the involved parties, all processed data will also be uploaded to the long-term storage in the open access repository Zenodo and/or DTU Data.

5 ETHICS

The only ethical concern regarding the data acquired is the depiction of human pilots —members of the consortium— while performing tests. While GPPR rights are not waivable, a detailed description of the data purpose and usage will be provided, and a consent form will be completed by the persons depicted. If necessary, similarly to the recent dataset by NTNU[4], we will blur all faces from the dataset.

6 OTHER ISSUES

No other issues need to be addressed.

3 https://inside.autoassess.eu

4 https://github.com/ntnu-arl/ballast_water_tank_dataset